

Mass Spectrometry Data Analysis in the Proteomics Era

Francesca Forner¹, Leonard J. Foster² and Stefano Toppo^{*,3}

¹Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, Martinsried, Germany

²UBC Centre for Proteomics, Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, Canada

³Department of Biological Chemistry, University of Padova, Padova, Italy

Abstract: With the advent of whole genome sequencing, large-scale proteomics has rapidly come to dominate the post-genomic age. As such, tandem mass spectrometry has emerged as the most promising and powerful technique in this area but analysis of raw spectra remains one of the principle bottlenecks to making effective use of the technology. Analytical approaches for identifying proteins from MS/MS data fall into two categories: comparing measured fragment spectra to theoretical spectra from sequence databases and *de novo* peptide sequencing. Available methods still have weaknesses, highlighting the need for new powerful algorithms that are able to exploit the enormous volume of data generated by proteomic experiments. Recent efforts have also been directed towards the identification of post-translational modifications, biomarker discovery and quantitative proteomics. Overall, the intended goal of this review is to give as thorough as possible an overview of state-of-the-art approaches and tools developed to analyze tandem mass spectra in different fields and discuss future directions aimed at overcoming the limits of present methods.

Keywords: Mass spectrometry, proteomics, database searching algorithms, spectra analysis, *de novo* sequencing, quantitative proteomics.

1. INTRODUCTION

After the historic 'completion' of the human genome sequencing in 2000, the scientific community has been catapulted into the post-genomic era and a multitude of so-called '*omics*' approaches have arisen to exploit the avalanche of genomic sequences. One of the most prominent of these is certainly proteomics, which aims to take advantage of genome information to study all the proteins in a cell or organism. Proteomics has been hampered by an imperfect correlation between predicted open reading frames (ORFs) and true coding genes. However, since the real business of cells is largely carried out by proteins rather than nucleic acids, our understanding of biology depends on us overcoming this and other limitations to unravel protein function. Thanks to continued technological improvements in mass spectrometry, proteomics is now at a high throughput stage and is going through the same critical phase that genomics went through in data management and interpretation.

To introduce the reader to the mass spectrometry landscape, we will first provide a technical overview of current instrumentation, followed by the basic aspects of peptide mass spectrometry such as patterns of peptide fragmentation (tandem MS) and the determination of peptide charge state. The bulk of this review will then be dedicated to tandem mass spectra interpretation by either database matching or *de novo* sequencing, highlighting weaknesses of current methods and future areas of development. We will conclude with a discussion of three areas we believe will be the major foci

in the next few years: biomarker discovery, identification of post-translational modifications and protein quantitative measurements. Throughout this text we have endeavored to discuss all publicly available bioinformatic tools currently used in proteomics but some may have been accidentally overlooked.

1.1. Proteomics World

A proteome is defined as the complete set of proteins, including those that are alternatively spliced and post-translational modified, that is expressed in the lifetime of a cell. It is also used in less general sense, to represent the complement of proteins expressed by a cell or organelle at any one time [1]. Thus, proteomics is the study of all proteins in a particular state or condition of a cell or organelle. The last few years have seen a steady expansion of proteomics applications to cover everything from cell biology and biochemistry to clinical diagnostics and several areas in between [2-19] (Table 1). The reader is referred to other sources for reviews of these topics [20].

1.2. Mass Spectrometry Technology and Instrumentation

Mass spectrometry is now a fundamental platform for proteomic research, thanks to its unsurpassed capacity for accurate protein identification and quantitation. In its simplest description a mass spectrometer measures the mass-to-charge ratio (m/z) of ionized molecules and its basic components are represented in Fig. 1. In principle, any ionizable molecule can be measured in a mass spectrometer but we will focus on peptides and proteins for the purpose of this review. Peptides are introduced into a mass spectrometer at the ion source in the form of liquid solutions (or solid mixtures), then desolvated (or desorbed from the matrix) and transferred into the gas phase as gas-phase ions. This ionization process is of primary importance since mass spectrometry

*Address correspondence to this author at the Department of Biological Chemistry, University of Padova, V.le G. Colombo 3, I-35121 Padova, Italy; Tel: +39 049 827 6958; Fax: +39 049 807 3310; E-mail: stefano.toppo@unipd.it

Table 1. Applications of Proteomics

| | Focus | Ref. |
|----------------------|---|---------|
| Modification "omics" | Study of post-translational modifications | [8] |
| Organelle proteomics | Cell organelle protein mapping | [2,3] |
| Clinical proteomics | Diagnosis of diseased status | [4-17] |
| Interactomics | Composition of protein complexes Protein-protein interaction networks | [18,19] |

ters can only measure charged species. There are two principle methods used for ionizing peptides: electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI).

The electrospray process was first described by Fenn and co-workers in 1985 [21,22]. Essentially, an acidic solution containing the analytes is sprayed through a thin needle kept at high voltage into the ion source where the analyte ions are transferred into the gas phase after solvent evaporation. At low pH and positive voltage, peptides are protonated at the N-terminal amine moiety and at the basic side chains of lysine, arginine and histidine residues. As a result, multiply charged peptides are formed and this is a unique feature of electrospray ionization (ESI) compared to other MS ionization techniques. Matrix-Assisted Laser Desorption/Ionization (MALDI) was first described by Karas in 1988 [23]. Peptides are mixed with a solution (matrix) containing an UV-absorbing organic acid and deposited in well plates. Peptide

ions (mostly protonated) are produced by bombarding the sample with short-duration (1-10 ns) pulses of UV light from a nitrogen laser [24].

The most widespread mass analyzers are quadrupoles, ion traps, Time Of Flight (TOF) and Fourier Transform-Ion Cyclotron Resonance (FT-ICR). Hybrid combinations of these analyzers are very common, such as the triple-quadrupole [25], quadrupole-time of flight (qTOF) [26], time of flight-time of flight (TOF-TOF) and ion trap/FT-ICR. In the triple-quadrupole configuration a precursor ion can be selected in the first mass analyzer (Q1), fragmented in the collision cell (Q2) and the resulting product ions separated by scanning in the second mass analyzer (Q3). In Q2, collisions between analyte ions and an inert gas produce fragment ions that provide important structural information [27-29].

In TOF analyzers ions are accelerated to high kinetic energy and enter a field-free region where each ion travels at a

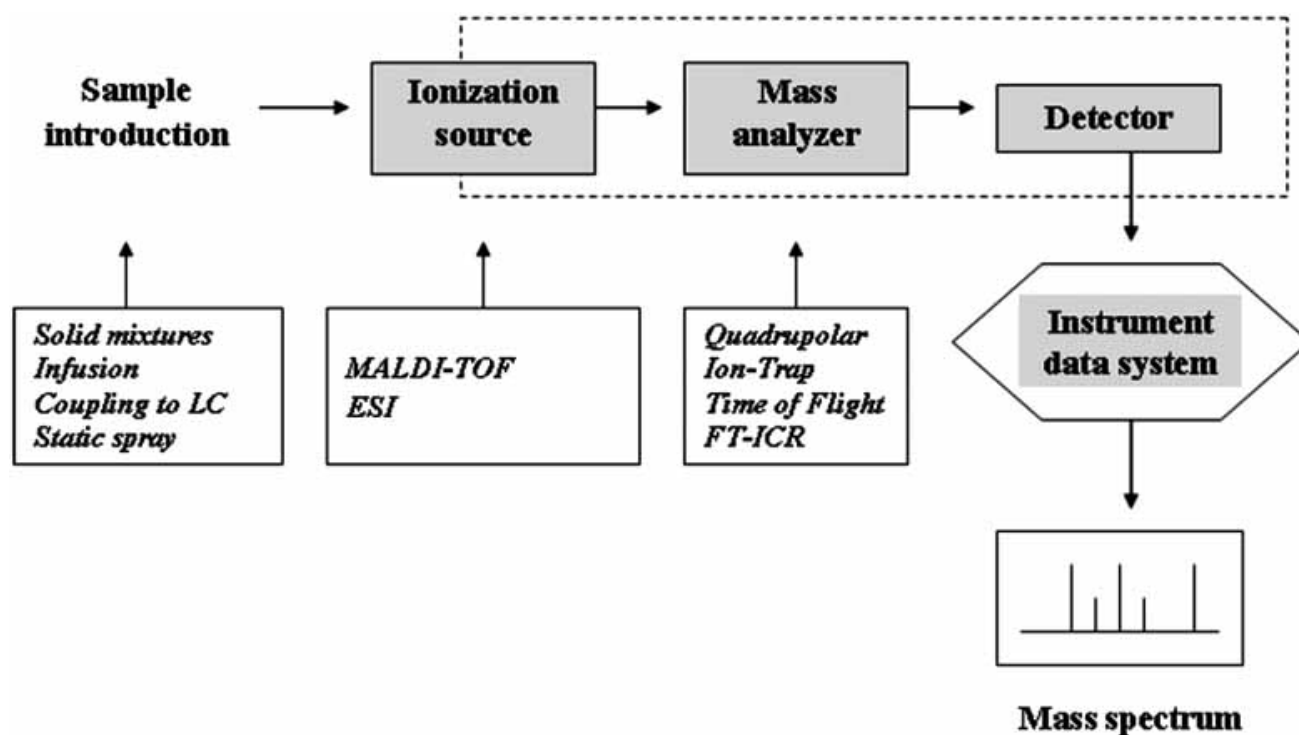


Fig. (1). Overview of a mass spectrometry system. The basic components of a mass spectrometry system here represented are: sample inlet, ion source, mass analyzer, detector, vacuum system (dotted line), instrument-control and data analysis systems. The instruments used in proteomic experiments are composed of different combinations of inlet systems, ion sources and mass analyzers. The inlet system drives analyte molecules into the source, where gas-phase ions are generated and led into the mass analyzer through precise control of electromagnetic fields. The behaviour of gas-phase ions in the electromagnetic field is in some way correlated with their m/z (mass over charge ratio).

velocity inversely proportional to its m/z ratio [30] – in effect the time it takes to 'fly' a certain distance is the base value that is measured. These analyzers can deliver resolutions near 10,000 (measured as the Full Width of a peak at Half its Maximum height, FWHM) and mass accuracies down to 10 ppm while detecting as little as 10 fmol of a peptide but require approximately one second to measure each spectrum. For examples of clinical proteomics applications of MALDI-TOF instruments (see ref. [31,32]). In three-dimensional ion traps, radio frequency (RF) voltages are applied so that all ions initially oscillate inside the mass analyzer. Ions at particular m/z values can then be selected and fragmented by modulating the RF-voltage, and the fragments scanned out and detected. By themselves ion traps have low resolution and mass accuracy (near 500 FWHM and 500 ppm respectively) but are extremely sensitive (1 fmol) and fast (spectra acquired in hundreds of milliseconds). FT-ICR analyzers essentially trap ions in the core of a very strong magnetic field that measures the precession of ions in a cyclotron resonance past a detector. The orbital speed of ions in the ICR is proportional to their m/z and thus a fast Fourier transform operation can be used to convert the measured frequency to a mass spectrum. While slow (one spectra per second), such detectors are able to measure m/z values more accurately (<1 ppm) and with greater resolution (>100,000 FWHM) than any other method [33]. An Orbitrap detector has been described more recently that uses a similar detection mechanism to FT-ICR but in the absence of a magnetic field [34]. For a comprehensive review of mass spectrometer technology the reader is directed to reference [35].

The four parameters alluded to above, resolution, mass accuracy, duty cycle and sensitivity, determine the effectiveness of a mass spectrometer. As has been mentioned, FT-ICR detectors have by far the highest resolution and mass accuracy but until recently have had such a slow duty cycle as to render them useless for proteomic studies. Ion traps, on the other hand, have the fastest duty cycles (the speed with which a cycle of MS and MS/MS spectra can be acquired) and sensitivity but without careful validation their low resolution and mass accuracy can easily lead to their data being misinterpreted [36,37]. qTOF instruments have intermediate specifications in each category and for many years have been the workhorses of the more advanced proteomics laboratories. The development of an ion trap/FT-ICR hybrid [38], and even more recently of an ion trap/Orbitrap hybrid [36], has revolutionized proteomics by combining the best features of each component into one instrument. High mass accuracy is extremely beneficial for database-dependent peptide identifications because it reduces the number of sequence candidates [39]: FT-ICR and Orbitrap instruments can determine the unique amino acid composition of a peptide in many cases without even fragmenting them.

Regardless of the sub-specialty of proteomics, mass spectrometers can generate reams of mass spectra which all require interpretation to identify and quantify peptides. The remainder of this review will focus on algorithms and software developed to aid the analysis of these data.

2. GENERAL ASPECTS OF PEPTIDE FRAGMENTATION

In tandem mass spectrometry (MS/MS) the physical fragmentation process is a delicate step that influences the

final spectrum quality and, consequently, its interpretation. Various groups have attempted to unravel the underlying physical events that should explain the peptide fragmentation process from both the quantitative and qualitative aspects.

Most commercial MS instruments use a gas-phase [40,41] collision-induced dissociation (CID) at low energy (<100eV). The collisions of the inert gas impart enough energy to the peptide ions that they fragment in a pattern that depends on both the collision energy used and the amino acid composition of the peptide [40,42,43]. Low energies preferentially cleave the peptide backbone bonds whereas high energies generate complex MS/MS spectra that contain additional fragments resulting from the dissociation of side chains bonds and that can be difficult to analyze [44]. The accepted nomenclature for the fragment ions formed in collision induced dissociation was proposed by Roepstorff [45] and expanded by Biemann [46] (Fig. 2). The most abundant fragments are usually *b* and *y* ions (Fig. 2), complementary fragments derived from charge capture either on the N- or on the C-terminus respectively [47]. For this reason most current automated methods and algorithms consider only these ions species. In quadrupole instruments *y* ions usually predominate, whereas in ion traps both *y* and *b* ions can be observed.

There are many aspects of tandem mass spectra that must be taken into account before a thorough interpretation of the signals can be attempted. The widely accepted "mobile proton model" introduced by Dongré *et al.* [48] states that protons can migrate along the peptide backbone prior to fragmentation. Indeed, molecular orbital calculations have demonstrated that protonation of the amide nitrogen weakens the amide bond and can lead to fragmentation at this bond [49,50]. The mobile proton model also explains the role of remote competitive mechanisms (like the aspartic acid effect) when basic amino acid side chains sequester the proton. These mechanisms seem straightforward but in practice the number of observed fragments is always less than the theoretical number and the intensities of fragment peaks can vary from spectrum to spectrum [51] so clearly there are other factors at work. The reader is referred to the studies of Tabb *et al.* and of Kapp *et al.* for a description of the state-of-the-art statistical characterization of tandem mass spectra [40,52]. For instance, certain peptide bonds seem more prone to fragmentation than others and lead to signals of higher intensity, whereas other fragments are never observed in MS/MS spectra. In proline-containing peptides enhanced bond cleavage is often observed at the amino-terminal side of proline residues [42,50,53]. Paizs and Suhai pointed out that an even more detailed understanding of MS/MS fragmentation mechanisms and of fragment ions abundances is needed in order to improve protein identification through bioinformatic tools [47]. Various studies have tried to pinpoint which factors are most important for interpreting fragment spectra but there is still no agreement on what those might be [40,54]. Such efforts have met with some success, however, as a recent study demonstrated that a machine learning approach could reduce the peptide identification error by 50-90% without any loss in sensitivity [54]. A web-based interface to this approach, SILVER, has been devised to assist manual curation of tandem mass spectra [55]. Attempts have also been made to predict the distribution of intensities based on the mobile proton model [44] and a sepa-

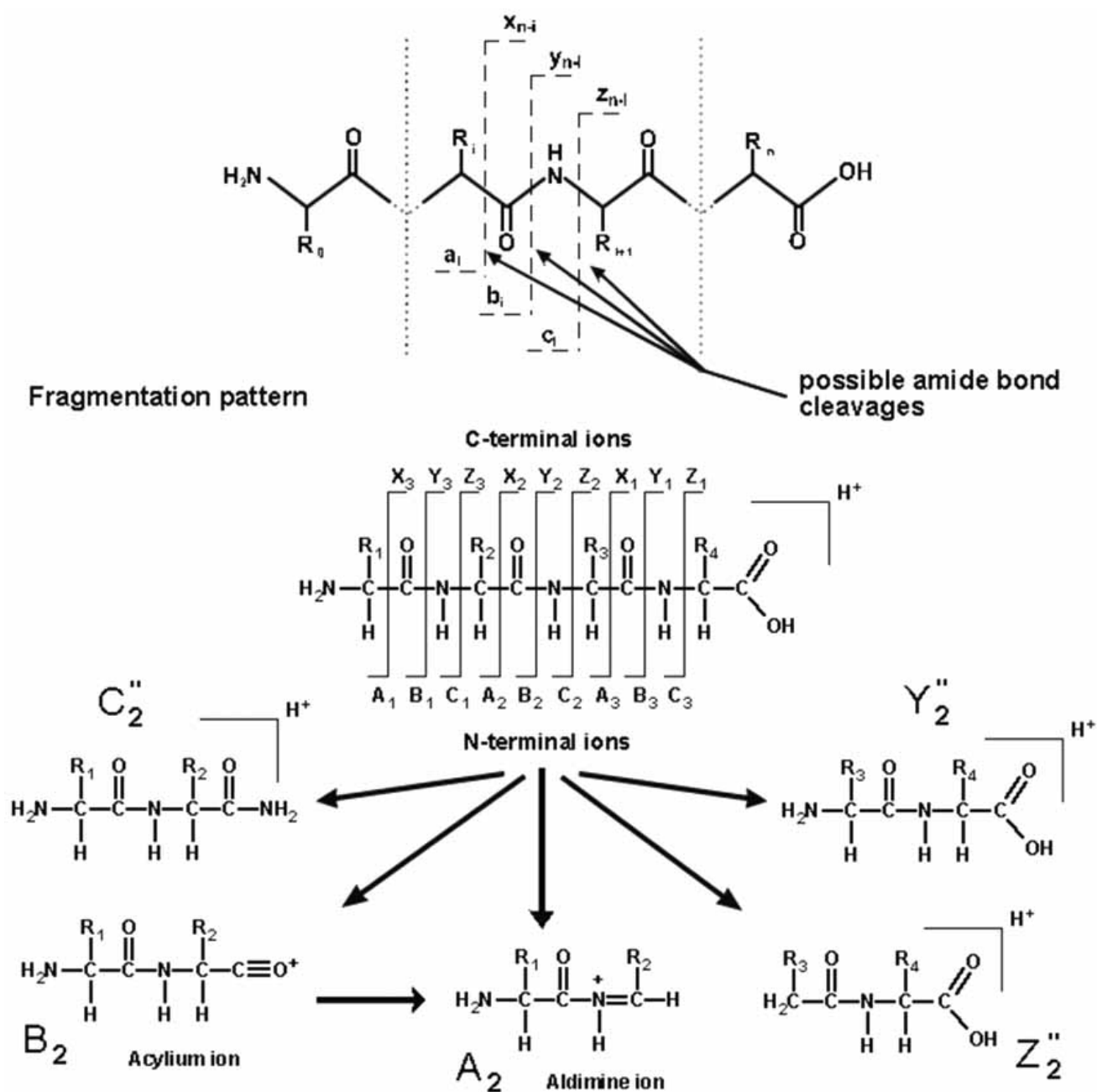


Fig. (2). The cleavage of the peptide ion occurs, preferentially, along the backbone bonds at low-energy collisions. The resulting two fragments contain either the N-terminus or the C-terminus of the starting peptide. The standard nomenclature for the C-terminal fragments is x, y and z whereas the corresponding N-terminal fragments are denoted as a, b and c depending on the position where the breakage effectively occur at the amide bond level. Ideally, the final result of the dissociation process is a heterogeneous population of all the possible fragments species that can be obtained from a peptide. If the peptide is n -amino acid long and the y -ion series is considered, the resulting MS/MS spectrum should contain $n-1$ fragments that differ of one amino acid mass from each other. The numbering of each fragment starts from N-terminus for (a,b,c) series and from C-terminus for (x,y,z) series. In tandem mass spectrometry the strongest signals derive from y -ions and the corresponding b -ions.

rate model has been proposed to describe the peak distribution as a function of physical parameters like collision kinetics and bond energies [56]. While tools for incorporating fragment ion intensities do exist, the approach has yet to become incorporated into widely popular MS/MS interpretation algorithms.

3. CHARGE STATE DETERMINATION

In a typical tandem MS experiment the first piece of data that is collected about a peptide is the m/z of the precursor or parent ion. Consequently, if the charge state of the ion is known, the mass of the original peptide is equal to the number of charges multiplied by the mass difference between the

observed m/z value in the spectrum and the proton weight ($H^+ \approx 1$ Da) that confers a net positive charge of one to the ion. In MALDI experiments peptides acquire only one proton and while this rule is not absolute, it is a very good first approximation and no independent determination is needed. However, in electrospray ionization at low pH peptides typically acquire two or three protons (i.e., two or three charges) but the number can vary widely depending on the size and composition of the peptide so the charge state cannot be assumed. For example, a peptide with 1500 Da molecular weight will display m/z 1501 $[M+H]^+$ in MALDI, and might display m/z 751 $[M+2H]^{2+}$ or m/z 501 $[M+3H]^{3+}$ in electrospray. The correct determination of the unknown charge state of the peptide ion clearly influences the measurement of peptide mass. In high resolution instruments like QTOFs or FT-ICRs the charge state can be easily determined from an examination of the isotope cluster, the result of naturally occurring heavier isotopes of the atoms comprising the peptide. Of the elements commonly encountered in peptides (C, O, H, N) the carbon isotope with one additional neutron, ^{13}C , is the most abundant, occurring at approximately 1.1%. Therefore,

in a peptide containing one hundred carbon atoms there is a very good likelihood that one of them is ^{13}C . Extend this to all potential isotopes of all elements in a peptide and the result is a cluster of peptides with the same amino acid composition that differ from one another by one m/z unit when singly charged, one half a m/z unit when double charged, etc. (Fig. 3). It is this difference between the isotope peaks that is exploited to determine the charge state in higher resolution instruments.

Unfortunately the charge state inference based on isotope distribution is generally not possible with low-resolution mass spectrometers like ion traps, but given the high importance of accurate charge state assignment, several approaches have been proposed. In the pioneering work of Mann *et al.* two algorithms were described for extracting mass information from spectra containing multiply-charged ions, thereby named the “averaging algorithm” and the “deconvolution algorithm” [57]. Deconvolution, based on the reasonable assumption that charge is quantal and can therefore only have integer values, transformed a cluster of multi-

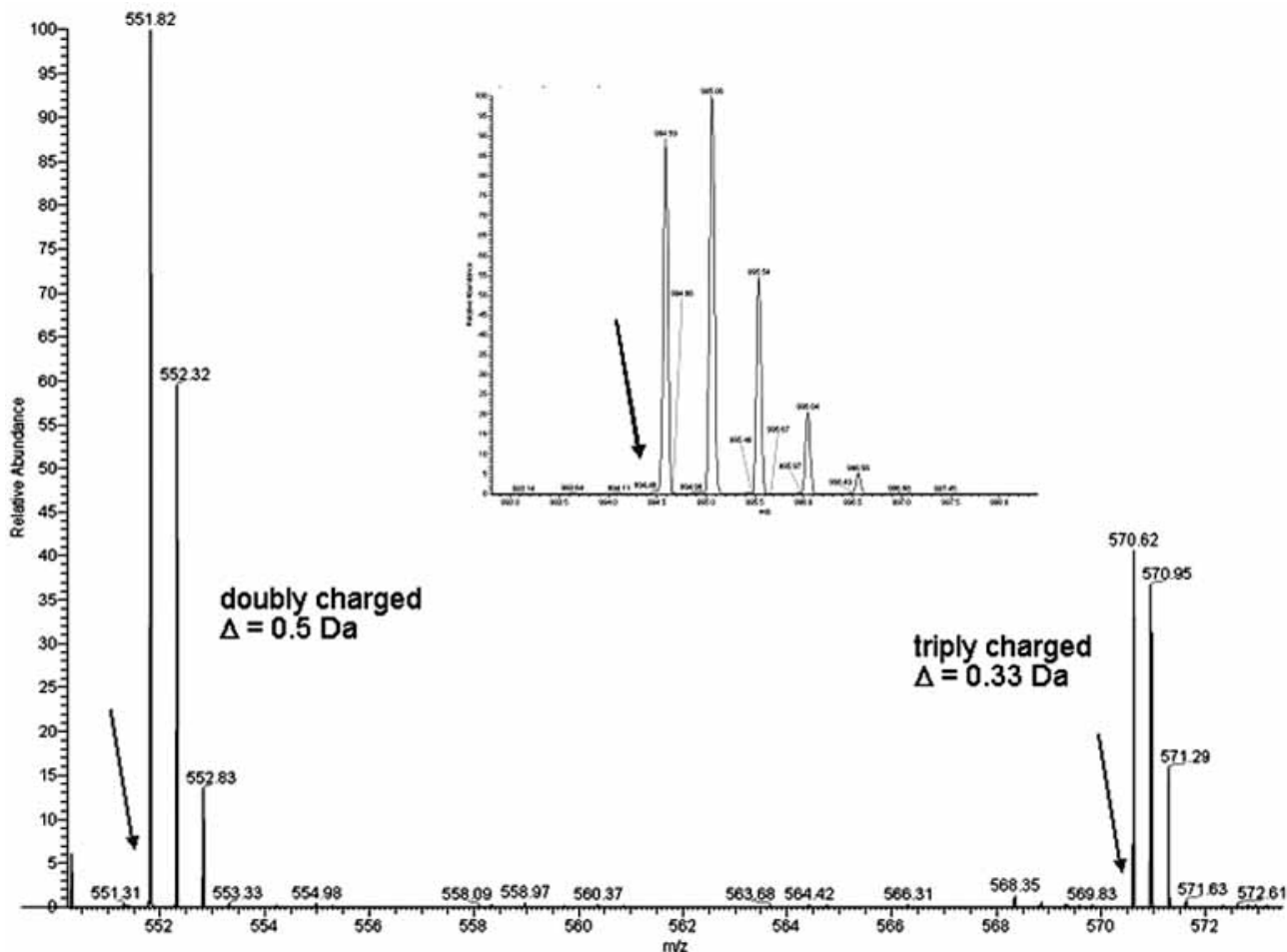


Fig. (3). MS spectrum of two different peptides acquired at 50000 resolution in a ion-trap-FT mass spectrometer. The doubly charged peptide has m/z 551.82 (monoisotopic mass), whereas the triply charged has m/z 570.62 (monoisotopic mass). In general, depending on the molecular mass, the relative height of the isotopic peaks will show different ratios. In particular, with increasing mass the monoisotopic peak becomes less predominant (see the isotope envelope for a 1987 Da peptide in the inset). Arrows indicate the monoisotopic peak. When the isotope cluster is resolved, the measured mass is based on the m/z of the monoisotopic ion. Indeed, as the exact elemental composition of this ion is known, it represents the most accurate peptide mass measure. When the cluster can not be resolved, peaks from different isotopes collapse into a single m/z peak with insufficient resolution and an average mass has to be calculated.

ply charged ions into a singly charged peak located at the molecular mass of the parent compound. The performance decreases with increasing mass due to increased peak multiplicity. The recently published *Thrash* algorithm is particularly effective at identifying isotopic clusters and in deriving mass values from complex electrospray spectra, thanks to the implementation of several features such as the location of isotopic clusters, resolving of overlaps, charge determination, least-squares isotopic abundance distribution-fitting and a novel signal-to-noise calculation [58]. Isotopic clusters are identified in “moving data windows” across the MS spectrum by considering user-defined intensity thresholds.

Accurate determination of the peptide m/z ratio is also extremely important for avoiding errors in precursor ion mass measurements that can lead to systematic errors in *de novo* peptide sequencing (see below). The complementarity of the b - and y -ion series has been proposed to identify the right molecular weight of the parental fragment [59] and has been further exploited to determine the ion charge state [60]. The method is based on the assumption that for each fragment mass F_i , its complement is calculated as $G_i = m(P) - F_i - d$ where $m(P)$ is the parent peptide mass and $d = m(y\text{-ion}) - m(b\text{-ion})$ is the difference of the offsets between y -ions and b -ions. This difference is a constant that depends on the ion series where F_i and its complement G_i are observed, since complementary fragments in b and y series have a mass difference that depends on the charge state. The algorithm tries to find all the possible complementary ions in the spectrum and match them to different hypothesized charge states. The correct charge state is predicted when the category that contains the most abundant complementary ions is found.

Recently, a new method for the deconvolution of tandem mass spectra and charge state assignment has been described and called CRAM (Charge Ratio Analysis Method) [61]. The method identifies the charge state of multiply charged ions without any prior knowledge of the nature of the charge-carrying species that may be different from the proton H^+ . The theory is based on the observation that the ratio of any two multiply charged ions ($(R_z)_a$ and $(R_z)_b$) approximately corresponds to the inverse ratio of their charges (z_b and z_a) (equation 2). Since charges must be integer values, the comparison with a pre-calculated table of all possible ratios of integers reveals the charges of the two ions and the mass (m) of the charge carrying species (equation 3). Consequently the real ion mass (M) may be calculated (equation 1):

- 1) $R_z = (M \pm zm) / z$ R_z is the mass-to-charge ratio of the ion, m is the mass of the charge-carrying species, M is the peptide mass and z is the charge number
- 2) $(R_z)_a / (R_z)_b = z_b / z_a$ a and b indexes are the two multiply charged ions
- 3) $((R_z)_a z - (R_z)_b z) = \pm m (z_a - z_b)$.

In summary, the easiest method for determining charge state is to simply use a higher quality mass spectrometer with the resolving power to separate isotope clusters. In the absence of such an instrument though, the initial deconvolution approach [57] or one of its later manifestations seems most effective.

4. PEPTIDE AND PROTEIN IDENTIFICATION: METHODS AND STATISTICAL EVALUATION

Tandem mass spectrometers can now fragment several thousand peptides per hour [62-64] and this level of data production requires robust algorithms for effective and timely interpretation. Current software tools are a balancing act between statistically solid data on one hand and increased peptide identifications on the other, with the latter usually winning out. In general there are three distinct approaches that have been developed for interpreting tandem mass spectra [62,63,65-72]: a knowledge-based procedure where the potential protein sequences are already known, an *ab-initio* approach and the sequence tag methods (Fig. 4). The first of these is also called the Shared Peak Count approach (SPC) and involves database search methods that try to find the highest similarity between experimental and theoretical tandem mass spectra, the latter derived from protein sequence databases. This approach is by far the most widely used despite limits and drawbacks that will be discussed later. The *ab-initio* methods, on the other hand, try to infer the peptide sequence directly from tandem mass spectra considering the typical fragmentation pattern and trying to reconstruct the full-length peptide sequence after exploring all possible solutions from the measured peaks. This approach suffers heavily from a lack of both sensitivity and specificity and is still time consuming from a computational point of view. The so-called Peptide Sequence Tag approach (PST) is a hybrid method that implements *de novo* sequencing algorithms and sequence similarity searches of small sequence tags versus protein databases. While it is thought to be more specific and sensitive than either of the first, PST has not been widely used due to a lack of automation. The choice of approach depends on the specific task and some differences will be further outlined giving a general overview of the state-of-the-art in the field (see Table 2 for a summary of pros and cons of the methods mentioned in this review and Table 3 for their web resources).

4.1. Database Search Methods: The Shared Peak Count Approach (SPC)

The basis for the most successful methods for tandem mass spectra interpretation is the comparison of acquired spectra versus a database of theoretical peptide fragments [69]. Virtually all commercialized search algorithms use SPC because of the relatively simple and robust implementation. The input data for an SPC engine typically contains the following information: 1) some limits on the terminal amino acids expected based on the endopeptidase used, 2) the measurement accuracy of the mass spectrometer used, 3) the database of protein sequences from the narrowest taxon available, 4) for each fragmented peptide two pieces of information: a) the precursor ion mass and b) the m/z ratios and intensities of all fragments originating from said precursor (the spectrum). Conceptually the first step is then an *in silico* digest of the protein sequence library using the specified cleavage rules (e.g., C-terminal of R and K for trypsin, except when followed by a P). Then for each fragment spectra the list of *in silico*-digested peptides is reduced to only those with a molecular mass equal to the precursor ion within the error limits – the ‘mass-matched peptides’. For

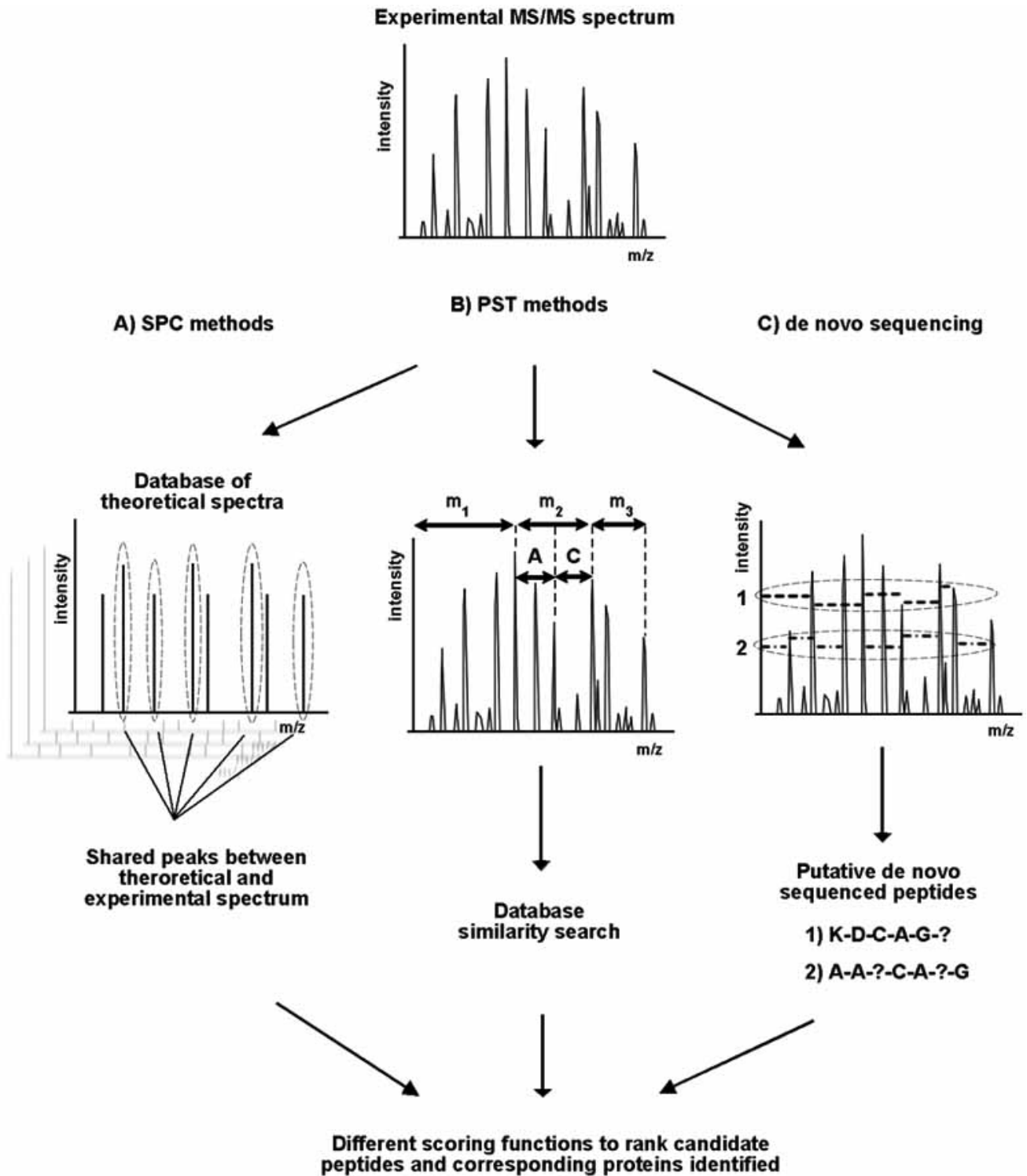


Fig. (4). schematic representation of the different peptide and protein identification methods used in tandem mass spectrometry. SPC methods (A) use the shared peak counts (SPC) that are based on the comparison of the experimental tandem mass spectrum with a pre-calculated library of theoretical spectra derived from a protein database. In peptide sequence tag methods, the so-called PST approach, (B) a local *de novo* sequencing of unambiguous regions of the experimental spectrum is performed to extract small sequence tags flanked by unresolved regions, m_1 and m_3 , whose masses are known. A subsequent error-tolerant, mass-informed similarity search against a protein database is done to retrieve candidate sequences that better fit into the experimental spectrum. The last approach is *de novo* sequencing (C). The goal of these methods is to extract a putative global sequence exclusively from the interpretation of the tandem mass spectrum on the basis of fragmentation rule constraints. Putative candidates, that may be incomplete due to intrinsic problems inherent to tandem mass spectra acquisition, are extracted and ranked accordingly to different statistical models. Presently, the distinction between PST and *de novo* sequencing methods is disappearing. More effective hybrid methods employ a more efficient sequence tag generation, derived from *de novo* sequencing techniques, coupled with a powerful error-tolerant similarity search against protein databases (see text for details).

Table 2. Summary Table of Partial Pros and Cons of Methods Mentioned in the Text

| Categories | Pros | Cons |
|---|--|--|
| Deconvolution Methods | Extracting mass information from spectra containing multiply-charged ions | - |
| CRAM | Effective at identifying isotopic clusters | - |
| Thrash | No prior knowledge of the nature of the charge-carrying species | - |
| SPC Methods | Reliable and widely used | Unable to identify unknown proteins that are not present in the searching database |
| Sequest | Wide spread and not affected to incomplete fragmentation spectra | It does not give a statistical confidence of the hits |
| Mascot | Wide spread and probability based | Not reviewed yet |
| X!TANDEM | Efficient search engine | - |
| SONAR,SALSA | - | Non-probability based |
| OMSSA, ProbiD, PROFOUND, ProbiD, pFind, Phenyx, SCOPE | Probability based | - |
| X!HUNTER | Match against a library of experimental spectra | - |
| De Novo Sequencing Methods | They do not need a searching database. They, potentially, can identify unknown proteins. | Unable to interpret complex spectra and low resolution spectra |
| SHERENGA, PepNovo, Lutefisk | Sequence graph theory | Training set |
| AuDeNS | Sequence graph theory, user intervention | - |
| PRIME | Identification of b- and y- ion series | - |
| PEAKS | Combinatorial method | - |
| RAId | Combinatorial method based on a two-step strategy | - |
| Sequit | Specific instrument application | - |
| PST Methods | - | Limited to small and easy readable portions of the spectrum |
| PeptideSearch, ProteinProspector | Easy tag detection | Limited |
| Hybrid Methods | | |
| GutenTag | Multiple sequence tags detection | - |
| Popitam | Graph theory for multiple sequence tags detection | - |
| Others | | |
| ProbiDTree | find possible multiple peptides from the same fragmentation spectrum | - |
| Post-Processing Statistical Assessment Methods | Independent statistical validation of results for widely used SPC methods | - |
| PEP_PROB/PeptideProphet | Confidence measure of peptide identification and automatic determination of the cut-off | - |
| PROT_PROB/ProteinProphet | Confidence measure of protein identification and automatic determination of the cut-off | - |
| DTASelect, CHOMPER, INTERACT | Confidence measure of peptide and protein identification | Empirical scoring threshold |
| PROVALT, MATH | Confidence measure of peptide and protein identification | Tailored for MASCOT only |
| Database Searching | | |
| DBDigger, PFSM | Reduce computational time | - |
| MultiTag | sequence similarity search optimized to be error tolerant | - |

(Table 2) contd.....

| Categories | Pros | Cons |
|--|---|------|
| FASTS, CIDentify, MS-BLAST | Based on known sequence similarity search algorithm | - |
| OpenSea, PepNovoTag, LocalTag, InSpecT, SPIDER, DeNovoID | Evolved sequence similarity search engines managing multiple sequence tag from a single spectrum and mass-based information. <i>De novo</i> sequencing techniques | - |
| PTM Methods | | |
| GlycoFragment, GlycoSearchMS | Tool for generating and searching all possible fragments of complex carbohydrates in a database | - |
| ProMoST, ProteoMod | pI and mass shift calculation in modified proteins | - |
| P-Mod | discover unanticipated protein modifications from database identified peptides | - |
| Quantitation Methods | | |
| ZoomScan, RelEx, EXTRACT-CHRO, XPRESS | Relative quantification | - |
| MSQuant | Relative quantification, adapted to any survey scan-based quantification | - |
| AQUA, MassView | Absolute quantification | - |

each of these mass-matched peptides a hypothetical fragment spectrum is generated and compared to the measured spectrum. To simplify the analysis many SPC engines assume that the fragment spectrum contains only *b*- and *y*-ions (Fig. 5). Some scoring function is then used to evaluate the closeness of this match for each mass-matched peptide: the peptide with the best score is considered the most likely to have given rise to the measured spectrum. The main differences between SPC engines, where they are fully understood, mostly occur in the scoring schemes used to assess the likelihood of a match and they fall into two classes: The first class comprises tools that correlate acquired mass spectra against theoretical spectra derived from a sequence database and calculate the quality of the match rather than its probability of occurrence (non-probability based models). The second class of methods calculates the probability that the match between fragment ions in the experimental and predicted spectra is a random event (probability based models).

4.1.1. Non-Probability Based Models

SEQUEST [73-75] is the oldest and the best known commercial SPC engine used in proteomics. It utilizes a cross correlation (Xcorr) function to assess the quality of the match between a tandem mass spectrum and amino acid sequence information in a database. No variable peak intensity information is used but SEQUEST does model all *y*-ions to be a uniform intensity, with *b*-ions a uniform lower intensity. The two-tiered score system is based firstly on the peak intensity of fragment ions matching the predicted sequence ions retrieved in the database and secondly on the cross-correlation of the experimental and theoretical spectra. This value is a sensitive metric based on the features of the measured tandem mass spectrum and the quality of its fit to the model spectrum. An empirical reward is also made for the presence of a consecutive ion ladder in a spectrum (e.g., y_3 , y_4 , y_5 , y_6) since specific fragmentation events rarely occur independently of one another. SEQUEST also computes a parameter called ΔC_n , which is the normalized score ob-

tained from the difference between the first and the second-ranked sequences. ΔC_n is a measure of the uniqueness of the match and is dependent on database size and the presence of highly similar sequences; a ΔC_n value greater than 0.1 is a widely used confidence limit for identifying a true positive hit. The SEQUEST algorithm is robust enough to overcome the incomplete fragmentation of a peptide but likely returns significant false positives since larger peptides score higher than high-quality smaller peptides and potentially noisy spectra that are highly populated can have high cross-correlation scores due to random matches of background noisy spikes in the database. Xcorr is not a probabilistic score and does not give a statistical confidence of the hits found even though it has proved to be surprisingly robust for low signal-to-noise spectra. For these reasons SEQUEST remains a primary search algorithm whose output can be further validated by various statistical approaches [64].

SONAR [76] and SALSA [77,78] belong to the same category of descriptive models of fragmentation pattern and similarity searches. SONAR uses an algorithm similar SEQUEST but the score is based on the dot product of experimental and theoretical spectra. SALSA uses a different matching criteria of theoretical and experimental spectra whose alignment starts from the highest experimental peak superimposed to the first ion in the theoretical spectrum regardless their absolute positions on the *m/z* axis.

4.1.2. Probability Based Models

Perhaps the most widely used software in the SPC category of probability based models is MASCOT [79], available both as a limited web service and as a commercial standalone platform. The actual algorithm implementation has not been published but the working core system is based on the MOWSE score [80] that computes the normalized frequencies of peptide (M + H)⁺ distributions from the database. It is probability-based and the predicted fragments are matched to the experimental fragments starting from the

Table 3. Web Resources and Software Mentioned in the Review with References: 1. Web Resource 2. Downloadable or Upon Request Software 3; Commercial Versions

| SPC Based Methods | | |
|--------------------------------------|-----------|---|
| SEQUEST ³ | [73-75] | http://fields.scripps.edu/sequest/index.html , http://www.thermo.com |
| SONAR ^{1,3} | [76] | http://65.219.84.5/service/prowl/sonar.html |
| SALSA ³ | [77,78] | http://www.mc.vanderbilt.edu/lieblerlab/salsa.php , http://www.thermo.com |
| Probid ² | [86] | http://projects.systemsbiology.net/probid |
| ProFound ^{1,3} | [87] | http://65.219.84.5/service/prowl/profound.html |
| MASCOT ^{1,3} | [79] | http://www.matrixscience.com/ |
| X!TANDEM ^{2,3} | [81-83] | http://www.thegpm.org/TANDEM/ , http://www.proteomesoftware.com |
| OMSSA ^{1,2} | [84] | http://pubchem.ncbi.nlm.nih.gov/omssa/ |
| pFind ^{1,2} | [88] | http://pfind.jdl.ac.cn |
| Aldente ¹ | [91] | http://www.expasy.org/tools/aldente/ |
| Phenyx ^{1,3} | [91] | http://phenyx.vital-it.ch/pwi/login/login.jsp |
| Result Validation Programs | | |
| PeptideProphet ² | [103] | http://peptideprophet.sourceforge.net/ |
| ProteinProphet ² | [106] | http://proteinprophet.sourceforge.net/ |
| DTASelect ² | [107] | http://fields.scripps.edu/DTASelect/ |
| CHOMPER ² | [108] | http://www.ludwig.edu.au/jpsl/jpslhome.html |
| AMASS ² | [118] | http://www.ia.ac.cn/personal/fuxin.li/amass.htm |
| INTERACT ² | [109] | http://tools.proteomecenter.org/Interact.php |
| De Novo Sequencing Methods | | |
| SHERENGA ³ | [59] | http://www.chem.agilent.com (Spectrum Mill suite) |
| PepNovo ^{1,2} | [140] | http://peptide.ucsd.edu/pepnovo.html |
| Lutefisk ² | [141,142] | http://sourceforge.net/projects/lutefiskxp |
| AuDeNS ² | [143] | http://www.ti.inf.ethz.ch/pw/publications/software/audens |
| PRIME ² | [144] | http://csbl.bmb.uga.edu/downloads/prime/prime.html |
| PEAKS ³ | [145] | http://www.bioinformaticssolutions.com/products/peaks/index.php |
| Sequit ³ | [152] | http://www.sequit.org/ |
| PST Based Methods | | |
| PeptideSearch ^{1,2} | [154] | http://www.narrador.embl-heidelberg.de/GroupPages/PageLink/peptidesearchpage.html |
| ProteinProspector ¹ | [156] | http://prospector.ucsf.edu/ |
| GutenTag ² | [158] | http://fields.scripps.edu/GutenTag/ |
| Similarity Searching Programs | | |
| CIDentity ² | [141,142] | http://ftp.virginia.edu/pub/fasta/CIDentity/?M=A |
| FASTS ^{1,2} | [162] | http://fasta.bioch.virginia.edu/fasta_www/cgi/search_frm.cgi?pgm=fs |
| MS-BLAST ¹ | [164] | http://dove.embl-heidelberg.de/Blast2/msblast.html |
| InSpecT ^{1,2} | [170] | http://peptide.ucsd.edu/ |
| SPIDER ¹ | [172] | http://bif.csd.uwo.ca/spider |
| DEenovoID ¹ | [173] | http://proteomics.mcw.edu/denovoID |
| Other Resources | | |
| SILVER ^{1,2} | [55] | http://llama.med.harvard.edu/Software.html |
| X!HUNTER ¹ | [194] | http://www.thegpm.org/HUNTER |
| DBParser ² | [189] | http://www.proteomecommons.org/archive/1109121060785/ |
| VEMS ² | [188] | http://yass.sdu.dk/ |
| RADARS ³ | [76] | http://65.219.84.5/RADARS.html |

(Table 3) contd.....

| Databases and Organizations | | |
|------------------------------------|-------|---|
| HUPO | - | http://www.hupo.org/ |
| PRIDE | [193] | http://www.ebi.ac.uk/pride/ |
| GPM | [81] | http://www.thegpm.org/ |
| Unimod | [213] | http://www.unimod.org |
| Resid | [214] | http://pir.georgetown.edu/cgi-bin/resid |
| DeltaMass | - | http://www.abrf.org/index.cfm/dm.home |
| Quantitation Resources | | |
| ZoomScan ¹ | [228] | http://proteomics.mcw.edu |
| RelEx | [233] | http://fields.scripps.edu/relex/ |
| MsQuant ¹ | [230] | http://msquant.sourceforge.net/ |
| PTM Resources | | |
| Lu and Chen software ¹ | [98] | http://hto-c.usc.edu:8000/msms/suffix/ |
| P-Mod ¹ | [210] | http://www.mc.vanderbilt.edu/lieblerlab/p-mod.php |
| ProSight PTM ¹ | [215] | https://prosigthptm.scs.uiuc.edu |
| FindPept ¹ | [216] | http://www.expasy.org/tools/findpept.html |
| ProMoST ¹ | [218] | http://proteomics.mcw.edu/promost/index.jsp |
| ProteoMod ¹ | [219] | http://biochem.iisc.ernet.in/proteomod.html |
| GlycoSuiteDB ¹ | [221] | https://tmat.proteomesystems.com/glycosuite/ |
| GlycoFragment ¹ | [222] | http://www.dkfz.de/spec/projekte/fragments/ |
| GlycoSearchMS ¹ | [222] | http://www.dkfz.de/spec/glycosciences.de/sweetdb/ms/ |

Tools mentioned in the text that are either not available or no more traceable as web URL have not been reported.

most intense peaks. It probably includes a limited stochastic model adapted from the MOWSE score together with heuristic and probability-based scoring in order to capture some properties related to signal intensity and consecutive fragment matches.

X!TANDEM [81-83] has been developed as an alternative to MASCOT and SEQUEST. The algorithm compares each spectrum to all likely candidate peptides in a protein database where all fragment ions are assumed to be of equal intensity. One of X!TANDEM's major strengths is its automatic search for modified peptides (i.e. post-translational modifications) on proteins it has otherwise identified. This approach allows a dramatic decrease in computational time compared to the brute force approach of trying all possible combination of peptide modifications. The preliminary score is the dot product of the acquired and modeled spectra where peak intensities are considered and the final score is based on a hypergeometric distribution that is exploited to assess an expectation value of the match. It is open source software and packaged in the SCAFFOLD suite from Proteome Software.

Another freely-available method that implements a similar statistical approach is OMSSA [84], a fast search algorithm whose results are scored using a BLAST-like statistical model based on assumptions taken directly from the experimental setup and allowing for experimental noise. The matching criteria employ a conservative approach; if a measured fragment ion is matched to one fragment then it is no longer considered for matching other ions. The scoring function calculates a Poisson distribution of random matches

for each spectrum, allowing the significance of a hit to be expressed as the probability of the hit being random, where a low probability implies a significant hit. Yet another statistical approach is found in Probit [85], which calculates the statistical significance of each peptide identification and reports the risk that a particular identification is a false positive, mitigating the influence of database size on the final score.

ProBID [86] and PROFOUND [87], on the other hand, use a Bayesian model to calculate the probability that the candidate peptide is a true match. In addition, ProBID separates the scoring function into distinct components that are computationally very efficient, as they only consider *b*- and *y*-ion information. Another novel feature of ProBID is that it includes the ratio of matched consecutive fragment pairs as a component of its probability.

One common assumption in all these SPC algorithms is the concept that every matching peak between the experimental and theoretical spectrum should be given equal weight. This undoubtedly leads to stochastic mismatches and increases the false-positive rates if not rigorously evaluated [37] so pFind [88] has been specifically devised to address this issue by exploiting the correlative information among fragments. A kernel function is applied to the dot product of shared peaks [89] to exhaustively count all possible combinations of correlated ions and reduce the computational complexity. The algorithm, available as a web resource [90], is based on a more rigorous formalization of the correlative property of co-occurring fragments and has demonstrated its effectiveness in terms of identification accuracy.

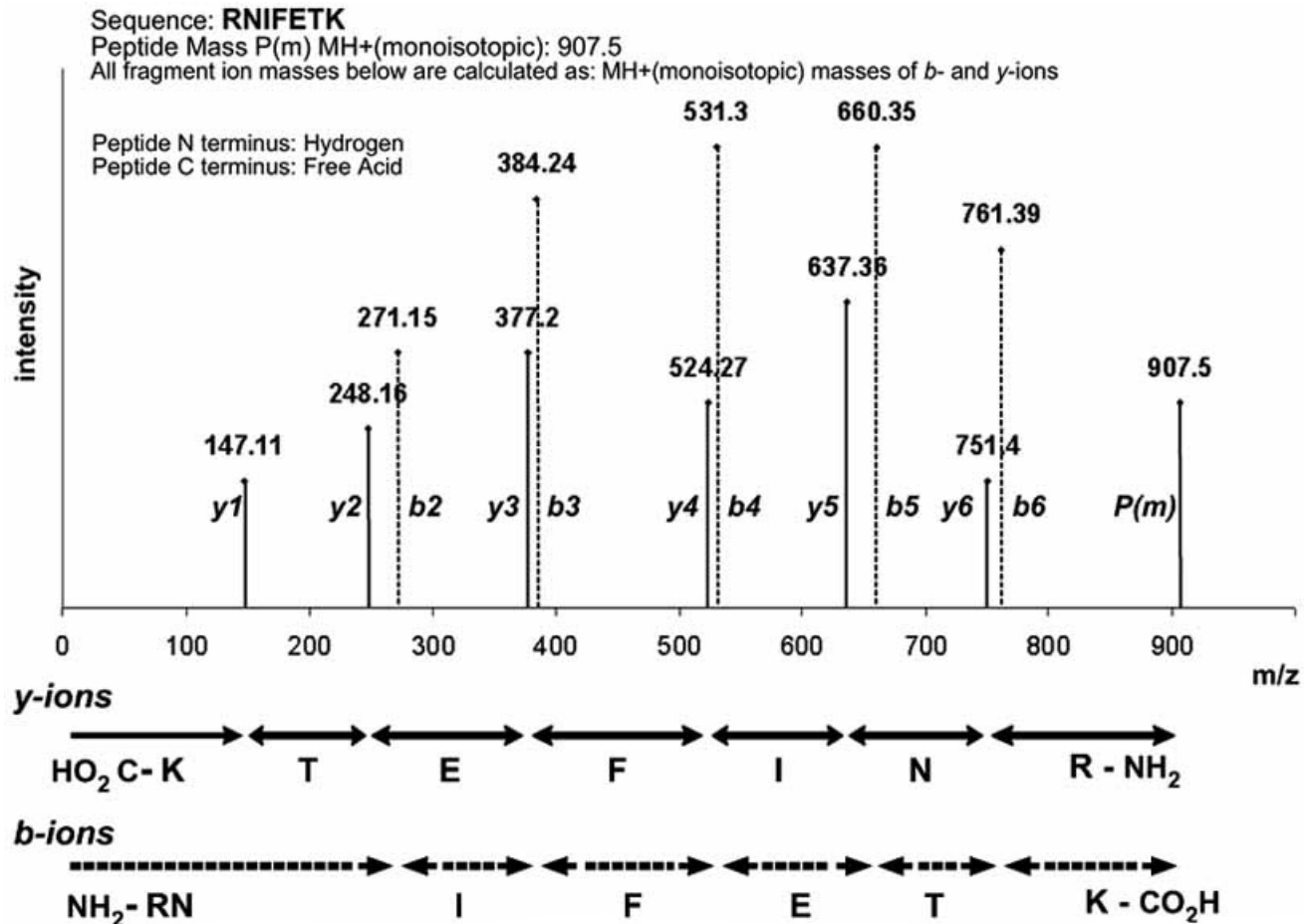


Fig. (5). Theoretical MS/MS spectrum of the sequence RNIFETK. The x-axis reports the mass-to-charge ratio of the fragments and the y-axis represents the intensity of the peaks whose heights do not account for any particular real case but have been chosen to make the example clearer. The singly charged series of the monoisotopic b - and y -ions are shown. The distance of two consecutive peaks in an ion series corresponds to the mass of one single amino acid supposing the ion series to be complete and containing all of the possible fragments derived from the precursor peptide as in the example. The y -ion series is written from the C- to the N-terminus and the reconstructed peptide sequence is the reversed form of the precursor peptide. On the contrary, the b -ion series has the same orientation from the N- to the C-terminus of the fragmented precursor (see text for details).

A group of algorithms based on stochastic approaches have also been developed to estimate probabilities for the peptide fragmentation process. Partial knowledge of the underlying rules governing the fragmentation process and the occurrence of ion types in training sets are common features that present methods have implemented. The general framework is usually based on the calculation of the probability of observing the experimental spectra given a particular fragmentation pattern and its distribution in the experimental training set. Finally a stochastic model distribution is applied to find the best fit between the observed and theoretical tandem mass spectra.

Rising from the ashes of the Aldente tool developed by GeneBio and SIB (Swiss Institute of Bioinformatics), Phenyx is based on the Hough Transform approach for realigning the experimental spectra and deals very efficiently with false-positive identifications produced by background noise. It is available both as a freely accessible web resource and as a commercial standalone platform. It incorporates the true probabilistic scoring system OLAV [91] that is based on

a stochastic approach and signal detection theory. One of the differences between Phenyx and the previously discussed methods is that in Phenyx the score is the sum of 12 basic scores, such as the presence of ion types (b , y , etc), intensities, peptide modifications and co-occurrence of ion series computed by a Hidden Markov Model (HMM). The scoring function computes the maximum likelihood that the identified peptide is a true positive match.

Working along the same lines, SCOPE [92] uses a stochastic model based on a two step strategy: in the first phase a fragmentation model of all the possible peptide patterns is generated and its empirical probability of occurrence is calculated. In the second phase the measurement of the tandem mass spectra derived from the first step is calculated according to the instrument accuracy. Then, the probability of observing a collection of spectral peaks given this model is calculated.

The main limit of stochastic approaches is that their scoring function is learned on a set of validated data and thus the resulting parameters are a function of the instruments and

sample preparation used to prepare the validated data. Therefore, the performance of these methods suffers from the lack of available training sets that cover all possible experimental conditions.

4.1.3. Alternative Approaches to Spectral Interpretation

Several alternative approaches have been described to enhance or complement classical approaches of present SPC methods. Fragmenting peptides is time consuming and the interpretation of the resulting spectra is fraught with difficulties so characterizing a protein sample without fragmenting peptides is an attractive goal. Towards this end Cargile and Stephenson employed accurate mass measurements and peptide isoelectric point (pI) as identification criteria [93]. The intrinsic power of this method is mainly based on its capability to discriminate peptides with the same monoisotopic masses but with differing isoelectric points allowing a more efficient identification process of the peptides. Along a similar vein, Smith *et al.* [94] have championed Accurate Mass and Time tags, using the elution time of reversed phase HPLC and high accuracy mass measurements to identify peptides in much the same way.

While the use of physical properties for peptide identification is attractive, no such method has achieved broad use yet and most groups still use tandem mass spectrometry. Zubarev *et al.* took this one step further by fragmenting each peptide using two orthogonal methods, electron capture dissociation (ECD) and collisionally-activated dissociation (CAD), to generate complementary pairs of fragments [95]. These pairs are exploited both for preliminary screening of tandem mass spectra and for improving the accuracy of peptide detection in database searching. Additionally, since such an experiment must be done in an FT-ICR instrument, ultra-high mass accuracy measurements provides further assurance in the identification of peptides. Such data has also allowed the implementation of a novel sequence tag detection algorithm based on the combined data extracted by CAD and ECD. The scoring system used is database independent since it relies on the maximum length of the peptide sequence tag provided by the complementary pairs of sequence tags from the combined CAD and ECD data [96].

Another interesting aspect of tandem mass spectrometry that contributes to the complexity in this field is that concurrent fragmentation of multiple precursors in the same tandem mass spectrum may occur. This factor, together with incomplete fragmentation in the first place, errors in sequence databases and the incompleteness of current databases, certainly contributes to the low frequency of matching fragment spectra to peptides. ProbIDTree [97] attempts to address some of these issues as it has been designed specifically to find possible multiple peptides from the same fragmentation spectrum. It explores, through an iterative database searching process, putative matches with candidate peptides to which probability scores are assigned. Tentatively matched peptides are organized in a tree structure from which their adjusted probability scores are calculated and used to determine the correct identifications.

4.1.4. Reducing Computation Time in Database Searching

One of the major bottlenecks of tandem mass spectrometry algorithms is certainly the increasing computational time required for performing database searches of large protein

libraries, especially when non-specific enzyme cleavages and several post-translational modifications are allowed. The need of reducing computation time to speed up the database search is a timely issue as protein databases continue to grow, thus several approaches to accelerating searches have been described. One solution has been combining suffix tree data structures to preprocess the protein sequence database and spectrum graphs to preprocess the tandem mass spectrum [98]. This algorithm efficiently searches the suffix tree against the spectrum graph for candidate peptides and ranks the hits accordingly to a SEQUEST-like scoring function. An alternative solution, adopted by DBDigger [99] with a marked decrease in search time, is to reorganize the database identification so that candidate sequences are generated only once for each collection of spectra rather than once for each spectrum. Yet another approach uses a new data structure called "peptide finite state machine" (PFSM) [100] where data from experimental spectra are organized in terms of monoisotopic and centroided spectra which may be used to rapidly search a known set of protein sequences, regardless of the number of spectra searched or the number of potential modifications examined.

The downside of heuristic approaches such as these aimed at accelerating the searching process is that they may degrade the final result. The false positive rate may increase whereas true hits may be lost or, more generally, sensitivity and specificity may suffer. The need to find a reasonable trade-off between more efficient searching algorithms and statistical evaluation of the results has created the basis for both the development of novel methods specifically designed to seriously improve current statistical estimates in protein identification and the assessment of comparative and independent benchmarking experiments able to test the effectiveness of present methods.

4.1.5. Independent Statistical Validation of SPC Results

SPC engines remain popular despite suffering from several common limitations, one of the most obvious being that they always report a hit for a given query despite the quality of the match. For a low quality measured spectrum this can leave the score of the match in the twilight zone where false and true positive hits are equally distributed. In addition, they act at the peptide level whereas the main goal in mass spectrometry-based proteomics is to give a statistical measure of whole protein detection in terms of peptide coverage and probability that a specific protein has been unequivocally detected. To address these issues different statistical models have been applied to improve the discrimination power of the above-mentioned search algorithms. Since it is not yet clear which solution fits best [64,101], we will discuss all published approaches to statistically validate database search results.

Acting at the level of peptide statistical significance there are PEP_PROB [102] and PeptideProphet [103,104]. The first approach implements a hypergeometric model that has been demonstrated to fit the frequency distribution of fragment ion matches and predicts a probability of generating any given number of fragment ion matches to an experimental tandem mass spectrum at random almost independently by database size. PeptideProphet, on the other hand, uses an expectation maximization algorithm to derive a mixture model of correct and incorrect peptide assignments by fitting

the derived curve of the calculated and empirical discriminant scores for all the spectra in the sample. Bayesian statistics are then applied to compute the probability that the match is correct (Fig. 6).

Their counterparts at the protein level identification, PROT_PROB [105] and ProteinProphet [106], have been developed to assess the protein detection probability by assembling peptides into proteins. The real goal of a high-throughput proteomics approach is to determine the identity of the proteins present in the original sample and the protein sequence coverage (the fraction of the entire protein sequence observed as peptides in the mass spectrometer) is a major factor when validating search results. The connectivity between peptides and precursor proteins is difficult to recreate when a complex protein mixture is analyzed, consequently assembling peptides into proteins is not straightforward and depends on different factors as the presence of degenerate, modified and repeated peptides (Fig. 6).

PROT_PROB implements two independent statistical models based on binomial and multinomial hypotheses, which use the hypergeometric probabilities and cross-correlation scores from SEQUEST respectively. Weaknesses of this approach manifest in cases where protein assignments are based on unusual and incomplete peptide fragments. ProteinProphet starts from the probabilities assigned by PeptideProphet to the identified peptides and then uses these values in a Bayesian model that is based on the concept of "number of sibling peptides" (NSP) for peptide matches to a

protein that has other matching peptides. An NSP is high when numerous peptides match the same protein.

More widely-used approaches for assembling peptide identifications into protein include DTASelect [107], CHOMPER [108] and INTERACT [109], all of which use empirically determined scoring threshold values to filter putative correct hits and then sort peptides by protein identity. The intrinsic limit of this approach resides in the empirical nature of the subjective choice of the threshold scores for correct identification that may change depending on the database size and the data set [110]. Despite lacking a probability-based measure, the ability of these methods to compare peptide identifications from multiple experiments and their filtering capabilities provide a certain level of confidence for the user. In addition, the popularity of MASCOT and SEQUEST means that many resources have been devoted to provide training and instructions for these analysis tools.

As was mentioned, the Xcorr value from SEQUEST is not a statistical value and shows bias towards longer peptides, large databases and noisy spectra. In order to decouple the peptide and database size, two groups have tried to normalize Xcorr scores by dividing the calculated value by the Xcorr of the input spectrum against itself [111,112]. This approach has also been recently applied to assess peptide identification based on their length in large scale proteomics projects [113]. Other statistical tools that have employed different probabilistic models considering numerous features of the resulting matches [114-121] have proven to give similar

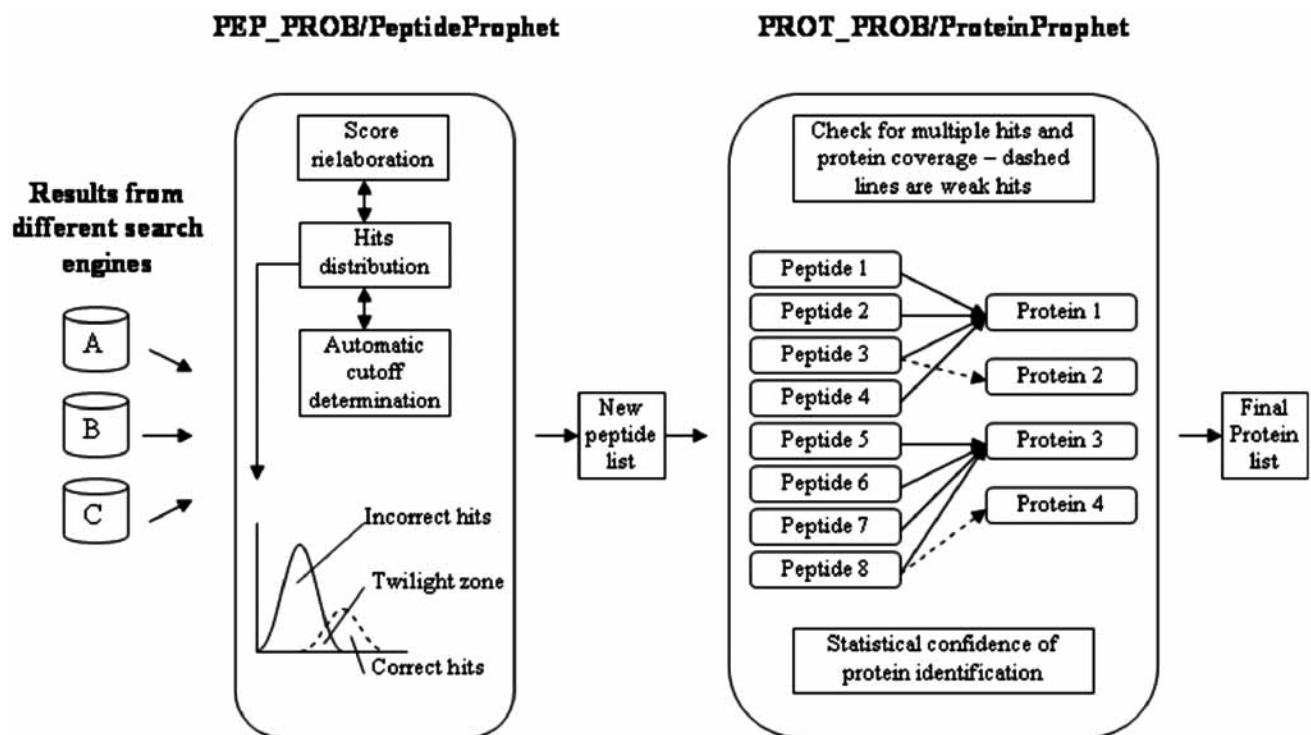


Fig. (6). A schematic and hypothetical working pipeline of PEP_PROB/PeptideProphet and PROT_PROB/ProteinProphet. Results from different search engines are elaborated from PEP_PROB/PeptideProphet to estimate a dynamic threshold able to better discriminate incorrect from correct peptide hits based on the characteristic of the data and their distribution. Once grouping is complete, the assigned peptides corresponding to an individual protein, and their probabilities, are combined by PROT_PROB/ProteinProphet that compute a single protein confidence measure which is effective at distinguishing the correct from incorrect protein identifications. These tools especially aim at increasing statistical significance of low scoring hits.

results. Because of this, several machine learning methods have been designed for post-database search validation of peptides ranging from neural networks [122,123] to support vector machine SVM [124] and Quadratic discriminant functions [125]. Unfortunately, one major caveat of machine learning approaches is that they may not produce accurate results when applied to data significantly different from that used for training.

Unlike SEQUEST, MASCOT uses a probability-based scoring scheme for peptide identification. However, it does not correlate these scores to determine the significance of the proteins to which they match so various statistical inference methods have been also applied to MASCOT outputs [126-128] to improve both sensitivity and specificity. The heuristic PROVALT [129] has been specifically tailored for MASCOT to overcome this common limitation of most peptide identification algorithms. A second heuristic approach, MATH [130] based on Bayes' Law, has been successfully applied to MASCOT by using mass tolerance settings and database size to improve sensitivity and specificity.

The whole overview of the different statistical representations of scoring functions suggests the need for evaluation methods to assess low scoring hits obtained from orthogonal approaches [131], which in turn requires a normalized score for comparing among different methods. The expectation value for protein identification has been suggested as a good candidate for this metric, considering only peptides that pass a preset threshold value for a survival function. The stochastic occurrence of peptides is governed by the Poisson distribution and may be used for the qualitative analysis of the differences between scoring systems [132]. The consensus approach has been employed to some success in shotgun proteomic projects using SEQUEST and MASCOTs, along with chemical properties and scores describing the nature of the fragmentation spectrum, to enhance the specificity and sensitivity of protein identification [133].

4.1.6. Benchmarking and the Importance of Using Rigorous Limits

With the advent of faster mass spectrometers and higher dimensional chromatography, proteomics seems to have become a numbers game – more proteins reported as identified equals a better study. While such studies MAY have taken more effort, they are less likely to make a significant contribution to science than more careful and focused studies. As with any developing field, proteomics has been slow to adopt global standards for things such as what constitutes an identified protein in a large-scale study [134]. The tools that protein identification hinges on are the database search engines, yet with the multitude of engines discussed here and new engines being introduced all the time there is still no comprehensive benchmarking of the results each one produces. One challenge in doing a comprehensive benchmarking is coming up with a gold standard that will be acceptable to the majority in the field. The obvious solution to this, at least for a set of 'good' spectra, is to take synthetic peptides and record their fragment spectra in different mass spectrometers, knowing without question which peptide gave rise to which fragment spectra, and then to test each search engine with this dataset. Equally obvious though is the time and financial cost involved in synthesizing and individually fragmenting several thousand peptides. However, some benchmarking of

current tools has been performed, within the limits imposed by the variety of the methods [131,135-137]. A substantial variability (RSD 4-25%) was documented among database scores provided by several searching algorithms by submitting replicates of tandem mass spectra [137]. There, post-search average of XCorr scores (SEQUEST) led to better results than spectral averaging. Other tests suggest that MASCOT, X!TANDEM, and SONAR perform significantly better in terms of specificity than SEQUEST [131,136], except in the case of low resolution ion trap data where MASCOT and SEQUEST performed similarly [136].

In the absence of rigorous standards and benchmarking, the onus then falls on the investigators to ensure the quality of their data. However, evidence suggests that many, if not a majority, of groups are using criteria that allow a very high number of false positive identifications to enter the scientific record. Stephenson *et al.* [37] used a set of MS/MS spectra collected from a rat sample to search a completely random database of proteins and even using the 'rigorous' criteria used by many investigators almost fourteen hundred proteins were identified, all of which must have been wrong. Such problems are compounded when investigators expand the search space by using wide mass windows or by allowing non-specific enzyme cleavage. These issues were addressed by Mann *et al.* [138], who collected a set of MS/MS spectra on a linear trapping quadrupole-Fourier transform hybrid mass spectrometer and searched them against a database using a variety of search parameters. The conclusion that these investigators arrived at was that trypsin cleavage is indeed specific and that the vast majority of non-tryptic peptides identified using low mass accuracy restrictions were incorrect.

One independent method for evaluating the quality of protein identifications in a large-scale dataset, regardless of the criteria or instrument used, is called reversed database searching [71,139]. This simply means that in addition to querying a set of fragment spectra against the desired sequence database, the spectra are also used to query the same database that has been reversed or randomized. The concept of reversed database searching is that any match between a spectra and a reversed peptide must be a false positive identification, with the exception being those peptides that are found in both the forward and reversed databases. The ratio of identifications in the reversed database to the forward database is an estimate of the false positive rate for the criteria used. By evaluating false positive rates under different search conditions (e.g. by adjusting the peptide mass accuracy window), a set of parameters that gives an acceptable rate of incorrect identifications can be empirically determined. Optimally this rate should be chosen to be less than the number of proteins identified (i.e., if 1000 proteins are identified then the false positive rate should be less than 1 in 1000).

4.2. De Novo Sequencing Methods

Unlike the methods already discussed, which rely on knowing all the possible protein sequences in order to interpret spectra (a.k.a., database-dependent), *de novo* sequencing methods rely exclusively on deriving the peptide sequence from a tandem mass spectrum using only the knowledge of the phenomena governing the peptide fragmentation process (database-independent, Fig. 4). These approaches are gener-

ally error prone and produce ambiguous results due to different problems inherent to the intrinsic nature of proteins and fragmentation techniques. Indeed, these caveats are common to every analysis regardless of the approach used, be it database-dependent or database-independent, but their impact on *de novo* sequencing is more critical since no prior knowledge of sequence is available to help discriminate between false and positive hits.

The first and most obvious problem is in trying to distinguish between amino acids with identical or nearly identical masses, as in the case of the isobaric leucine/isoleucine and glutamine/lysine respectively. In addition, certain amino acids have the same mass as pairs of other amino acids. Other major hurdles can be poor quality spectra where some expected fragment ions are missing, the directionality of the ion series (*b*- and *y*-ions) and charge states of the fragments. Despite these drawbacks, *de novo* sequencing can be extremely valuable since it is database-independent and can be used with proteins from any organism, sequenced or not.

Present *de novo* methods may be divided into two distinct categories depending on which solution is chosen to represent the mass spectra. The first class includes methods adopting graph representation of tandem mass spectrum, whereas the second embraces different solutions generally described as combinatorial problems.

4.2.1. *De Novo Sequencing – Graph Theory Approaches*

SHERENGA [59] is among the progenitors of *de novo* sequencing methods in this first category. The algorithm is based on peptide sequence reconstruction through optimal path scoring in the 'directed acyclic graph' representations of the tandem mass spectra. It automatically learns fragment ion types and intensity thresholds from a collection of test spectra generated from any type of mass spectrometer. The training set allows the algorithm to produce a likelihood-based scoring list of ranked paths corresponding to potential peptide sequences. SHERENGA is most useful for interpreting sequences of peptides resulting from unknown proteins and for validating database results.

A similar statistical model has been implemented and improved in the PepNovo [140] algorithm. It tests the likelihood that the observed peaks may be obtained from either known peptide fragmentation rules or a random fragmentation process. The algorithm steps are designed to solve the spectrum graph using a probabilistic network to reduce the computational space. This network takes into account correlations between fragment ions, dependencies due to the relative position of the cleavage site in the peptide and influence of flanking amino acids to the cleavage site. Then, solutions are ranked accordingly to the random distribution hypothesis.

Lutefisk [141,142] is another tool that is based on sequence graph theory. In the first phase the raw data undergo a smoothing process to extract a weighted profile. The data are subsequently converted into a graph of their corresponding ion masses. Finally, the graph is resolved following the best path in the graph to gain the longest partial sequence and allowing for gaps that compensate incomplete fragmentation patterns. The obtained sequences are searched for in protein databanks with an error-tolerant similarity searching algorithm CIDentify [141,142] that will be described later.

One of the first methods to employ a heuristic component was AuDeNS [143]. The user may interact with this tool by assigning relevance values to the input peaks and thresholds to preprocess the spectrum and find solutions along the graph path. The problem of the potentially exponential number of solutions is then solved since only those within a user-specified threshold are tested and reported.

One central problem in interpreting tandem mass spectra is the uncertainty in correctly assigning the various ions of different types. To address this problem, PRIME [144] treats *b*- and *y*-ion series as type-1 and type-2 edges of the graph respectively and then a dynamic programming algorithm tries to rigorously solve this graph partition problem. The program can be used to pre-screen acquired spectra and separate the different types of ions to improve *de novo* sequencing accuracy.

4.2.2. *De Novo Sequencing – Combinatorial Method*

PEAKS [145], one of the first and most widely used algorithms in this second category, is based on a four step strategy. The first stage is dedicated to the preprocessing of the data by noise filtering and peak centroiding. In the second step an almost exhaustive candidate search is computed and the best 10000 hits retained. The solution space explored consists of all possible combinations of amino acids for a given precursor ion. The basic assumption of the scoring system is based on an empirical measure and starts from the observation that the greater number of high abundance peaks that are matched by those ions of a sequence, the more likely the predicted sequence is the correct sequence. The last two steps are different rounds of scoring assessment and normalization.

A similar approach has been recently proposed in the RAID [146] program that is divided into two distinct modules. The first is a pure *de novo* sequencing method that uses a similar work-flow as PEAKS to obtain a list of candidates that are clustered and ranked on the basis of their fit to the experimental fragmentation and intensity pattern. The second module is based on an exact string match between the top candidates against a pre-calculated peptide library extracted from the protein databank. If a candidate turns out to be a substring of some library peptides, the exact match is reported. On the other hand, if no significant library hit is found the top ranking candidates from the *de novo* sequencing results become potential candidates for new peptides that are not yet in the database.

Another new strategy has been described for the determination of amino acid sequences of unknown peptides based on a two-step process that relies on highly accurate mass determination [147]. In the first step the amino acid composition of the peptide target and a small number of accurate fragment ion mass values are determined. The second phase is based on simple permutation and calculation of all possible amino acid sequences that are scored accordingly to the correspondence between expected and observed fragment ion signals of the permuted sequences. The efficiency process is warranted only if the peptide composition is reduced to a small list and this is possible only in the case of ultrahigh mass accuracy measurements.

To overcome the limitations of *de novo* sequencing methods used for interpreting low resolution instrument results,

a method has been designed that takes advantage of the reproducibility and predictability in fragment ion intensities of less expensive and more widely accessible ion trap instruments [148]. In the first step, a divide-and-conquer algorithm generates a collection of sequence candidates from the experimental tandem mass spectrum [56] and in the second step these candidates are further refined by comparing their simulated tandem mass spectra to the experimental spectrum. The main shortcoming of this approach is due to its empirical construction of the theoretical spectrum that is instrument dependent.

Other methods using different solutions include formulation of mathematical models [149], the resolution of suboptimal paths in a matrix spectrum graph [150], a novel concept of two-dimensional fragment correlation [151], specific instrument application [152] and implementation of genetic algorithms that construct peptide sequences matching the target spectrum optimally [153].

Most of these methods are coupled with error-tolerant sequence similarity searches [154] following the example of the hybrid methods, discussed later on. *De novo* peptide sequencing and PST generation, though related, are historically distinct problems. The main difference is the “global” approach of *de novo* methods versus the “local” approach of PST methods for interpreting the tandem mass spectrum [155]. In *de novo* sequencing the main goal is roughly to find a single and representative “global” path that is constituted by an ordered list of sequence tags extracted by the tandem mass spectrum accordingly to a probabilistic model. Then, the number of incorrect sequence tags due to random local matches is limited. On the contrary, the PST generation usually appears to be a special case of *de novo* sequencing. In this case, the minimal requisite is to satisfy the “covering” principle. Tags are generally local and independent solutions of resolved portions of the tandem mass spectrum that do not necessarily belong to or fit an optimal global path. Random tag identifications of local ambiguities may increase the false positive rate if not correctly ranked. In recent years, this distinction has disappeared and PST methods have improved their tag generation by taking advantage of recent advances in *de novo* sequencing algorithms coupled with improved sequence similarity searches as described below.

4.3. Hybrid Methods: The Peptide Sequence Tag approach (PST) Using *De Novo* Tag Sequencing and Database Similarity Searching

The principle of peptide sequence tags was actually the first method developed for identifying peptides from tandem mass spectra. The starting assumption of this method is that for each spectrum it is possible to determine at least a clear and consecutive short ion series of either *b*-ions or *y*-ions whose peak distance, in the mass-to-charge ratio scale, corresponds to the mass of one amino acid (Fig. 4). One of the first implementations of this principle was PeptideSearch [154] and the first step of the algorithm really depends on a sort of *de-novo* sequencing limited to a small and easy readable portion of the spectrum. The small amino acid sequence obtained this way is called peptide sequence tag (PST) and it is used in a protein database search together with several other pieces of information: flanking regions of the tag, whose sequences have not been detected, are exploited in the

database search since their masses are known and correspond to the N- and C-terminus of the whole peptide, the mass of the precursor ion, and information about the terminal amino acids derived from the endopeptidase specificity. In the case of the sequence tag belonging to the *b*-series, the N-terminus mass (M_1) corresponds to the mass-to charge ratio of the smallest ion in the series of the partial sequence tag whereas the C-terminus mass (M_3) corresponds to the difference between the precursor peptide mass (P_m) and the largest ion in the series of the partial sequence tag (M_2). For example, if the recognized peptide tag is FET, as shown in Fig. 4, M_1 mass-to-charge ratio is 384.2 and M_3 is $(P_m - M_2) \rightarrow (907.5 - 761.39) = 146.11$ m/z. No *a priori* assumption about the directionality of the sequence tag, either belonging to the *b*- or *y*-ion series, is made so that both database searches are performed by the algorithm. The sequence match is scored based on a random probability match of each assigned region (the N-, C-terminus masses and the sequence of the tag) and an expectation probability associated to protease cleavage specificity. For example amino acid probability is set to 1/20 whereas isobaric isoleucine and leucine and almost weight equivalent glutamic acid and lysine are set to 1/10 since they appear two times more frequently than other amino acids due to their weight equivalence within a specified mass accuracy. Total random probability is set:

$$P_{\text{random}} = P_{\text{NtermCleavageSpecificity}} \times P_{M_1} \times \prod_{i=1}^{L_{\text{TagLength}}} P_i \times P_{M_3} \times P_{\text{CtermCleavageSpecificity}}$$

and the resulting nonrandom probability is set to:

$$P_{\text{nonrandom}} = (1 - 2 \times P_{\text{random}})^N$$

where N is the number of amino acids in the database and the random probability is multiplied by 2 since the orientation of the sequence tag is not known. The method proved to have some intrinsic limitations due to the probability model used. For example, amino acids occur at vastly different frequencies in a database. This observation is not taken into account by the algorithm as probabilities are set to 1/20 for each amino acid. Probabilities assigned to the cleavage specificity and the molecular weights of the M_1 , M_3 regions are arbitrary. A second limit concerns the correct sequence tag recognition that may fail in the case of incomplete spectra where not all fragmentation patterns are available. Another limit is the presence of post-translational modifications or isoforms in the sequence tag that are inevitably not identified during the searching process if this information is missing in the database. The algorithm tries to overcome this problem by matching at least two out of the three extracted regions (N-terminus mass, sequence tag, C-terminus mass) and evaluating putative alterations in the remaining unmatched region.

Another implementation of PeptideSearch algorithm is available in the ProteinProspector package [156,157] but other methods that employ a hybrid approach have also been developed. One of the aspects usually not considered, because of the lack of a good theory, is the problem of interpreting peak intensities, as described above, even though attempts have been made to exploit this seemingly valuable source of information. GutenTag [158] uses an empirically derived model of fragment ion peak intensities to achieve a higher specificity in sequence tag detection. It retains multi-

ple sequence tags from a tandem mass spectrum that are scored accordingly to the empirical knowledge of ion fragmentation and the best are efficiently searched against the protein database in a single pass. The same ion fragmentation model is applied to estimate the candidate sequence peak intensities. The final score is calculated using a normalized dot-product algorithm that compares the theoretical and observed peak intensities of respectively the candidate sequences and the tags. Multiple tags that match the same complete peptide sequence contribute to increase the final score. This sequence tag implementation is useful for identifying peptides with unknown post-translational modifications or sequence variations.

Popitam [159] employs a different approach that exploits a graph structure in association with database searching. The real innovation of this method lies in the identification step of the tags extracted from the graph. While identification methods based on tag searching typically try to extract tags and then use them to find peptides in the database, Popitam utilizes the database to direct the search and to detect relevant sections in the graph from which the peptides can be scored. The first step of the algorithm consists in transforming the source MS/MS peaks into potential singly charged *b*-ion fragments that are structured in a graph in the same way most of the *de novo* sequencing algorithms do. The graph is then compared with theoretical sequences extracted from the database leading to a similarity score for each extracted peptide sequence. In Popitam, the aim is to find all possible sequence tags from the database that fit in the graph. The final score of the extracted peptide from the database reflects a sort of "fitness" of the tag in the graph path. The combinatorial problem of the optimal path in the graph has been partially solved by implementing the Ant Colony Optimization (ACO) metaheuristics. Some apparent limits have arisen in the efficient identification of incomplete ion series lacking one or more ion fragments but these have been partially solved by shifting from a Full Path Algorithm approach to a Tag Algorithm which allows extracting weighted sections from the graph rather than entire paths. This method has been designed to meet the still unanswered needs of present PST limits enhancing the potentiality of algorithms to detect post-translational modifications and protein isoforms.

The subtle boundary between pure *de novo* sequencing methods and PST approaches is likely to begin to blur and will eventually disappear. A new era of efficient tag extraction techniques borrowed from *de novo* sequencing approaches should stimulate this field. Present SPC computational costs will also come down, particularly for the task of post-translational modifications detection in the face of huge sequence databases. In addition, these approaches generally have to overcome the problem of assessing the false positive rate detection. On the contrary simple PST-based approaches may still miss correct hits, albeit with generally low false positive rates, due to their limited capacity to generate a sufficient number of ranked tags to search for. Pure *de novo* sequencing methods have to overcome the interpretation drawbacks of low quality spectra, unusual fragmentation and ionization patterns that lead to incorrect reconstruction of the complete peptide sequence. For these reasons the interest has been shifting towards developing novel algorithms that may benefit from the strong points of each technique, such as in Popitam and GutenTag. The most promising approach seems

to combine different algorithms developed in different areas. A more efficient integration of *de novo* sequencing techniques for peptide sequence tag identification coupled with efficient sequence database searching algorithms based on SPC approaches could be a winning strategy to enhance tandem mass spectra analysis and peptide identification.

4.3.1. Similarity Searching

The first step of PST methods is to detect clear sequence tags that generally do not cover the entire tandem mass spectrum from where they are extracted. To complement these fragmented pieces of information, these small tags are used in similarity searches against a protein databank to find putative candidate peptides that may fit the complete spectrum. These techniques have been extended and applied even to *de novo* sequencing methods used in hybrid approaches with the intended goal to assess the final rank of putative identified peptides as discussed later on.

MultiTag [160] is a sequence similarity search algorithm specifically designed to help the characterization of proteins from organisms with unsequenced genomes allowing for matches of multiple error-tolerant sequence tags that may be very short, between two and four amino acids in length. Previous methods have been based on modified FASTA [161] (FASTS [162] and CIDentify [141]) and BLAST [163] (MS-BLAST [164]) algorithms. These methods can deal with the peculiarities and ambiguities of sequences obtained from MS/MS data as isobaric amino acids (leucine / isoleucine and glutamine / lysine) and amino acids whose mass is the same mass as certain dipeptides (glycine + glycine = asparagine). Modified substitution matrices, the use of undefined amino acid symbols and the chance of performing searches against nucleic acids database that are *in silico* translated into all six reading frames are part of the common features of these tools. FASTS searches with peptide sequences of unknown order, evaluating all possible arrangements of the peptides, while CIDentify allows for multiple query sequences whose scores are summed at the end of the process. A further modification to the original algorithm is the use of the *k*-tuple [165] set to one because of the short query sequences used in the search. In addition, the algorithm does not allow for gaps in the match to avoid spurious hits. MS-BLAST is more efficient in large database searching thanks to the original algorithm implementation of BLAST and has proved to perform well in proteome-wide identification of unknown proteins [166]. MultiTag, compared with BLAST and FASTA based methods, is less generic and is an example of a specifically designed tool for exact sequence matches of very short peptides, which are represented by an N-terminal mass, usually three amino acids, and a C-terminal mass. It employs a robust statistical evaluation of true positive hits based on expectation values of the multiple error-tolerant matches and recent benchmarking has demonstrated that the algorithm performs very well in nucleotide database searches [167].

Another extension of the classical similarity searching tools has been proposed in OpenSea [168], using what the investigators called a "mass-based" algorithm to align ambiguous *de novo* sequences to protein database entries. The algorithm identifies, in the first step, a list of clear sequence tags, typically generated by a *de novo* sequencing program, that are not broken by ambiguous mass regions. These tags

are used as queries in a string database search. Candidate hits are then subjected to mass-based alignments so that amino acids encompassing the short tag match in both the query and database sequences are converted into their corresponding masses. A series of consecutive local alignments on either side of the tag match are made to form a complete alignment using a "breadth-first search" algorithm. All the combinations of amino acids of candidate sequences are searched within a certain mass tolerance to fit the observed mass intervals in the query mass spectrum. The search continues up to two database residues versus two query residues and the final score is a linear combination of the summed peptide alignment scores. Mass-based alignment of *de novo* sequences is a further step towards a more accurate identification of sequence variations and post-translational protein modifications [169].

As with SPC methods, substantial efforts have also been directed towards reducing the computational cost of sequence tag searches. One particularly successful approach has been to exploit a robust filtration technique to rapidly eliminate candidate sequences in a database search while retaining the true positive hits [155]. The direct consequence of this approach is a dramatic decrease in computational time without significantly affecting sensitivity and specificity. The method makes use of two modified versions of the PepNovo algorithm for *de-novo* sequencing (see 4.2.1 *De novo* sequencing – graph theory approaches) to generate the tags from the spectrum as accurately as possible: PepNovoTag extracts all substrings from the PepNovo reconstruction results and LocalTag searches all sub-paths of the spectrum graph extracted from PepNovo. The final scores of the extracted tags are assigned using a logistic regression model. The database search is performed by InSpecT [170], a novel database searching algorithm that implements the concept of filtering the database by selecting a small fraction of putative true positive hits from the rest. In its present incarnation, InSpecT uses a Tag generation model that is a modified version of *de-novo* spectrum interpretation based on a direct acyclic graph that retains up to 50-100 best tags of length 3 for the subsequent database search. It was originally designed to improve large database searches for possibly modified peptides by implementing an ordered tree data structure called 'trie' to speed up computational analysis. The candidate peptides are ranked according to their relative likelihoods of generating the mass spectrum. The final score is an optimal linear combination of four components as follows: candidate scores (S), explained intensity of a candidate (I) that is the fraction of total ion current belonging to annotated spectral peaks, explained peaks (P) that is the fraction of the peaks that are annotated and b/y ion score (B) that is the fraction of b and y ions found in the spectrum $Q_{final_score} = w_1S + w_2I + w_3P + w_4B$.

The weights w_i have been tuned on a training set to discriminate correct matches from incorrect matches. Thanks to the modular nature of the method, future developments in tagging, filtering and scoring will eventually solve some of the weaknesses of this tool.

The tag-based approaches coupled with efficient tag generation techniques based on *de novo* sequencing methods and efficient database searching is certain to be one of the main future directions for tandem mass spectra analysis. However,

de novo sequencing very often gives only partially correct tags and new efforts have been addressed to define a strategy for this type of homology-tolerant database searches [171], as in SPIDER [172], that is able to efficiently match sequence tags with errors to database sequences. A different approach has been chosen for DeNovoID [173], a web tool tailored towards improving database searches of *de novo* sequencing results. It is based on a new paradigm, relative to common database searching algorithms, as it depends solely on the amino acid composition of the peptide and not its sequence [174]. The program converts and maps the composition of each peptide to a vector, allowing peptides to be compared using vector algebra. The match between a query and a database peptide is measured as the Euclidean distance between the vectors representing the query and database peptides. Many of the problems associated with standard sequence search methods are avoided using composition instead of sequence, as well as an algorithm that is peptide length independent.

The key feature of these hybrid approaches is that they have the potential to be of enormous use for unannotated isoforms, post-translational modifications and cross species homologues that are usually missed by automatic tools and generally need greater computational power. In small-scale proteomic projects manual intervention of an expert is the only feasible solution to assess the results but further efforts need to be addressed to make effective automatic detection of ambiguous cases even for large-scale proteomic projects.

5. HANDLING LARGE DATA SETS: STANDARDS, SPECTRA PRE-PROCESSING AND CLUSTERING

With the advent of large-scale proteomic projects, novel technical hurdles have emerged and the developers' community has rapidly focused their efforts on both the management of large data sets following acknowledged standards and the pre-processing of spectra to reduce the computational time of protein identification.

In the general work-flow, the precursor (MS) and tandem (MS/MS) mass spectra data are acquired and stored by the mass spectrometer as raw binary files. Usually, mass spectrometers vendors provide software to process raw files (*e.g.* smoothing and centroiding of peaks) and to convert them into text file formats (.dta, .pkl, .mgf) consisting of a list of precursor m/z values and charge followed by m/z values of the MS/MS fragments and their intensity.

After data acquisition, the mass spectra processing consists basically of the following steps: 1) spectrum denoising 2) peak deconvolution 3) precursor ion charge state recognition 4) fragment ion assignment. These processed files are the fundamental units of database searches, and some labs develop in-house software to handle them. Therefore, in published works the extent of data processing and adopted criteria are sometimes unclear, in some cases because they are unknown to the researcher due to the use of proprietary software. The definition of a common set of guidelines for data evaluation and processing is still in its infancy, but the huge amount of data being produced makes these standards crucial, as highlighted also during the last international Human Proteome Organization (HUPO) congress (Munich, 2005) [175].

The HUPO Proteomics Standards Initiative (PSI) has devoted considerable effort towards community standards for data representation (<http://psidev.sourceforge.net/>) [176, 177]. In particular, PSI-MS is focused on mass spectrometry and suggests mass spectrometry data formats to make data exchangeable among different research teams. PSI-MS supports the mzData standard (developed in XML) for capturing MS peak lists acquired in different formats. A list of MS companies and organizations currently working on the implementation of mzData to convert their raw data files can be found in the above-mentioned web site. In the same context, Gårdén *et al.* are maintaining an open resource (*PROTEIOS*) at Lund University (Sweden) for managing proteomics data obtained from different sources [178]. Their application supports the mzData format and is freely available at <http://www.proteios.org>.

Recently, the mzXML format has been developed to provide a common data interface between mass spectrometers and data analysis pipelines [179]. The “MS native output” of different mass spectrometers can be converted to mzXML format to facilitate the most common proteomics applications, including database searching, *de novo* sequencing and quantification. As the XML format does not support binary data, the mass spectrometry acquired *m/z* and intensity pairs are encoded in mzXML in base64. An open-source full version and the description of the mzXML schema can be found on the project homepage <http://sashimi.sourceforge.net>. The mzXML format will provide a common language for different mass spectrometers and most importantly will enable the exchange and the objective comparison of data acquired by independent laboratories. Very recently HUPO PSI has decided to fold the mzData format into the more comprehensive mzXML.

There are several other valuable and freely available web tools for multiple protein sequence annotation that deserve to be mentioned because of their utility in retrieving information for multiple protein sequences. Information such as annotation (from different databases), structure, function etc. can be retrieved by means of the integrated database *MPSS* [180]. Related resources are *DASTY* - a DAS client (Distributed Annotation System) for the visualization of protein feature sequence information (<http://www.ebi.ac.uk/das-srv/uniprot/dasty/>) - and UniProt DAS which acts both as a reference and as an annotation server [181].

The algorithm *Fulspec* [182] takes into account the chromatographic properties of detected analytes prior to MS/MS fragmentation in a flexible manner. Basically, the precursor ion is selected based on real-time monitoring of the chromatographic separation. Exclusion of broad, tailing peaks and putative low-quality mass spectra precursors is thereby achieved. However, as is the case of other algorithms, extensive validation and incorporation into instrument software has yet to be implemented. A statistical approach has also been proposed for the systematic evaluation of MS/MS mass spectra to be used off-line to address instrumental settings improvement [119].

From a naïve point of view one might be surprised by the small percentage of acquired MS/MS spectra that are eventually used for database identification. The reasons for this inefficiency are numerous but include non-specific cleavages, post-translational modifications and other not yet clari-

fied events, in addition to spectra generated from non-peptide ions (i.e., 'junk'). This means that a lot of information is lost but it also represents a major opportunity for reducing the size of the MS/MS dataset prior to the actual database search. Burlingame *et al.* [183] examined 3269 collision-induced dissociation spectra acquired on a quadrupole time-of-flight hybrid mass spectrometer and concluded that 72% of spectra from this type of instrument were identifiable through various means. While a Herculean effort and very informative, this manual approach is obviously not generally applicable to high-throughput experiments. Yates *et al.* [184] used Quadratic Discriminant Analysis and Support Vector Machines to filter out bad spectra, achieving an impressive 75% reduction in bad spectra in exchange for only a 10% reduction in good spectra.

Another approach for reducing the amount of data to facilitating database searches and data validation involves the grouping of raw data into a lesser number of representative clusters. *Pep-Miner* was designed to cluster MS/MS spectra from multiple LC-MS/MS runs [185] by assigning spectra to a cluster based on the mass difference between peaks. Moreover, only the highest peaks in each spectrum were considered and similarity scores were calculated. The major limitation was due to the fact that different charge states of the same peptide ended up in different clusters because of the non-similarity of the MS/MS spectra. *Pep-Miner* can perform correlation and prediction of LC retention times throughout different LC runs in order to improve peptide identification (particularly useful when the fragmentation quality is poor). Retention time prediction is based on amino acid hydrophobicity and learns from a set of training peptides. The algorithm has been tested on the proteome of lung cancer cells and on cancer MHC peptides.

Reducing computational tasks such as database searching will no doubt reap benefits by increasing efficiency but several issues still remain. In the analytical pipeline the bottleneck is at the level of data acquisition in the mass spectrometer, but in the bigger scheme of things the biological workup and the functional validation (when it is done) take far more time.

6. THE PITFALLS OF PROTEIN IDENTIFICATION METHODS

While automated methods for identifying proteins from mass spectra are far more developed than for lipids or polysaccharides, they still have a long way to go before they are as robust as DNA sequencing. As mentioned, a large fraction of acquired fragment spectra do not identify any peptides and while a majority of these un-used spectra are simply of poor quality, many appear to be interpretable. Of course, for PST and SPC approaches the peptide that gave rise to a spectrum must be in the database being searched in order for it to be identified. Even for organisms whose genomes have been sequenced this is still a problem but annotation groups around the world are constantly updating the sequence repositories with newer gene predictions so presumably this problem will be overcome. For species where incomplete sequence information is available one solution has been proposed to efficiently identify cross-species proteins in an integrated approach making use of Expressed Sequence Tags (EST) [186]. Nevertheless, protein databases are growing

exponentially thanks to whole genome sequencing projects, mitigating the shortcomings of missing information in many cases.

Even with fully sequenced genomes though, much information about polymorphisms and splicing variants remains undocumented [187]. This can cause difficulties in peptide detection since error-tolerant matches are still an open issue and their statistical significance difficult to assess correctly. One solution might be to query a collection of several species at once, although specificity would then be diminished. A large, comprehensive and multi-species database, by presenting a much larger search space, would increase the false positive rate since one spectrum is more likely to match more than one entry. The degree of ambiguity is strictly dependent on the database size and the number of tandem mass spectra used. The combinatorial explosion of the search space becomes even more problematic when considering post-translational modifications.

De novo sequencing, on the other hand, suffers from the uncertainty of incomplete tandem mass spectra that may yield either a false-positive or a partial sequence of the precursor peptide. Integrated and hybrid approaches of pure *de novo* sequencing used in conjunction with efficient database searches developed for PST approaches have certainly improved present standards and overcome some limits of both PST and *de novo* tools taken separately but more efforts are still necessary and required.

6.1. The Future of Peptide/Protein Identification

The present focus in evaluating large datasets of fragment spectra seems to be shifting towards platforms that integrate different approaches and tools, as in VEMS [188] and DBParser [189]. The idea is certainly to get to a consensus-based framework that may gather information from different analysis tools. The underlying idea is that low-scoring, identified proteins may get significant and effective hits if concurrent and independent methods agree in the detection of the same protein [131]. Importantly, the field of proteomics is now largely in agreement about one simple rule: proteins should only be considered identified on the basis of two independent, non-overlapping and unique peptide sequences.

As already said, HUPO is the international umbrella organization representing proteomics researchers and is spearheading several projects such as HPPP [190,191]. HUPO's primary goal is to promote new experimental guidelines and standards to stabilize present protocols in lab procedures and analysis tools [192]. Working in this direction, the novel repository platform PRIDE (Proteomics IDentification database) [193] has been specifically designed to collect experimental data from proteomics experiments in a format arrived acknowledged by HUPO's PSI. Recently PRIDE has been recognized as the reference database for proteomics projects and now includes the experiments coming from the GPM (Global Proteome Machine) [81] database, which shares the same goals. In the future, the wide availability of controlled and assessed experimental data will certainly give raise to novel methods able to exploit this valuable information through more sophisticated machine learning approaches. A pilot study has already been described on the occurrence of "proteotypic peptide sequences" [194], those peptides in a protein sequence that are most likely to be confidently ob-

served by present proteomics methods. An experimental web service that implements the search of an input spectrum against a library of experimental spectra (that were confidently assigned to a particular peptide sequence), has been also created and called X!HUNTER. If successful, one attractive outcome of this project should be a much better understanding of the fragmentation process.

7. BIOMARKER DISCOVERY

Certainly the most active field of proteomics involves the search for biomarkers diagnostic for specific diseases. This work focuses mainly on proteomic profiling of blood [195], but other body fluids are being investigated for the same purpose, *e.g.* urine [196]. The HUPO's Plasma Proteome Project (HPPP), started in 2002, represents the joint effort of thirty five research groups all around the world who have been involved in the annotation of plasma-derived proteins and their bioinformatics analysis [191] (see also <http://www.plasmaproteomedatabase.org>). Plasma is particularly interesting for disease biomarker discovery because it interfaces with many different organs in the body but the pitfalls involved in its analysis are many: the concentration range between albumin and the 'interesting' proteins, the presence of very high levels of intracellular proteins as a result of cell lysis, highly variable results between individuals. The data from all thirty five research groups were collected on different instruments using multiple configurations and so they are highly variable. To address some of the complexity, Ping *et al.* undertook a functional annotation of the plasma proteomic data from all groups involved [197]. The main goal of the HPPP is to develop reproducible methods for the collection and analysis of samples with the hope of discovering discriminating, genuine biomarkers. While an enormous amount of efforts has been invested in this endeavor, the goal remains elusive. What is clear, however, is that collaboration between research teams with different expertise is crucial, especially with more bioinformaticians and statisticians. Geurts *et al.* [198] have pointed out that classical statistical tests are simply not robust enough for the analysis of the volumes of data that can come out of proteomic experiments. They go on to say, though, that this limitation could be overcome by the development of machine learning methods (tree-based ensembles and kernel-based methods). Indeed, the authors themselves developed decision tree-based software and applied it to SELDI-TOF-MS data for the diagnosis of rheumatoid arthritis and inflammatory bowel diseases. In order to study ovarian cancer, Yu and Chen developed Bayesian neural network models for SELDI-TOF-MS data [199] whereas Tibshirani *et al.* proposed the *peak probability contrast* technique [200]. After peak smoothing and centroiding, the authors extracted a number of clusters of peaks in the spectra by means of principal component analysis (PCA). The most discriminatory quantile height was determined and based on those features that were derived from new samples for the evaluation of spectra. Another interesting application of PCA was proposed by Bryant *et al.* [201] to discriminate between hepatic protein mixtures from rats treated with either methapyrilone or SB-219994. Although carried out as a preliminary study on a limited number of gel-bands, the ability of PCA to identify differences in complex related protein mixtures was here demonstrated. A different algorithm was proposed by Bensmail *et al.* and per-

formed well with SELDI-MS data from the sera of seventy individuals infected with the virus HTLV-1 [202]. Cluster analysis based on a Bayesian-Fourier approach was used to distinguish the proteomes of adult T-cell leukemia-affected patients from spastic paraparesis-affected patients, two diseases both associated with HTLV-1 infection. Based on the mass information only, they obtained two different clusters, whereas three different clusters (normal, leukemia, paraparesis) emerged when either the intensity dataset or the mass+intensity combined datasets were used. Yet a different application of cluster analysis was shown by Lancashire *et al.* in the identification of the bacterium *Neisseria meningitidis* from closely related taxa (*Neisseria gonorrhoeae*) [203] based on SELDI-TOF-MS proteomic data. Here, artificial neural network analysis was used to train 3 kDa blocks over the data mass range and by modeling the ion intensity profiles of each sample.

We have not covered exhaustively all the different approaches to statistical analysis and evaluation of large datasets but the increasing interest in this field, as evidenced by current literature, is encouraging. This trend seems likely to continue and provide proteomics investigators with the right tools to achieve higher and higher levels of reliability.

8. POST-TRANSLATIONAL MODIFICATIONS

Another popular and perhaps the most powerful biomedical application of mass spectrometry is the investigation of post-translational modifications (PTMs), although the implementation remains a non trivial task (see reviews [68,204-208] for details on this topic). PTMs occur in the cell after the protein has been translated from the mRNA template, and this post-processing of amino acids is very frequent. The functional groups can only be identified at the protein sequence level and serve to modulate protein integrity and interactions, as targeting signals for delivery of proteins into specific locations inside the cell and for transport across membranes. Some examples of important protein modifications are phosphorylation (pTyr, pSer, pThr), glycosylation (N-linked and O-linked), acetylation, methylation, fatty acid modification, sulfation, deamidation, ubiquitination, sumoylation. As proteomic studies produce reams of new information concerning PTMs, the Swiss-Prot knowledge base is trying to standardize the annotation of PTM features contained within it [209]. Mapping PTMs by mass spectrometry remains challenging for numerous reasons; first, sufficient quantity of protein is required, and usually affinity purification is needed in order to enrich the sample with the modified proteins. Moreover, testing for post-translational modifications involves a considerable computation cost for search engines, especially in the case of complex samples. Additional complexity is added by the possibility of combined PTMs on different amino acids of the same peptide. Lu and Chen found that the database search process could be accelerated by pre-processing the protein sequence database and the tandem mass spectra prior to matching [98]. For the same purpose, MacCoss *et al.* extracted the subset of database-identified proteins and put them into a new database. They then repeated the searches by considering several PTMs [205]. Hansen *et al.* developed an algorithm (*P-Mod*) to discover unanticipated protein modifications from database identified peptides [210]. MS/MS data files were screened to find mass shifts between identified peptides and measured

values, and finally scored. Local sequence preferences of short residues around the PTM sites were also used to train sequence models by support vector machines [211,212]. Searle *et al.* used the OpenSea algorithm [168] to first align de-novo sequencing-derived peptides to protein sequences in databases and then assumed regions of non-match to be either amino acid substitutions or modifications [169]. The method was tested on the PTM-rich lens tissue and was able to confirm several previously known PTMs, as well as to identify new ones. In addition to Swiss-Prot, which maintains the annotation of PTMs for given proteins there are three general repositories of PTMs, Unimod [213], DeltaMass and Resid [214].

The “Top-Down” approach consists of the MS analysis of intact proteins and represents one of the new directions in proteomics. PTMs have been characterized in a “Top-Down” approach [215] based on searching of protein fragment ions against databases annotated with both native and PTM-modified proteins (from the Resid nomenclature). The “absolute mass search” function matched protein, *b/y* and *c/z* masses to database entries, and derived probability scores based on the Poisson model. This function could then be combined with sequence-tag based searches.

To identify PTMs from peptide mass fingerprinting data, Gattiker *et al.* developed *FindPept*, which is integrated in the ExPASy suite of proteomic tools [216]. The program supports plain text, .dta and .pkm data file formats, and is also particularly useful since it can report an unlimited number of missed cleavages (the frequency of trypsin missed/unspecific cleavages is relatively high). It was shown that the tryptic rule of cleavage after Lys and Arg (except if followed by Pro) is not exhaustive, and that charged adjacent residues can have a strong influence on the cleavage efficiency [217].

To aid in the interpretation of 2D-gel separated proteins, Halligan *et al.* developed *ProMoST* [218], a program that calculated protein pI and molecular masses and examined the effect of single or multiple PTMs on the relative position of migration on 2D gels. This tool could provide valuable assistance in the comparison of observed spots with patterns of spots predicted for different PTMs. A similar web resource is *ProteoMod*, based on the calculation of pI shifts between PTM-modified and unmodified proteins to derive the number of phosphorylations [219]. Other PTMs were also described therein.

Phosphorylation and glycosylation are recognized as fundamental components of signaling and recognition pathways. Glycans are undoubtedly the most complex structures in this context. Indeed, the type, number, position, branching and further modifications of individual sugar residues can vary widely (see [220] for a nomenclature of glycan MS/MS fragments). The *GlycoSuiteDB* is an annotated relational database that contains glycan structures of many biological sources which are derived from the literature [221]. Lohmann *et al.* developed *GlycoFragment* to calculate all possible fragments of complex carbohydrates in a database, and *GlycoSearchMS* to compare experimental MS fragments with calculated fragments [222]. Tang *et al.* described an algorithm (*GLYCH*) for the characterization of glycans using MS/MS [223] that also considered cross-ring ions resulting from internal cleavages.

It has been estimated that one-third of all the human proteins could be phosphorylated, therefore large-scale phosphoproteomic studies stand to add great value of our understanding of cell processes. Phosphorylation analysis has often been reported as problematic due to suppression effect of phosphopeptide ionization in the presence of unmodified peptide and lower ionization efficiencies in general. Nevertheless, Steen *et al.* used mixture of standard peptides to demonstrate that most of these arguments are questionable and that powerful affinity purification procedures could compensate for the low percentage of modified/unmodified peptides in biological mixtures [208]. Zhou *et al.* developed an *in silico* prediction method named *GPS* based on the clustering of all the phosphorylation sites of 52 protein kinase families. They then calculated the distance of peptide sequences to the clusters and scored potential phosphorylation sites for the kinase represented by that cluster [224]. Zaho's *et al.* algorithm reconstructed all possible peptide sequences of a precursor ion and used them to derive the most probable phosphoprotein [225].

9. MASS SPECTROMETRY FOR PROTEIN QUANTIFICATION: CHALLENGES IN PROTEOMICS

Before the advent of mass spectrometry for the study of large pools of proteins, protein quantification relied mainly on the use of stains and antibodies. Nevertheless, the comparison of staining intensities of gel-electrophoresis-separated proteins suffers from low accuracy and provides only rough estimates of protein levels. Antibody-based approaches, though highly sensitive, are hampered by questions of specificity. Thanks to ongoing improvements and research into new technologies, mass spectrometers in combination with software tools can now quantify hundreds or thousands of proteins in tissues, cells, microorganisms, organelles and body fluids. The number of studies based on quantitative proteomics is increasing very rapidly, and exciting insights into cell biology and diagnostics have already been achieved using these approaches. In clinical applications, differentially expressed proteins in normal versus diseased samples can be identified. Mass spectrometry-based quantitative proteomic approaches fall into two main categories: absolute and relative quantification.

In absolute quantification, either the real concentration of single proteins within the sample or the protein copy number can be determined, in principle. To date it is not possible to accurately quantify relative amounts of different proteins within the same sample (and therefore in one single analysis), because the peptide/protein ionization efficiencies and mass spectrometry signals depend on the amino acid composition. Despite some recent achievements in absolute quantification, the measurement of protein abundance in complex samples remains challenging.

In the most widespread approaches, the relative amount of peptides and proteins can be determined between two or more different samples (*e.g.* different cell states, healthy *vs* diseased, normal *vs* treated, tissue A *vs* tissue B *vs* tissue C, etc.). One of the most popular approaches in relative quantification is to determine peptide amounts from the ratio of isotopically "light" and "heavy" peptide analogues. The ratio of the peak intensities is directly correlated with the ratio of abundance of the peptide in the two samples.

9.1. Relative Quantification

There are several ways to incorporate heavy isotopes into proteins or in peptides, including the use of cell culture media containing heavy isotopes (*e.g.* ^{15}N or ^{18}O), metabolic labeling in media enriched in stable isotope-containing amino acids (*e.g.*, SILAC) and chemical reactions with isotopically-labeled tags (*e.g.*, ICAT, iTRAQ, HysTag). A large number of methods, some very similar to one another, have been described for chemical or metabolic incorporation of mass tags (see [226] for a comprehensive review of these) but the one thing common to all these methods is the requirement for extracting the quantitative data from raw mass spectra after analysis. By digesting proteins in ^{18}O -containing media (such as H_2^{18}O), oxygen atoms at the C-terminus of forming peptides are exchanged for the heavy isotopic form [227]. Two independent samples can be differentially labeled with ^{18}O and ^{16}O (unlabeled) for relative peptide quantification. An interesting program (*ZoomScan*) has been developed that uses high resolution zoom scan spectra acquired with ion trap instruments to calculate the ratios of ^{18}O labeled to ^{16}O unlabeled peptides [228]. Three different methods were used to derive the theoretical isotopic envelope distribution of database-identified peptides. These distributions were then used to evaluate the contribution of labeled and unlabeled peptides to the peak areas. In the SILAC approach [229], cells are cultured in media containing ^{13}C and ^{15}N enriched essential amino acids ($^{13}\text{C}_6\text{-Arg}$, $^{13}\text{C}_6\text{-Lys}$, $^{13}\text{C}_6^{15}\text{N}_4\text{-Arg}$). After several cell doublings, all the proteins in the cell will incorporate these amino acids. This approach was very effective in comparing protein expression in normal *vs* treated cell cultures. Normal and labeled lysates were first mixed, and then digested with trypsin. Mass spectrometry data were searched with MASCOT, specifying the labeled amino acids as variable modifications. *MSQuant* [230] is a more generic package that was originally developed to quantify SILAC [229] data but it has since been adapted to any survey scan-based quantitation. *MSQuant* works with a Mascot output to link search results to the underlying raw data. One unique feature of *MSQuant* is how it calculates differences between the control and experimental conditions. Most such calculations are based on integrating the area under the chromatographic peak of each of the ions being followed. However, for isotopically-enriched tags that co-elute with the normal isotopic abundance forms (*e.g.*, deuterated leucine does not co-elute with hydrogenated leucine from reversed phase material whereas ^{13}C -labelled leucine does coelute with normal isotopic abundance leucine) *MSQuant* also calculates a weighted average of the ratios in each survey scan comprising the chromatographic peak. This effectively reduces the measurement error for each ratio and provides a value for that error in the variability across the peak. For a recently published application of SILAC see [231]. In another approach, amino acids containing ^{15}N were included in animal diets to obtain isotopically-labeled proteins in all the animal tissues [232]. Proteins were searched against two different databases; one normal database and one that had been ^{15}N -corrected to take into account the mass shift due to the heavy isotope incorporation. Ion chromatograms were extracted from raw data using a modified version of the program *EXTRACT-CHRO* [233]. The program *EXTRACT-CHRO* was tested on yeast cultured in ^{15}N -enriched minimal growth media. After deriving the peptide sequences of inter-

est from SEQUEST results, the program was used to predict the isotope envelopes for both the ^{14}N and ^{15}N -labeled peptides. This information was further exploited to estimate the m/z range in order to extract ion chromatograms for 100 MS precursor scans surrounding the ion. Quantitative peak analysis and evaluation of the extracted LC pairs were performed using the program *RelEx* (RELative EXpression). *RelEx* performed peak smoothing, background subtraction and calculation of labeled to native peptide ratios. In a different version, *RelEx* was modified to perform quantification from MS/MS spectra rather than from MS scans [234]. Here, successive windows of 10 m/z were isolated and subsequently fragmented throughout the mass spectrum. The software attempted the reconstruction of the ion chromatogram from fragment ion intensities in the MS/MS spectra, resulting in increased signal/noise. Nevertheless, the large mass window that had to be used by the database search algorithms likely resulted in a very high number of false positive identifications. A cross-correlation algorithm was also introduced to derive the peptide mass from its MS/MS spectrum. According to the authors, further improvements are required to handle MS/MS spectra convolution caused by the co-elution and fragmentation of isobaric peptides.

Another method is to chemically modify proteins by labeling some reactive amino acid groups with stable isotope labels. Usually, two protein samples are reacted with either the light or the heavy form of the label. A more comprehensive view of the chemical tagging methods in quantitative proteomics can be found in a recent review by Ong and Mann [226]. Some examples will help clarifying this approach. In the Isotope-Coded Affinity Tags (ICAT) method [235] two different protein mixtures were tagged using biotin-derivatized affinity tags containing a heavy/light isotope linker and a thiol-reactive group. Relative protein levels could be derived by measuring peak ratios of the heavy/light peptide forms. The XPRESS software developed for ICAT experiments [109] (<http://tools.proteomecenter.org/XPRESS.php>) reconstructs the elution profiles of the heavy/light isotopes to derive area ratios. In a similar approach brain plasma membranes from two different brain regions were targeted with thiol-reactive peptides (HysTag) containing either deuterium-labeled or normal isotopic abundance alanine [236]. Relative quantification of tagged vs non-tagged common peptides in the two samples was performed with *MSQuant* (see above) by adapting the software to take the HysTag mass shift into account. More recently, Applied Biosystems has introduced isobaric Tags for Relative and Absolute Quantitation (iTRAQ) [237]. The major difference between iTRAQ and the other stable isotope methods mentioned here is that iTRAQ quantitation is at the level of tandem mass spectra rather than on survey scans since each of the different labels is isobaric and only the fragments are of different masses.

9.2. Absolute Quantification

The basic principle underlying relative quantitation can be extended to measure absolute levels of analytes by spiking known amounts of isotopically-labelled standards into a sample. The unlabeled experimental peptide and the labeled standard peptide undergo the same sample preparation and mass spectrometric detection, therefore the ratio of their ion intensities should be accurate enough to derive a quantitative

measurement of the native peptide based on the concentration of the spiked peptide. To this purpose, peptides have been synthesized with incorporated stable isotopes (e.g. ^{13}C and ^{15}N enriched Leucine) that mimic native peptides in protein samples. These peptides can be used as internal standards to provide absolute protein quantification by mass spectrometry. In addition, the *AQUA* (Absolute QUAntification) strategy was applied to mimic native peptides containing post-translational modifications [238]. Standard peptides incorporating O^{18} into carboxyl-terminal moieties of tryptic peptides were spiked with known amounts of simple protein mixtures [239] to test the reliability of *in solution* and *in gel* absolute quantification of proteins. Here, the quantification was based on doubly charged precursors, and the areas of monoisotopic peaks of the highest intensity y -ion fragments were exclusively considered. The amount of peptides in the mixtures was determined from area ratios by using commercial software. As the monoisotopic peaks of ^{16}O and ^{18}O peptides are spaced from each other by 2 Da, a deconvolution algorithm was applied to handle overlaps of isotopic clusters and calculates their individual contributions to the peak areas [240]. Interestingly, the authors showed that MALDI-MS and nanoESI-MS produced consistent results in ^{18}O quantification, although longer acquisition times of nanoESI provided better accuracy in the measurement of peak areas.

Internal standards are particularly useful for the quantification of selected proteins in reasonably simple mixtures. The LC (Liquid Chromatography) traces of selected peptide ions can be extracted in a narrow time window and the ratio between the native and the internal standard peptide abundance determined with the help of software packages. Clearly, due to the complexity and costs, only a limited number of internal standard peptides can be introduced in one sample, so this approach can only provide the accurate quantification of a limited number of proteins. Based on published studies, data analysis is usually performed by in-house-developed programs that extract ion chromatograms and derive peak areas from raw data in different ways. Unfortunately, a thorough documentation of the software packages used by individual labs is not always provided together with the published proteomic data. The well-established method of external calibration used in both gas-chromatography mass spectrometry (GC-MS) and LC-MS (for the analysis of small organic molecules such as pesticides, pollutants, amino acids, drugs, hormones, lipids and carbohydrates, hydrocarbons [34,241,242]), was recently applied to a proteomic study to estimate the concentration of rhodopsin [243]. Standard curves were obtained from mass spectrometry of synthetic peptides spiked at different concentrations, demonstrating acceptable recoveries and measurement accuracies of both non-labeled external standard and labeled internal standard peptides.

Very recently, a number of independent studies have demonstrated the reliability of label-free ion-intensity-based peptide quantification in related samples. It has been shown that by spiking complex protein samples such as the human serum with different amounts of standard proteins, a linear response can be observed. It is tempting to jump to the conclusion that by carefully monitoring the experimental conditions, ion intensities could be correlated with peptide concentration but further validation of this concept is required. Due to their variable amino acid composition, peptides are

very different from each other in terms of chemical properties, so their behavior and yields differ in terms of chromatographic separation and ionization in the mass spectrometer source. Certainly, the lack of a labeling step would be highly advantageous in that exogenous modifications of biological samples would no longer be required. Moreover, the door would be opened to the quantitative analysis of proteins extracted directly from animal fluids and tissues, rather than from cell culture models. Although isotopically-labeled tags have been successfully linked to proteins extracted from tissues (see above), the effective yields of these reactions are almost never quantitative in complex protein mixtures. On the other hand, it has been recently shown by independent works that comparison of the ion intensities (XIC, eXtracted Ion Chromatogram) for the same peptide in different LC/MS analyses was a reliable means of comparison [226,244-246], addressing concerns about the nonlinearity of absolute quantification due to suppression effects in the ion source. An interesting work demonstrated that absolute quantification could be the method of choice given sufficiently careful sample preparation [246]. Spectra intensity normalization was performed based on molecules of constant concentration among different samples by means of the *MassView* software. In another study, by spiking the human serum with different concentration of horse myoglobin, a good linear correlation of peak areas with myoglobin concentration was obtained [244]. Here, correlation was improved by normalizing the peak areas with a correction factor representative of experimental variability.

In an alternative approach to absolute quantitation the number of identified peptides per protein was shown to be roughly proportional to the protein concentration within the sample [247]. In a previous work, the authors introduced the protein abundance index (PAI), based on the number of identified peptides per protein normalized by the theoretical number of observable peptides. The normalization removed the bias of this method towards larger proteins [248] and the PAI metric was further refined by expressing it relative to the logarithm of the protein concentration.

A common drawback reported by researchers dealing with complex protein mixtures is that MS analyses of replicates from the same sample are not as reproducible as expected, neither qualitatively nor quantitatively. This is, of course, the major limitation of absolute quantification. Nevertheless, thanks to improvements in the instrument technology, promising results are being achieved in this regard. Interestingly, it was shown that in 9 LC-MS/MS replicate runs of a yeast extract sample, most proteins were identified either in one or in all of the runs [245]. A statistical model was developed to predict how many runs would in theory be necessary to identify all the peptides contained in a complex mixture, in other words, to reach saturation. Indeed, it must not be forgotten that there is always a sampling bias toward abundant ions, which repeatedly trigger data dependent acquisition.

Statistics plays a key role in the evaluation of significant differences among biological samples, especially in the field of biomarker discovery. The importance of statistical analysis becomes clear if we consider that intrinsic biological variability and some degree of randomness in the mass spectrometry detection might produce misleading results. It is

therefore of primary importance that both qualitative and quantitative measurements are evaluated in replicates. The need for reporting mass spectrometry data with measures of confidence and description of methods for error estimation (also in quantitative studies) has also been widely agreed at the last HUPO annual world congress (Munich, 2005) [175]. Recently, a statistics-based comparative study of human plasma spiked with low concentrations of marker proteins was undertaken by combining peptide identification scores (PMSS, peptide match score summation [249]) with the local pooled error test [250,251]. Based on PMSS, marker proteins were detected with 90% confidence with only two or three replicates. This same analysis performed similarly well when using the spectrum counts as quantitative indicator [245]. Indeed, several semi-quantitative indicators of protein abundance were implemented in the data processing program (not provided). In another study, LC peak overlap (i.e. reproducibility of retention times through LC runs) was shown to increase considerably with the increasing number of analyses [252], reinforcing the concept that replicates improve data consistency. A platform integrating several modules and performing peak detection, peak alignment and peptide quantitation was evaluated for biomarker screening in the blood. The intensities of 400 peaks from two LC-MS replicates were aligned in a (\log_{10}) scatter plot to demonstrate a good correlation (R^2 0.84).

CONCLUSION

In conclusion, we have tried to give the readers a feeling for the plethora of tools available to proteomics researchers for the interpretation of mass spectra. The applications range from the biomarker detection, *de novo* sequencing of proteins to identification of post-translational processes, protein interactions and quantification of expressed proteins. It should now be clear that it is necessary to integrate several informatic tools in order to rationalize mass spectrometry data. Indeed, while tremendous efforts have been expended to improve current instrumentation and spectral analysis tools, the informatic tools still lag far behind the ability to acquire high quality data. We believe that with the proliferation of so many methods for spectra analysis, rigorous scrutiny is required to evaluate any real contributions to current state-of-the-art, a task ideally suited to HUPO's PSI. While Mascot and SEQUEST are the current workhorses, open-source engines such as the GPM/X!TANDEM seem likely to dominate in the future as their open-source nature will allow interested parties to incorporate new analysis tools and through natural selection will allow the tools to evolve to their full potential. Within three years we expect that this will translate practically into researchers being able to assign two thirds of high quality spectra to peptides or modifications, rather than the maximum one quarter to one third that can currently be utilized. We also believe that a tighter integration of different strategies, contributed from current *de novo*, SPC and PST methods, could strongly enhance this process (e.g., Popitam in section 4.3). This integration is most likely to be driven by open source platforms, perhaps leading to the proliferation of meta-servers that will be able to collect and integrate the results of different algorithms with the aim of creating a consensus (e.g, SCAFFOLD in section 4.1.2). This approach will certainly make PTM detection, biomarker discovery and protein quantification more

feasible than ever and accurate mass measurements, combined with high quality fragment mass spectra will provide a veritable mine of reliable data that could act as a spectral library for all new mass spectrometry peptide sequence identifications.

ACKNOWLEDGEMENTS

The authors wish to thank Matthias Mann for his constructive comments and critical reading of the manuscript. LJF is the Canada Research Chair in Organelle Proteomics and a Michael Smith Foundation Scholar. This work was funded in part by the Canadian Institutes for Health Research Operating Grant #MOP-77688

REFERENCES

- [1] Proteomics, transcriptomics: what's in a name? *Nature* **1999**; 402: 715.
- [2] Mootha VK, Bunkenborg J, Olsen JV, *et al.* Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* **2003**; 115: 629-40.
- [3] Forner F, Foster LJ, Campanaro S, Valle G, Mann M. Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. *Mol Cell Proteomics* **2006**; 5: 608-19.
- [4] Nedelkov D, Kiernan UA, Niederkofler EE, Tubbs KA, Nelson RW. Investigating diversity in human plasma proteins. *Proc Natl Acad Sci USA* **2005**; 102: 10852-7.
- [5] Callesen AK, Mohammed S, Bunkenborg J, *et al.* Serum protein profiling by miniaturized solid-phase extraction and matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom* **2005**; 19: 1578-86.
- [6] Fujii K, Nakano T, Kawamura T, *et al.* Multidimensional protein profiling technology and its application to human plasma proteome. *J Proteome Res* **2004**; 3: 712-8.
- [7] Zhang X, Leung SM, Morris CR, Shigenaga MK. Evaluation of a novel, integrated approach using functionalized magnetic beads, bench-top MALDI-TOF-MS with prestructured sample supports, and pattern recognition software for profiling potential biomarkers in human plasma. *J Biomol Tech* **2004**; 15: 167-75.
- [8] Jain KK. Role of oncoproteomics in the personalized management of cancer. *Expert Rev Proteomics* **2004**; 1: 49-55.
- [9] Drake RR, Cazare LH, Semmes OJ, Wadsworth JT. Serum, salivary and tissue proteomics for discovery of biomarkers for head and neck cancers. *Expert Rev Mol Diagn* **2005**; 5: 93-100.
- [10] Davidsson P, Sjogren M. The use of proteomics in biomarker discovery in neurodegenerative diseases. *Dis Markers* **2005**; 21: 81-92.
- [11] Henley SM, Bates GP, Tabrizi SJ. Biomarkers for neurodegenerative diseases. *Curr Opin Neurol* **2005**; 18: 698-705.
- [12] Zhang J, Goodlett DR, Quinn JF, *et al.* Quantitative proteomics of cerebrospinal fluid from patients with Alzheimer disease. *J Alzheimers Dis* **2005**; 7: 125-33; discussion 73-80.
- [13] Stanley BA, Gundry RL, Cotter RJ, Van Eyk JE. Heart disease, clinical proteomics and mass spectrometry. *Dis Markers* **2004**; 20: 167-78.
- [14] Vivanco F, Martin-Ventura JL, Duran MC, *et al.* Quest for novel cardiovascular biomarkers by proteomic analysis. *J Proteome Res* **2005**; 4: 1181-91.
- [15] De Celle T, Vanrobaeys F, Lijnen P, *et al.* Alterations in mouse cardiac proteome after *in vivo* myocardial infarction: permanent ischaemia versus ischaemia-reperfusion. *Exp Physiol* **2005**; 90: 593-606.
- [16] Fernando P, Deng W, Pekalska B, *et al.* Active kinase proteome screening reveals novel signal complexity in cardiomyopathy. *Mol Cell Proteomics* **2005**; 4: 673-82.
- [17] Garcia A, Watson SP, Dwek RA, Zitzmann N. Applying proteomics technology to platelet research. *Mass Spectrom Rev* **2005**; 24: 918-30.
- [18] Natsume T, Yamauchi Y, Nakayama H, *et al.* A direct nanoflow liquid chromatography-tandem mass spectrometry system for interaction proteomics. *Anal Chem* **2002**; 74: 4725-33.
- [19] Cho S, Park SG, Lee do H, Park BC. Protein-protein interaction networks: from interactions to networks. *J Biochem Mol Biol* **2004**; 37: 45-52.
- [20] Yates JR, 3rd, Gilchrist A, Howell KE, Bergeron JJ. Proteomics of organelles and large cellular structures. *Nat Rev Mol Cell Biol* **2005**; 6: 702-14.
- [21] Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989**; 246: 64-71.
- [22] Whitehouse CM, Dreyer RN, Yamashita M, Fenn JB. Electrospray interface for liquid chromatographs and mass spectrometers. *Anal Chem* **1985**; 57: 675-9.
- [23] Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* **1988**; 60: 2299-301.
- [24] Karas M, Gluckmann M, Schafer J. Ionization in matrix-assisted laser desorption/ionization: singly charged molecular ions are the lucky survivors. *J Mass Spectrom* **2000**; 35: 1-12.
- [25] Yost RA, Boyd RK. Tandem mass spectrometry: quadrupole and hybrid instruments. *Methods Enzymol* **1990**; 193: 154-200.
- [26] Morris HR, Paxton T, Dell A, *et al.* High sensitivity collisionally-activated decomposition tandem mass spectrometry on a novel quadrupole/orthogonal-acceleration time-of-flight mass spectrometer. *Rapid Commun Mass Spectrom* **1996**; 10: 889-96.
- [27] Hayes RN, Gross ML. Collision-induced dissociation. *Methods Enzymol* **1990**; 193: 237-63.
- [28] McLuckey SA, Goeringer DE, Glish GL. Collisional activation with random noise in ion trap mass spectrometry. *Anal Chem* **1992**; 64: 1455-60.
- [29] Chernushevich IV, Loboda AV, Thomson BA. An introduction to quadrupole-time-of-flight mass spectrometry. *J Mass Spectrom* **2001**; 36: 849-65.
- [30] Roepstorff P. MALDI-TOF mass spectrometry in protein chemistry. *EXS* **2000**; 88: 81-97.
- [31] Marvin LF, Roberts MA, Fay LB. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry in clinical chemistry. *Clin Chim Acta* **2003**; 337: 11-21.
- [32] Bienvenut WV, Deon C, Pasquarello C, *et al.* Matrix-assisted laser desorption/ionization-tandem mass spectrometry with high resolution and sensitivity for identification and characterization of proteins. *Proteomics* **2002**; 2: 868-76.
- [33] Peterman SM, Dufresne CP, Horning S. The use of a hybrid linear trap/FT-ICR mass spectrometer for on-line high resolution/high mass accuracy bottom-up sequencing. *J Biomol Tech* **2005**; 16: 112-24.
- [34] Gizzi G, Hoogenboom LA, Von Holst C, Rose M, Anklam E. Determination of dioxins (PCDDs/PCDFs) and PCBs in food and feed using the DR CALUX bioassay: results of an international validation study. *Food Addit Contam* **2005**; 22: 472-81.
- [35] De Hoffmann D, Stroobant V. *Mass Spectrometry: Principles and Applications*, 2nd Edition. New York: John Wiley & Sons, 2001.
- [36] Olsen JV, de Godoy LM, Li G, *et al.* Parts per million mass accuracy on an orbitrap mass spectrometer via lock-mass injection into a C-trap. *Mol Cell Proteomics* **2005**.
- [37] Cargile BJ, Bundy JL, Stephenson JL, Jr. Potential for false positive identifications from large databases through tandem mass spectrometry. *J Proteome Res* **2004**; 3: 1082-5.
- [38] Syka JE, Marto JA, Bai DL, *et al.* Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *J Proteome Res* **2004**; 3: 621-6.
- [39] Gorshkov MV, Zubarev RA. On the accuracy of polypeptide masses measured in a linear ion trap. *Rapid Commun Mass Spectrom* **2005**; 19: 3755-58.
- [40] Tabb DL, Smith LL, Brezi LA, Wysocki VH, Lin D, Yates JR, 3rd. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal Chem* **2003**; 75: 1155-63.
- [41] Hunt DF, Yates JR, 3rd, Shabanowitz J, Winston S, Hauer CR. Protein sequencing by tandem mass spectrometry. *Proc Natl Acad Sci USA* **1986**; 83: 6233-7.
- [42] Brezi LA, Tabb DL, Yates JR, 3rd, Wysocki VH. Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra. *Anal Chem* **2003**; 75: 1963-71.
- [43] Gu C, Tsapralis G, Brezi L, Wysocki VH. Selective gas-phase cleavage at the peptide bond C-terminal to aspartic acid in fixed-charge derivatives of Asp-containing peptides. *Anal Chem* **2000**; 72: 5804-13.

- [44] Wysocki VH, Tsapralis G, Smith LL, Breci LA. Mobile and localized protons: a framework for understanding peptide dissociation. *J Mass Spectrom* **2000**; 35: 1399-406.
- [45] Roepstorff P, Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom* **1984**; 11: 601.
- [46] Biemann K. Contributions of mass spectrometry to peptide and protein structure. *Biomed Environ Mass Spectrom* **1988**; 16: 99-111.
- [47] Paizs B, Suhai S. Towards understanding the tandem mass spectra of protonated oligopeptides. 1: mechanism of amide bond cleavage. *J Am Soc Mass Spectrom* **2004**; 15: 103-13.
- [48] Dongre AR, Somogyi A, Wysocki VH. Surface-induced dissociation: an effective tool to probe structure, energetics and fragmentation mechanisms of protonated peptides. *J Mass Spectrom* **1996**; 31: 339-50.
- [49] Csonka IP, Paizs B, Lendvay G, Suhai S. Proton mobility and main fragmentation pathways of protonated lysylglycine. *Rapid Commun Mass Spectrom* **2001**; 15: 1457-72.
- [50] Paizs B, Suhai S. Fragmentation pathways of protonated peptides. *Mass Spectrom Rev* **2005**; 24: 508-48.
- [51] Havilio M, Haddad Y, Smilansky Z. Intensity-based statistical scorer for tandem mass spectrometry. *Anal Chem* **2003**; 75: 435-44.
- [52] Kapp EA, Schutz F, Reid GE, et al. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal Chem* **2003**; 75: 6251-64.
- [53] Schwartz BL, Bursey MM. Some proline substituent effects in the tandem mass spectrum of protonated pentaalanine. *Biol Mass Spectrom* **1992**; 21: 92-6.
- [54] Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat Biotechnol* **2004**; 22: 214-9.
- [55] Gibbons FD, Elias JE, Gygi SP, Roth FP. SILVER helps assign peptides to tandem mass spectra using intensity-based scoring. *J Am Soc Mass Spectrom* **2004**; 15: 910-2.
- [56] Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal Chem* **2004**; 76: 3908-22.
- [57] Mann M, Meng CK, Fenn JB. Interpreting mass spectra of multiply charged ions. *Anal Chem* **1989**; 61: 1702-8.
- [58] Horn DM, Zubarev RA, McLafferty FW. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J Am Soc Mass Spectrom* **2000**; 11: 320-32.
- [59] Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. *De novo* peptide sequencing via tandem mass spectrometry. *J Comput Biol* **1999**; 6: 327-42.
- [60] Colinge J, Magnin J, Dessingy T, Giron M, Masselot A. Improved peptide charge state assignment. *Proteomics* **2003**; 3: 1434-40.
- [61] Maleknia SD, Downard KM. Charge ratio analysis method: approach for the deconvolution of electrospray mass spectra. *Anal Chem* **2005**; 77: 111-9.
- [62] Swanson SK, Washburn MP. The continuing evolution of shotgun proteomics. *Drug Discov Today* **2005**; 10: 719-25.
- [63] Nesvizhskii AI, Aebersold R. Interpretation of Shotgun Proteomic Data: The Protein Inference Problem. *Mol Cell Proteomics* **2005**; 4: 1419-40.
- [64] MacCoss MJ. Computational analysis of shotgun proteomics data. *Curr Opin Chem Biol* **2005**; 9: 88-94.
- [65] Delahunty C, Yates JR, 3rd. Protein identification using 2D-LC-MS/MS. *Methods* **2005**; 35: 248-55.
- [66] Johnson RS, Davis MT, Taylor JA, Patterson SD. Informatics for protein identification by mass spectrometry. *Methods* **2005**; 35: 223-36.
- [67] Lin D, Tabb DL, Yates JR, 3rd. Large-scale protein identification using mass spectrometry. *Biochim Biophys Acta* **2003**; 1646: 1-10.
- [68] Reinders J, Lewandrowski U, Moebius J, Wagner Y, Sickmann A. Challenges in mass spectrometry-based proteomics. *Proteomics* **2004**; 4: 3686-703.
- [69] Sadygov RG, Cociorva D, Yates JR, 3rd. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods* **2004**; 1: 195-202.
- [70] Shadforth I, Todd K, Crowther D, Bessant C. Determination of partial amino acid composition from tandem mass spectra for use in peptide identification strategies. *Proteomics* **2005**; 5: 1787-96.
- [71] Steen H, Mann M. The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* **2004**; 5: 699-711.
- [72] Standing KG. Peptide and protein *de novo* sequencing by mass spectrometry. *Curr Opin Struct Biol* **2003**; 13: 595-601.
- [73] Yates JR, 3rd, Eng JK, McCormack AL. Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem* **1995**; 67: 3202-10.
- [74] Yates JR, 3rd, Eng JK, McCormack AL, Schieltz D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* **1995**; 67: 1426-36.
- [75] Eng JK, McCormack AL, Yates JR, 3rd. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **1994**; 5: 976-89.
- [76] Field HI, Fenyo D, Beavis RC. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* **2002**; 2: 36-47.
- [77] Liebler DC, Hansen BT, Jones JA, Badghisi H, Mason DE. Mapping protein modifications with liquid chromatography-mass spectrometry and the SALSA algorithm. *Adv Protein Chem* **2003**; 65: 195-216.
- [78] Hansen BT, Jones JA, Mason DE, Liebler DC. SALSA: a pattern recognition algorithm to detect electrophile-adducted peptides by automated evaluation of CID spectra in LC-MS-MS analyses. *Anal Chem* **2001**; 73: 1676-83.
- [79] Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**; 20: 3551-67.
- [80] Pappin DJ, Hojrup P, Bleasby AJ. Rapid identification of proteins by peptide-mass fingerprinting. *Curr Biol* **1993**; 3: 327-32.
- [81] Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* **2004**; 3: 1234-42.
- [82] Craig R, Beavis RC. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom* **2003**; 17: 2310-6.
- [83] Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**; 20: 1466-7.
- [84] Geer LY, Markey SP, Kowalak JA, et al. Open mass spectrometry search algorithm. *J Proteome Res* **2004**; 3: 958-64.
- [85] Eriksson J, Fenyo D. Probit: a protein identification algorithm with accurate assignment of the statistical significance of the results. *J Proteome Res* **2004**; 3: 32-6.
- [86] Zhang N, Aebersold R, Schwikowski B. ProbiD: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2002**; 2: 1406-12.
- [87] Zhang W, Chait BT. ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal Chem* **2000**; 72: 2482-9.
- [88] Fu Y, Yang Q, Sun R, et al. Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics* **2004**; 20: 1948-54.
- [89] Tabb DL, MacCoss MJ, Wu CC, Anderson SD, Yates JR, 3rd. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal Chem* **2003**; 75: 2470-7.
- [90] Li D, Fu Y, Sun R, et al. pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics* **2005**; 21: 3049-50.
- [91] Colinge J, Masselot A, Giron M, Dessingy T, Magnin J. OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics* **2003**; 3: 1454-63.
- [92] Bafna V, Edwards N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* **2001**; 17(Suppl 1): S13-21.
- [93] Cargile BJ, Stephenson JL, Jr. An alternative to tandem mass spectrometry: isoelectric point and accurate mass for the identification of peptides. *Anal Chem* **2004**; 76: 267-75.
- [94] Mohan D, Pasa-Tolic L, Masselon CD, et al. Integration of electrokinetic-based multidimensional separation/concentration platform with electrospray ionization-Fourier transform ion cyclotron resonance-mass spectrometry for proteome analysis of *Shewanella oneidensis*. *Anal Chem* **2003**; 75: 4432-40.
- [95] Nielsen ML, Savitski MM, Zubarev RA. Improving protein identification using complementary fragmentation techniques in fourier transform mass spectrometry. *Mol Cell Proteomics* **2005**; 4: 835-45.

- [96] Savitski MM, Nielsen ML, Zubarev RA. New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. *Mol Cell Proteomics* **2005**; 4: 1180-8.
- [97] Zhang N, Li XJ, Ye M, Pan S, Schwikowski B, Aebersold R. ProbiDtree: an automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* **2005**; 5: 4096-106.
- [98] Lu B, Chen T. A suffix tree approach to the interpretation of tandem mass spectra: applications to peptides of non-specific digestion and post-translational modifications. *Bioinformatics* **2003**; 19(Suppl 2): II113-II21.
- [99] Tabb DL, Narasimhan C, Strader MB, Hettich RL. DBDigger: reorganized proteomic database identification that improves flexibility and speed. *Anal Chem* **2005**; 77: 2464-74.
- [100] Falkner J, Andrews P. Fast tandem mass spectra-based protein identification regardless of the number of spectra or potential modifications examined. *Bioinformatics* **2005**; 21: 2177-84.
- [101] Nesvizhskii AI, Aebersold R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discov Today* **2004**; 9: 173-81.
- [102] Sadygov RG, Yates JR, 3rd. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal Chem* **2003**; 75: 3792-8.
- [103] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **2002**; 74: 5383-92.
- [104] Keller A, Purvine S, Nesvizhskii AI, Stolyar S, Goodlett DR, Kolker E. Experimental protein mixture for validating tandem mass spectral analysis. *OMICS* **2002**; 6: 207-12.
- [105] Sadygov RG, Liu H, Yates JR. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal Chem* **2004**; 76: 1664-71.
- [106] Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* **2003**; 75: 4646-58.
- [107] Tabb DL, McDonald WH, Yates JR, 3rd. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* **2002**; 1: 21-6.
- [108] Eddes JS, Kapp EA, Frecklington DF, et al. CHOMPER: a bioinformatic tool for rapid validation of tandem mass spectrometry search results associated with high-throughput proteomic strategies. *Proteomics* **2002**; 2: 1097-103.
- [109] Han DK, Eng J, Zhou H, Aebersold R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat Biotechnol* **2001**; 19: 946-51.
- [110] Eriksson J, Fenyo D. The statistical significance of protein identification results as a function of the number of protein sequences searched. *J Proteome Res* **2004**; 3: 979-82.
- [111] MacCoss MJ, Wu CC, Yates JR, 3rd. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal Chem* **2002**; 74: 5593-9.
- [112] Lopez-Ferrer D, Martinez-Bartolome S, Villar M, Campillos M, Martin-Maroto F, Vazquez J. Statistical model for large-scale peptide identification in databases from tandem mass spectra using SEQUEST. *Anal Chem* **2004**; 76: 6853-60.
- [113] Yang B, Ying W, Gong Y, et al. Using cross-correlation normalized for peptide length to optimize peptide identification in shotgun proteomics. *Rapid Commun Mass Spectrom* **2005**; 19: 2983-85.
- [114] Moore RE, Young MK, Lee TD. Qscore: an algorithm for evaluating SEQUEST database search results. *J Am Soc Mass Spectrom* **2002**; 13: 378-86.
- [115] Li F, Sun W, Gao Y, Wang J. RScore: a peptide randomness score for evaluating tandem mass spectra. *Rapid Commun Mass Spectrom* **2004**; 18: 1655-9.
- [116] Dworzanski JP, Snyder AP, Chen R, Zhang H, Wishart D, Li L. Identification of bacteria using tandem mass spectrometry combined with a proteome database and statistical scoring. *Anal Chem* **2004**; 76: 2355-66.
- [117] Qian WJ, Liu T, Monroe ME, et al. Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome. *J Proteome Res* **2005**; 4: 53-62.
- [118] Sun W, Li F, Wang J, Zheng D, Gao Y. AMASS: software for automatically validating the quality of MS/MS spectrum from SEQUEST results. *Mol Cell Proteomics* **2004**; 3: 1194-9.
- [119] Rogalski JC, Lin MS, Sniatynski MJ, et al. Statistical evaluation of electrospray tandem mass spectra for optimized peptide fragmentation. *J Am Soc Mass Spectrom* **2005**; 16: 505-14.
- [120] Fridman T, Razumovskaya J, Verberkmoes N, Hurst G, Protopopescu V, Xu Y. The probability distribution for a random match between an experimental-theoretical spectral pair in tandem mass spectrometry. *J Bioinform Comput Biol* **2005**; 3: 455-76.
- [121] Zhao Y, Lin YH. A proteomic tool for protein identification from tandem mass spectral data. *Proteomics* **2005**; 5: 853-5.
- [122] Baczek T, Bucinski A, Ivanov AR, Kalisz R. Artificial neural network analysis for evaluation of peptide MS/MS spectra in proteomics. *Anal Chem* **2004**; 76: 1726-32.
- [123] Razumovskaya J, Olman V, Xu D, et al. A computational method for assessing peptide-identification reliability in tandem mass spectrometry analysis with SEQUEST. *Proteomics* **2004**; 4: 961-9.
- [124] Anderson DC, Li W, Payan DG, Noble WS. 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* **2003**; 2: 137-46.
- [125] Xu M, Geer LY, Bryant SH, et al. Assessing data quality of peptide mass spectra obtained by quadrupole ion trap mass spectrometry. *J Proteome Res* **2005**; 4: 300-5.
- [126] Chen Y, Kwon SW, Kim SC, Zhao Y. Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra. *J Proteome Res* **2005**; 4: 998-1005.
- [127] Shinkawa T, Taoka M, Yamauchi Y, et al. STEM: A Software Tool for Large-Scale Proteomic Data Analyses. *J Proteome Res* **2005**; 4: 1826-31.
- [128] Creasy DM, Cottrell JS. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics* **2002**; 2: 1426-34.
- [129] Weatherly DB, Astwood JA, 3rd, Minning TA, Cavola C, Tarleton RL, Orlando R. A Heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. *Mol Cell Proteomics* **2005**; 4: 762-72.
- [130] Rudnick PA, Wang Y, Evans E, Lee CS, Balgley BM. Large scale analysis of MASCOT results using a Mass Accuracy-based Threshold (MATH) effectively improves data interpretation. *J Proteome Res* **2005**; 4: 1353-60.
- [131] Kapp EA, Schutz F, Connolly LM, et al. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics* **2005**; 5: 3475-90.
- [132] Fenyo D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal Chem* **2003**; 75: 768-74.
- [133] Resing KA, Meyer-Arendt K, Mendoza AM, et al. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal Chem* **2004**; 76: 3556-68.
- [134] Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A. The need for guidelines in publication of peptide and protein identification data: working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol Cell Proteomics* **2004**; 3: 531-3.
- [135] Chamrad DC, Korting G, Stuhler K, Meyer HE, Klose J, Bluggel M. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics* **2004**; 4: 619-28.
- [136] Elias JE, Haas W, Faherty BK, Gygi SP. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat Methods* **2005**; 2: 667-75.
- [137] Venable JD, Yates JR, 3rd. Impact of ion trap tandem mass spectra variability on the identification of peptides. *Anal Chem* **2004**; 76: 2928-37.
- [138] Olsen JV, Ong SE, Mann M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol Cell Proteomics* **2004**; 3: 608-14.
- [139] Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res* **2003**; 2: 43-50.
- [140] Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* **2005**; 77: 964-73.

- [141] Taylor JA, Johnson RS. Sequence database searches *via de novo* peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* **1997**; 11: 1067-75.
- [142] Taylor JA, Johnson RS. Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry. *Anal Chem* **2001**; 73: 2594-604.
- [143] Grossmann J, Roos FF, Cieliebak M, *et al.* AUDENS: A Tool for Automated Peptide *de Novo* Sequencing. *J Proteome Res* **2005**; 4: 1768-74.
- [144] Yan B, Pan C, Olman VN, Hettich RL, Xu Y. A graph-theoretic approach for the separation of b and y ions in tandem mass spectra. *Bioinformatics* **2005**; 21: 563-74.
- [145] Ma B, Zhang K, Hendrie C, *et al.* PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* **2003**; 17: 2337-42.
- [146] Alves G, Yu YK. Robust accurate identification of peptides (RAId): deciphering MS2 data using a structured library search with *de novo* based statistics. *Bioinformatics* **2005**; 21: 3726-32.
- [147] Spengler B. *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* **2004**; 15: 703-14.
- [148] Zhang Z. *De novo* peptide sequencing based on a divide-and-conquer algorithm and peptide tandem spectrum simulation. *Anal Chem* **2004**; 76: 6374-83.
- [149] Bruni R, Gianfranceschi G, Koch G. On peptide *de novo* sequencing: a new approach. *J Pept Sci* **2005**; 11: 225-34.
- [150] Lu B, Chen T. A suboptimal algorithm for *de novo* peptide sequencing *via* tandem mass spectrometry. *J Comput Biol* **2003**; 10: 1-12.
- [151] Zhang Z, McElvain JS. *De novo* peptide sequencing by two-dimensional fragment correlation mass spectrometry. *Anal Chem* **2000**; 72: 2337-50.
- [152] Demine R, Walden P. Sequit: software for *de novo* peptide sequencing by matrix-assisted laser desorption/ionization post-source decay mass spectrometry. *Rapid Commun Mass Spectrom* **2004**; 18: 907-13.
- [153] Heredia-Langner A, Cannon WR, Jarman KD, Jarman KH. Sequence optimization as an alternative to *de novo* analysis of tandem mass spectrometry data. *Bioinformatics* **2004**; 20: 2296-304.
- [154] Mann M, Wilm M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* **1994**; 66: 4390-9.
- [155] Frank A, Tanner S, Bafna V, Pevzner P. Peptide sequence tags for fast database search in mass-spectrometry. *J Proteome Res* **2005**; 4: 1287-95.
- [156] Clauser KR, Baker P, Burlingame AL. Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem* **1999**; 71: 2871-82.
- [157] Chalkley RJ, Baker PR, Huang L, *et al.* Comprehensive Analysis of a Multidimensional Liquid Chromatography Mass Spectrometry Dataset Acquired on a Quadrupole Selecting, Quadrupole Collision Cell, Time-of-flight Mass Spectrometer: II. New Developments in Protein Prospector Allow for Reliable and Comprehensive Automatic Analysis of Large Datasets. *Mol Cell Proteomics* **2005**; 4: 1194-204.
- [158] Tabb DL, Saraf A, Yates JR, 3rd. GutenTag: high-throughput sequence tagging *via* an empirically derived fragmentation model. *Anal Chem* **2003**; 75: 6415-21.
- [159] Hernandez P, Gras R, Frey J, Appel RD. Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. *Proteomics* **2003**; 3: 870-8.
- [160] Sunyaev S, Liska AJ, Golod A, Shevchenko A. MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal Chem* **2003**; 75: 1307-15.
- [161] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **1988**; 85: 2444-8.
- [162] Mackey AJ, Haystead TA, Pearson WR. Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. *Mol Cell Proteomics* **2002**; 1: 139-47.
- [163] Altschul SF, Madden TL, Schaffer AA, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997**; 25: 3389-402.
- [164] Shevchenko A, Sunyaev S, Loboda A, Bork P, Ens W, Standing KG. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal Chem* **2001**; 73: 1917-26.
- [165] Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol* **1990**; 183: 63-98.
- [166] Habermann B, Oegema J, Sunyaev S, Shevchenko A. The power and the limitations of cross-species protein identification by mass spectrometry-driven sequence similarity searches. *Mol Cell Proteomics* **2004**; 3: 238-49.
- [167] Liska AJ, Sunyaev S, Shilov IN, Schaeffer DA, Shevchenko A. Error-tolerant EST database searches by tandem mass spectrometry and multiTag software. *Proteomics* **2005**; 5: 4118-22.
- [168] Searle BC, Dasari S, Turner M, *et al.* High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS *de novo* sequencing results. *Anal Chem* **2004**; 76: 2220-30.
- [169] Searle BC, Dasari S, Wilmarth PA, *et al.* Identification of protein modifications using MS/MS *de novo* sequencing and the OpenSea alignment algorithm. *J Proteome Res* **2005**; 4: 546-54.
- [170] Tanner S, Shu H, Frank A, *et al.* InsPecT: identification of post-translationally modified peptides from tandem mass spectra. *Anal Chem* **2005**; 77: 4626-39.
- [171] Kayser JP, Vallet JL, Cerny RL. Defining parameters for homology-tolerant database searching. *J Biomol Tech* **2004**; 15: 285-95.
- [172] Han Y, Ma B, Zhang K. SPIDER: software for protein identification from sequence tags with *de novo* sequencing error. *J Bioinform Comput Biol* **2005**; 3: 697-716.
- [173] Halligan BD, Ruotti V, Twigger SN, Greene AS. DeNovoID: a web-based tool for identifying peptides from sequence and mass tags deduced from *de novo* peptide sequencing by mass spectrometry. *Nucleic Acids Res* **2005**; 33: W376-81.
- [174] Halligan BD, Dratz EA, Feng X, Twigger SN, Tonellato PJ, Greene AS. Peptide identification using peptide amino acid attribute vectors. *J Proteome Res* **2004**; 3: 813-20.
- [175] HUPO 4th Annual World Congress, August 29-September 1, 2005, Munich. *Mol Cell Proteomics* **2005**; 4: S1-S414.
- [176] Orchard S, Montecchi-Palazzi L, Hermjakob H, Apweiler R. The use of common ontologies and controlled vocabularies to enable data exchange and deposition for complex proteomic experiments. *Pac Symp Biocomput* **2005**: 186-96.
- [177] Orchard S, Taylor C, Hermjakob H, Zhu W, Julian R, Apweiler R. Current status of proteomic standards development. *Expert Rev Proteomics* **2004**; 1: 179-83.
- [178] Garden P, Alm R, Hakkinen J. PROTEIOS: an open source proteomics initiative. *Bioinformatics* **2005**; 21: 2085-7.
- [179] Pedrioli PG, Eng JK, Hubley R, *et al.* A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* **2004**; 22: 1459-66.
- [180] Hao P, He WZ, Huang Y, *et al.* MPSS: an integrated database system for surveying a set of proteins. *Bioinformatics* **2005**; 21: 2142-3.
- [181] Jones P, Vinod N, Down T, *et al.* Dasty and UniProt DAS: a perfect pair for protein feature visualization. *Bioinformatics* **2005**; 21: 3198-9.
- [182] Kohli BM, Eng JK, Nitsch RM, Konietzko U. An alternative sampling algorithm for use in liquid chromatography/tandem mass spectrometry experiments. *Rapid Commun Mass Spectrom* **2005**; 19: 589-96.
- [183] Chalkley RJ, Baker PR, Hansen KC, *et al.* Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: I. How much of the data is theoretically interpretable by search engines? *Mol Cell Proteomics* **2005**; 4: 1189-93.
- [184] Bern M, Goldberg D, McDonald WH, Yates JR, 3rd. Automatic quality assessment of Peptide tandem mass spectra. *Bioinformatics* **2004**; 20(Suppl 1): I49-I54.
- [185] Beer I, Barnea E, Ziv T, Admon A. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics* **2004**; 4: 950-60.
- [186] Kwon KH, Kim M, Kim JY, *et al.* Efficiency improvement of peptide identification for an organism without complete genome sequence, using expressed sequence tag database and tandem mass spectral data. *Proteomics* **2003**; 3: 2305-9.

- [187] Colinge J, Cusin I, Reffas S, et al. Experiments in searching small proteins in unannotated large eukaryotic genomes. *J Proteome Res* **2005**; 4: 167-74.
- [188] Matthiesen R, Bunkenborg J, Stensballe A, Jensen ON, Welinder KG, Bauw G. Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. *Proteomics* **2004**; 4: 2583-93.
- [189] Yang X, Dondeti V, Dezube R, et al. DBParser: web-based software for shotgun proteomic data analyses. *J Proteome Res* **2004**; 3: 1002-8.
- [190] Muthusamy B, Hanumanthu G, Suresh S, et al. Plasma Proteome Database as a resource for proteomics research. *Proteomics* **2005**; 5: 3531-6.
- [191] Omenn GS, States DJ, Adamski M, et al. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **2005**; 5: 3226-45.
- [192] Adamski M, Blackwell T, Menon R, et al. Data management and preliminary data analysis in the pilot phase of the HUPO Plasma Proteome Project. *Proteomics* **2005**; 5: 3246-61.
- [193] Martens L, Hermjakob H, Jones P, et al. PRIDE: the proteomics identifications database. *Proteomics* **2005**; 5: 3537-45.
- [194] Craig R, Cortens JP, Beavis RC. The use of proteotypic peptide libraries for protein identification. *Rapid Commun Mass Spectrom* **2005**; 19: 1844-50.
- [195] Liotta LA, Ferrari M, Petricoin E. Clinical proteomics: written in blood. *Nature* **2003**; 425: 905.
- [196] Pang JX, Ginanni N, Dongre AR, Hefta SA, Opitek GJ. Biomarker discovery in urine by proteomics. *J Proteome Res* **2002**; 1: 161-9.
- [197] Ping P, Vondriska TM, Creighton CJ, et al. A functional annotation of subproteomes in human plasma. *Proteomics* **2005**; 5: 3506-19.
- [198] Geurts P, Fillet M, de Seny D, et al. Proteomic mass spectra classification using decision tree based ensemble methods. *Bioinformatics* **2005**; 21: 3138-45.
- [199] Yu J, Chen XW. Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data. *Bioinformatics* **2005**; 21(Suppl 1): i487-i94.
- [200] Tibshirani R, Hastie T, Narasimhan B, et al. Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics* **2004**; 20: 3034-44.
- [201] Bryant DK, Monte S, Man WJ, et al. Principal component analysis of mass spectra of peptides generated from the tryptic digestion of protein mixtures. *Rapid Commun Mass Spectrom* **2001**; 15: 418-27.
- [202] Bensmail H, Golek J, Moody MM, Semmes JO, Haoudi A. A novel approach for clustering proteomics data using Bayesian fast Fourier transform. *Bioinformatics* **2005**; 21: 2210-24.
- [203] Lancashire L, Schmid O, Shah H, Ball G. Classification of bacterial species from proteomic data using combinatorial approaches incorporating artificial neural networks, cluster analysis and principal components analysis. *Bioinformatics* **2005**; 21: 2191-9.
- [204] Cantin GT, Yates JR, 3rd. Strategies for shotgun identification of post-translational modifications by mass spectrometry. *J Chromatogr A* **2004**; 1053: 7-14.
- [205] MacCoss MJ, McDonald WH, Saraf A, et al. Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc Natl Acad Sci USA* **2002**; 99: 7900-5.
- [206] Reinders J, Sickmann A. State-of-the-art in phosphoproteomics. *Proteomics* **2005**; 5: 4052-61.
- [207] Zhong H, Zhang Y, Wen Z, Li L. Protein sequencing by mass analysis of polypeptide ladders after controlled protein hydrolysis. *Nat Biotechnol* **2004**; 22: 1291-6.
- [208] Steen H, Jebaranirajah JA, Rush J, Morrice N, Kirschner MW. Phosphorylation analysis by mass spectrometry: Myths, facts and the consequences for qualitative and quantitative measurements. *Mol Cell Proteomics* **2005**.
- [209] Farriol-Mathis N, Garavelli JS, Boeckmann B, et al. Annotation of post-translational modifications in the Swiss-Prot knowledge base. *Proteomics* **2004**; 4: 1537-50.
- [210] Hansen BT, Davey SW, Ham AJ, Liebler DC. P-Mod: an algorithm and software to map modifications to peptide sequences using tandem MS data. *J Proteome Res* **2005**; 4: 358-68.
- [211] Puntervoll P, Linding R, Gemund C, et al. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res* **2003**; 31: 3625-30.
- [212] Plewczynski D, Tkacz A, Wyrwicz LS, Rychlewski L. AutoMotif server: prediction of single residue post-translational modifications in proteins. *Bioinformatics* **2005**; 21: 2525-7.
- [213] Creasy DM, Cottrell JS. Unimod: protein modifications for mass spectrometry. *Proteomics* **2004**; 4: 1534-6.
- [214] Garavelli JS. The RESID Database of Protein Modifications: 2003 developments. *Nucleic Acids Res* **2003**; 31: 499-501.
- [215] Taylor GK, Kim YB, Forbes AJ, Meng F, McCarthy R, Kelleher NL. Web and database software for identification of intact proteins using "top down" mass spectrometry. *Anal Chem* **2003**; 75: 4081-6.
- [216] Gattiker A, Bienvenut WV, Bairoch A, Gasteiger E. FindPept, a tool to identify unmatched masses in peptide mass fingerprinting protein identification. *Proteomics* **2002**; 2: 1435-44.
- [217] Thiede B, Lamer S, Mattow J, et al. Analysis of missed cleavage sites, tryptophan oxidation and N-terminal pyroglutamylation after in-gel tryptic digestion. *Rapid Commun Mass Spectrom* **2000**; 14: 496-502.
- [218] Halligan BD, Ruotti V, Jin W, Laffoon S, Twigger SN, Dratz EA. ProMoST (Protein Modification Screening Tool): a web-based tool for mapping protein modifications on two-dimensional gels. *Nucleic Acids Res* **2004**; 32: W638-44.
- [219] Kumar Y, Khachane A, Belwal M, Das S, Somsundaram K, Tatu U. ProteoMod: a new tool to quantitate protein post-translational modifications. *Proteomics* **2004**; 4: 1672-83.
- [220] Dornon B, Costello CE. A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. *Glycoconjugate* **1988**; 5: 397-409.
- [221] Cooper CA, Joshi HJ, Harrison MJ, Wilkins MR, Packer NH. GlycoSuiteDB: a curated relational database of glycoprotein glycan structures and their biological sources. 2003 update. *Nucleic Acids Res* **2003**; 31: 511-3.
- [222] Lohmann KK, von der Lieth CW. GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. *Nucleic Acids Res* **2004**; 32: W261-6.
- [223] Tang H, Mechref Y, Novotny MV. Automated interpretation of MS/MS spectra of oligosaccharides. *Bioinformatics* **2005**; 21(Suppl 1): i431-i39.
- [224] Zhou FF, Xue Y, Chen GL, Yao X. GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem Biophys Res Commun* **2004**; 325: 1443-8.
- [225] Zhao Y, Lin YH. The development of an algorithm for the mass spectral interpretation of phosphoproteins. *Proteomics* **2005**; 5: 843-5.
- [226] Ong SE, Mann M. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol* **2005**; 1: 252-62.
- [227] Stewart, II, Thomson T, Figeys D. 18O labeling: a tool for proteomics. *Rapid Commun Mass Spectrom* **2001**; 15: 2456-65.
- [228] Halligan BD, Slyper RY, Twigger SN, Hicks W, Olivier M, Greene AS. ZoomQuant: an application for the quantitation of stable isotope labeled peptides. *J Am Soc Mass Spectrom* **2005**; 16: 302-6.
- [229] Ong SE, Blagoev B, Kratchmarova I, et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **2002**; 1: 376-86.
- [230] Schulze WX, Mann M. A novel proteomic screen for peptide-protein interactions. *J Biol Chem* **2004**; 279: 10756-64.
- [231] Kratchmarova I, Blagoev B, Haack-Sorensen M, Kassem M, Mann M. Mechanism of divergent growth factor effects in mesenchymal stem cell differentiation. *Science* **2005**; 308: 1472-7.
- [232] Wu CC, MacCoss MJ, Howell KE, Matthews DE, Yates JR, 3rd. Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis. *Anal Chem* **2004**; 76: 4951-9.
- [233] MacCoss MJ, Wu CC, Liu H, Sadygov R, Yates JR, 3rd. A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal Chem* **2003**; 75: 6912-21.
- [234] Venable JD, Dong MQ, Wohlschlegel J, Dillin A, Yates JR. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* **2004**; 1: 39-45.
- [235] Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **1999**; 17: 994-9.
- [236] Olsen JV, Andersen JR, Nielsen PA, et al. HysTag--a novel proteomic quantification tool applied to differential display analysis of membrane proteins from distinct areas of mouse brain. *Mol Cell Proteomics* **2004**; 3: 82-92.

- [237] Ross PL, Huang YN, Marchese JN, *et al.* Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **2004**; 3: 1154-69.
- [238] Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc Natl Acad Sci USA* **2003**; 100: 6940-5.
- [239] Havlis J, Shevchenko A. Absolute quantification of proteins in solutions and in polyacrylamide gels by mass spectrometry. *Anal Chem* **2004**; 76: 3029-36.
- [240] Havlis J, Thomas H, Sebela M, Shevchenko A. Fast-response proteomics by accelerated in-gel digestion of proteins. *Anal Chem* **2003**; 75: 1300-6.
- [241] Bogianni S, Curini R, Di Corcia A, Lagana A, Stabile A, Sturchio E. Development of a multiresidue method for analyzing herbicide and fungicide residues in bovine milk based on solid-phase extraction and liquid chromatography-tandem mass spectrometry. *J Chromatogr A* **2005**.
- [242] Benfenati E, Porazzi E, Bagnati R, *et al.* Organic tracers identification as a convenient strategy in industrial landfills monitoring. *Chemosphere* **2003**; 51: 677-83.
- [243] Barnidge DR, Dratz EA, Martin T, Bonilla LE, Moran LB, Lindall A. Absolute quantification of the G protein-coupled receptor rhodopsin by LC/MS/MS using proteolysis product peptides and synthetic peptide standards. *Anal Chem* **2003**; 75: 445-51.
- [244] Chelius D, Bondarenko PV. Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J Proteome Res* **2002**; 1: 317-23.
- [245] Liu H, Sadygov RG, Yates JR, 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **2004**; 76: 4193-201.
- [246] Wang W, Zhou H, Lin H, *et al.* Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem* **2003**; 75: 4818-26.
- [247] Ishihama Y, Oda Y, Tabata T, *et al.* Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* **2005**; 4: 1265-72.
- [248] Rappsilber J, Ryder U, Lamond AI, Mann M. Large-scale proteomic analysis of the human spliceosome. *Genome Res* **2002**; 12: 1231-45.
- [249] Allet N, Barrillat N, Baussant T, *et al.* *In vitro* and *in silico* processes to identify differentially expressed proteins. *Proteomics* **2004**; 4: 2333-51.
- [250] Jain N, Thatte J, Braciale T, Ley K, O'Connell M, Lee JK. Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* **2003**; 19: 1945-51.
- [251] Colinge J, Chiappe D, Lagache S, Moniatte M, Bougueleret L. Differential Proteomics via probabilistic peptide identification scores. *Anal Chem* **2005**; 77: 596-606.
- [252] Radulovic D, Jelveh S, Ryu S, *et al.* Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics* **2004**; 3: 984-97.