INFORMATICS DEVELOPMENT: CHALLENGES AND SOLUTIONS FOR MALDI MASS SPECTROMETRY

David Fenyö[1]* and Ronald C. Beavis[2]*
[1]*The Rockefeller University, 1230 York Avenue, New York, New York 10065*
[2]*University of British Columbia, British Columbia, Canada V6T 1Z3*

Received 31 May 2007; accepted 4 September 2007

Published online 23 October 2007 in Wiley InterScience (www.interscience.wiley.com) DOI 10.1002/mas.20152

Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) has been successfully applied to elucidating biological questions trough the analysis of proteins, peptides, and nucleic acids. Here, we review the different approaches for analyzing the data that is generated by MALDI-MS. The first step in the analysis is the processing of the raw data to find peaks that correspond to the analytes. The peaks are characterized by their areas (or heights) and their centroids. The peak area can be used as a measure of the quantity of the analyte, and the centroid can be used to determine the mass of the analyte. The masses are then compared to models of the analyte, and these models are ranked according to how well they fit the data and their significance is calculated. This allows the determination of the identity (sequence and modifications) of the analytes. We show how this general data analysis workflow is applied to protein and nucleic acid chemistry as well as proteomics. © 2007 Wiley Periodicals, Inc., Mass Spec Rev 27:1–19, 2008

Keywords: *MALDI; data analysis; peptides; proteins; nucleic acids*

I. INTRODUCTION

The development of matrix-assisted laser desorption/ionization (MALDI) (Karas & Hillenkamp, 1988) required the design and construction of novel time-of-flight mass spectrometers that were used to explore the physical effects that allowed MALDI to produce ionized gas-phase macromolecules. Initially, instruments constructed to use laser-desorption or plasma-desorption ion sources were adapted for this new type of ion source. Subsequent instrumental development produced purpose-built analytical instruments that used an understanding of the physical characteristics of the MALDI effect to improve the performance of the devices.

Paralleling the development of new instruments was the development of new informatics techniques for acquiring, storing, analyzing, and displaying the new types of mass spectra being measured. Data acquisition proved to be particularly challenging with the new instruments, because the operating characteristics of MALDI ion sources were very different for either laser-desorption or plasma-desorption systems. MALDI ion sources produced very sporadic output that varied in intensity with each laser shot taken, with integrated signal intensities changing by more than 100-fold. Given the dynamic range of the fast electronic transient recorders available at the time (numbers in the range 0–255), manual intervention was necessary to prevent signal distortions caused by detector signals that exceeded the measurable upper limit of the recording electronics. Developing automated systems to control the acceptance of data, as well as controlling the laser intensity and position on the sample, has been an ongoing process as instruments based on MALDI ion sources migrated from instrument development laboratories into application environments.

The storage and display of MALDI mass spectra also produced a new set of challenges, because of the nature of the measurements being made. Time-of-flight mass spectrometers were a rarity when MALDI was introduced; however, they proved to be ideal for recording MALDI mass spectra. Time-of-flight instruments are suited to ion sources that produce brief pulses of ions, followed by relatively long periods of measurement during which no additional ions can be allowed into the instrument's analyzer. The MALDI ion source produced ions in bursts during the initiating laser pulse; that is, ions were generated for approximately 10 nsec, and it did not produce any subsequent ions until the next laser pulse. Therefore, the detector output recording electronics could be triggered by a pulse generated from the laser's output to give the system a very reproducible ''start'' time. If the various high voltage supplies in the instrument remained constant over the period of a full measurement, then the results of multiple laser shots could be accumulated and summed without having to correct for jitter in the ''start'' time, relative to the initiation and termination of the ion-production effect. Because this situation was very similar to the application of time-of-flight analyzers to ions generated by laser or plasma desorption, the display software used initially for these instruments was applied to MALDI spectra. Subsequent generations of display and manipulation software evolved as the characteristics of the analyzers improved (e.g., increased mass resolution), and as the application of MALDI-derived data became more involved in matching spectra to the details of proposed macromolecular structures. The introduction of tandem mass spectrometers with MALDI ion sources also changed the types of problems that could be addressed with the instruments, resulting in changes in the software requirements to support the handling of the new types of information.

This review outlines a selection of the informatics developments that have been important in the interpretation and the use of

**Correspondence to:* David Fenyö, The Rockefeller University, 1230 York Avenue, New York, New York 10065, fenyo@rockefeller.edu; or Ronald C. Beavis, University of British Columbia, British Columbia, Canada V6T 1Z3, rbeavis@brc.ubc.ca

MALDI mass spectra. A number of Unified Modeling Language (UML) diagrams have been used to illustrate general processes and data structures that represent how elements of an informatics system work: any particular implementation of the described process might differ in detail.

II. DEVELOPMENT OF TIME-OF-FLIGHT DATA SYSTEMS

A. Laser Desorption

Laser desorption time-of-flight mass spectrometers used a pulse of focused laser light to stimulate the emission of ions from a substrate. Unlike MALDI, normal laser desorption substrates were only minimally prepared. For example, a layer of organic molecules of interest or a tissue section would be placed on a metal sample stage. The laser beam would be steered and focused so that the light pulse would deliver the desired irradiance to the organic layer. Ions formed from the surface would be extracted immediately with an electric field. Because the ion emission kinetic energy distribution was nearly thermal, there was minimal need to compensate for this initial distribution in the time-of-flight instrument to obtain good mass resolution.

The ions generated by this type of ion source were normally of precursor ion mass less than 2,000 Da (often $<$500 Da) with a charge $z = 1$. Each laser pulse generated a large number of ions, often generating significant radiation damage on the sample. If the damage was sufficiently large, then the laser would be repositioned to a fresh area of sample for subsequent laser interrogation. Because of this damage, it was necessary for an operator to maintain visual observation of the sample through a microscope, and to move the sample when damage had occurred. Therefore, the data system was designed to provide the operator with a great deal of control in the timing of each laser shot, as well as to provide immediate feedback to the operator regarding the quality of the data generated by each laser shot.

Figure 1 shows a generalized state diagram to illustrate the steps required for the measurement and storage of a laser desorption mass spectrum. It is important to note how many steps require operator intervention. At every stage of the process, the operator has the opportunity to intervene and correct (or bias) the quality of the resulting spectrum by adding or rejecting the transient records of the ions generated by each laser pulse. The requirement for operator intervention also has the effect of making the entire system relatively slow; even though the laser may be able to produce tens of pulses per second, the operator requires 1–10 sec to evaluate the results of each shot and to mechanically transmit commands to the data system.

Each of the recorded signal transients contained sufficient ion current to be a recognizable mass spectrum in its own right. The data from each transient, as well as the spectrum generated from the sum of selected transients, had to be stored to disk and recalled rapidly into memory for display. At the time MALDI was being developed, the computers used to control these instruments had very limited main memories (64 kB or less) and rather slow disk drives. Figure 2 illustrates the information that had to be created in order to record and display an individual transient or a summed mass spectrum.

The structure shown in Figure 2 can be used to understand the storage requirements for any time-of-flight analyzer. Because of the importance of this type of instrument in MALDI's development (and current application), understanding how the data are structured provides some insight into the requirements for related informatics systems. The *Spectrum* contains a

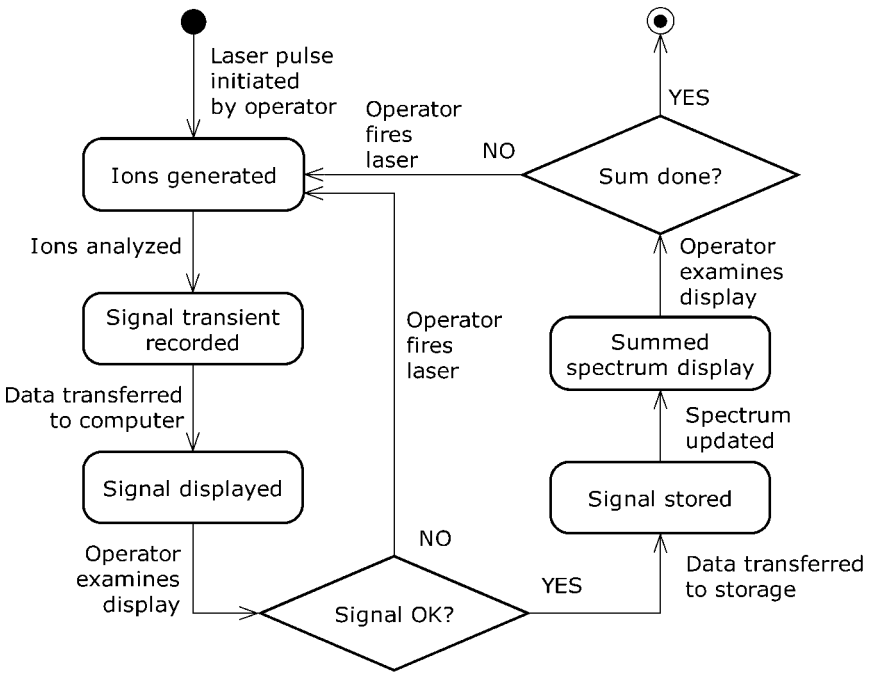

FIGURE 1. General scheme for the operation of a laser desorption mass spectrometer.

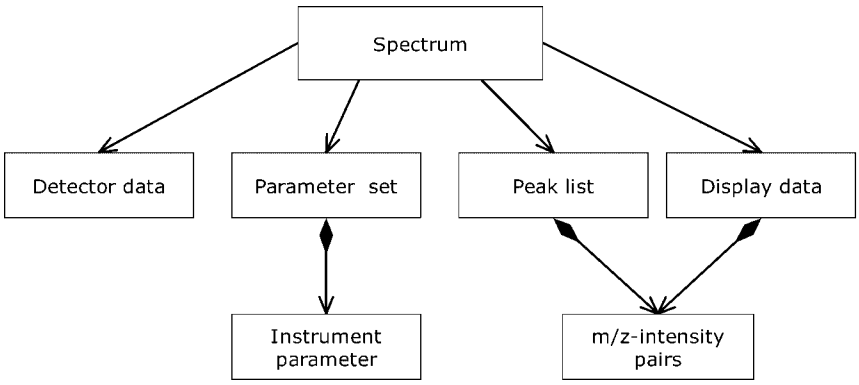

FIGURE 2. General class diagram for the storage of a time-of-flight mass spectrum.

Detector data object, which records all of the intensity versus time-of-flight information produced by the recording electronics. In many implementations, this object is simply a container for the binary large object (BLOB) transferred from the recording electronics to the computer. The internal structure of the BLOB is specified by the electronics manufacturer, either through documentation or an application programming interface (API). Either way, the BLOB is normally stored without modification, as the best record of the original data.

The *Spectrum* object also contains a *Parameter set* object, which records all of the information necessary to interpret the *Detector data* object. The contents of this object will vary depending on instrument design and software requirements. Examples of information stored in this object are the time period corresponding to each data point in the transient, the calibration information for converting a flight time into a molecular mass, and the setting for various important voltages in the ion source and mass analyzer.

The *Peak list* and *Display data* objects are collections of (m/z, I) pairs, where m/z represents the mass-to-charge ratio of a measured time-of-flight and I represents the intensity of the signal that corresponds to that time-of-flight. In both cases, the conversion between the measured time-of-flight (t) and the desired m/z value is performed with information from the *Parameter set* object. These two collections are created for very different purposes, however. The *Display data* object represents the pairs of data points required to generate the current view of the object required by any display software: it should contain approximately as many pairs of values as there are pixels on the x-axis of the output video display. If the object contains too many points, then the display will appear blurry because there will be multiple traces at each pixel. Too many points can also cause unacceptably slow screen updates when the display region is changed by the user. Normally, if the m/z interval corresponds to more than one (m/z, I) pair, then the pair with the highest intensity is chosen; that choice results in a clean display with a y-axis value that does not change as the user zooms in and out on the region that surrounds a particular peak.

The *Peak list* object contains the information that is often of the most value to a user: the (m/z, I) pairs associated with well-defined features (peaks) in the spectrum histogram. Just how peaks are defined and how they are interpreted depends on the instrumental resolution and how the information will be used: further description of how these features are located and annotated will be discussed in later sections.

B. Plasma Desorption

Plasma desorption ion sources used the high-energy particles generated by the spontaneous fission of ^{252}Cf to stimulate the emission of ions from a substrate (Macfarlane & Torgerson, 1976). The sample of interest was normally deposited as a thin layer on a metalized Mylar film and the ^{252}Cf source was placed behind the film, with the sample layer facing away from the ^{252}Cf source. The high-energy fission fragments spontaneously emitted from the ^{252}Cf passed though the Mylar and metal layers, traversed the sample, and deposited energy in a cylindrical track fin collisions with electrons in the sample. A pressure-pulse was created when this energy dissipated in the sample, and this pressure pulse caused formation and ejection of ions from the surface of the sample (Johnson et al., 1989). These ions would be extracted immediately with an electric field.

The ions generated by this type of ion source were normally of precursor ion mass less than 30,000 Da with precursor ion charges that typically ranged from $z = 1–3$; $z = 1$ was the dominant signal. Each fission fragment event generated a small number of ions; each event generated very local radiation damage on the sample. In the course of normal plasma desorption measurements, the radiation damaged area on a sample was a sufficiently small fraction of the total surface area that it could be ignored. Each fission fragment event produced so few ions that no discernable spectrum was obtained until results of thousands of events were summed. Therefore, the data system was designed to provide the operator no control over the data acquisition process: typically, the system was allowed to accumulate data until some spectrum quality threshold was reached; for example, the total number of fission fragment events recorded.

Figure 3 shows a generalized state diagram that illustrates the steps required for the measurement and storage of a plasma desorption mass spectrum. In contrast with the state diagram in Figure 1, there is no operator intervention or feedback required (or desired) for a plasma desorption data system. The fission fragment events occur at random intervals; only a small number

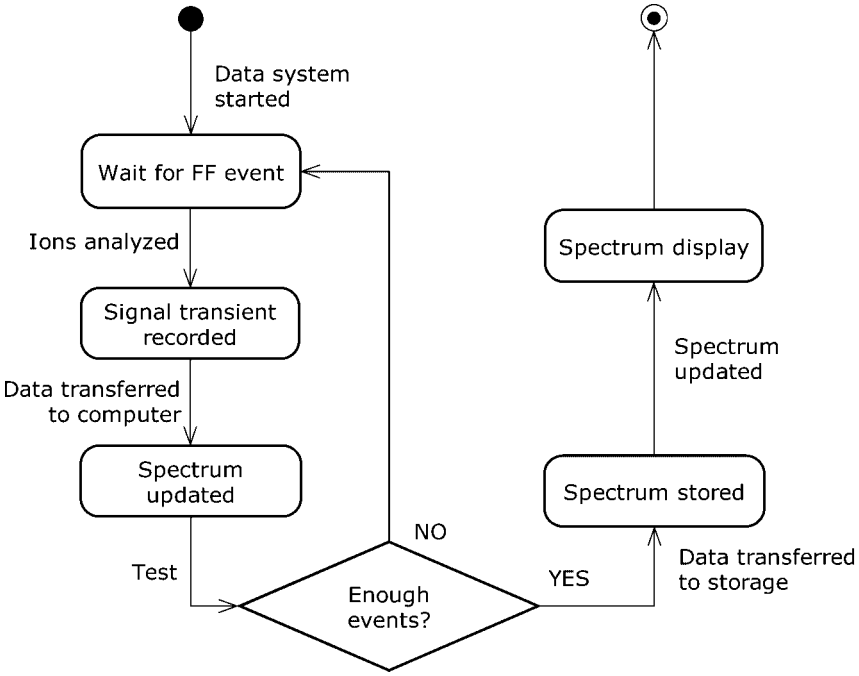

FIGURE 3. General scheme for the operation of a plasma desorption mass spectrometer.

of those that strike the sample layer generated intact precursor ions characteristic of the sample of interest. Because, accumulation of high-quality data required hours (or days), it was necessary to design a system that did not require any user intervention during normal operation. Because most laboratories constructed data systems and instruments themselves, the data systems used were often minimal.

From the time of its discovery in 1975 until the announcement of Tanaka's first MALDI spectra using a liquid matrix in 1987 (Tanaka et al., 1988), plasma desorption was the best ion source for the analysis of molecules with molecular masses greater than 2,000 Da. It was the only ion source that could be used to analyze intact proteins; it required approximately 1 nmol of relatively pure protein for an analysis. Although the upper mass limit was approximately 25,000, the peaks generated from proteins tended to be broadened by meta-stable ion decay during the ion's transit through the mass spectrometer flight tube. There was also significant background noise that required the subtraction of backgrounds that were much higher than the signals corresponding to the high mass analytes' precursor ions.

There are few (if any) active plasma desorption time-of-flight mass spectrometers in use today. The higher sensitivity and ease of operation of MALDI and electrospray ion sources have led to the abandonment of plasma desorption as a routine method to analyze peptides and proteins. From the viewpoint of informatics, its historical importance was that it was the first commercially available desorption-based mass spectrometer that did not require any user intervention during data acquisition. This simplicity of operation led directly to the development of the highly automated MALDI time-of-flight mass spectrometers that are so widely used today.

C. MALDI

The observation of the solid-phase MALDI effect at the University of Muenster (Karas & Hillenkamp, 1988) was made with a commercial prototype laser desorption mass spectrometer. The discoverers had considerable experience with this type of instrumentation, and the informatics design of its data acquisition system (Fig. 1) influenced the first generation of commercial instruments designed to take advantage of the MALDI effect for protein and peptide analysis. Because the MALDI effect could not be reproduced on the production models of the same instrument, it was assumed that it was the detailed construction of the prototype (particularly its ion extraction and laser optics) as well as the skill of the operators that made the effect difficult to replicate.

The first successful repetition of the MALDI effect at Rockefeller University (Beavis & Chait, 1989a,b,c) was made with an ion source design based on a simple ion source very similar to the plasma desorption ion sources in use at the time. The time-of-flight analyzer, detector, and informatics were also very similar to the plasma desorption style of data acquisition (Fig. 3), with only a very few modifications. Rather than firing the laser one shot at a time with an operator-controlled switch, the laser was allowed to run continuously and the data acquisition system took over control of which signal transients to sum together to create a final spectrum. Although the initial system still required a skilled operator to adjust the fluence of the laser, the data system operated without requiring any user intervention.

An example of this type of acquisition system is shown in Figure 4; based on one of the early systems developed at the Memorial University of Newfoundland to operate a MALDI time-of-flight instrument without user intervention. This system

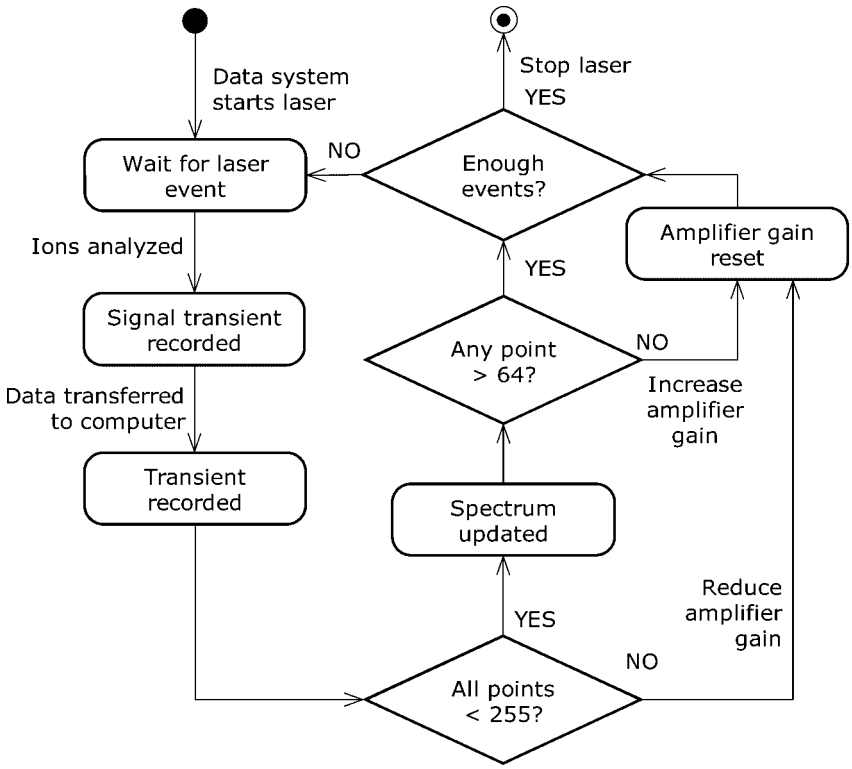

FIGURE 4. A general scheme for the automated operation of a MALDI mass spectrometry.

was a variant of the scheme in Figure 3, with the addition of several decision points that were used to compensate for the combination of the variation in MALDI ion current, shot-to-shot, and the limited dynamic range of the fast transient recorder used to digitize the detector output: it could only record numbers in the range 0–255.

This compensation was achieved by adjusting the gain of an amplifier that was placed between the detector output and the transient recorder input. MALDI signal variation tends to vary relatively slowly pulse-to-pulse: the signal from a region of the target will gradually decrease in intensity as laser shots gradually erode the matrix crystal surface. Therefore, when the most intense peak in a spectrum was low, the amplifier gain was increased to better use the dynamic range of the recorder. Similarly, if the signals were large enough to saturate the recorder, then the tops of intense peaks were flattened; the maximum value for each transient was 255, regardless of the absolute intensity of a particular peak. Because the inclusion of these flat-topped signals would result in significant distortion of any intense peak, it was important that these transients be discarded and the amplifier gain reduced so that the next shot would hopefully be recorded correctly.

Modern time-of-flight MALDI instruments use some variation of the scheme in Figure 4 to control the instrument's acquisition of data. These instruments include a number of different features that are designed to compensate for the gradual decrease in signal intensity as the MALDI crystals are damaged by the desorption laser, as well as the unpredictable changes in intensity that occur when the laser spot is relocated to a previously unirradiated region of the sample. The most common of these features is some form of feedback control of the laser intensity. If the signal intensity is too large, then the laser irradiance is reduced (e.g., by rotating a wheel-type neutral density filter), and if it is too small, then the laser irradiance is increased. Because the MALDI crystals' response to changes in irradiance exhibits considerable hysteresis, any system of this type must include a dampening function that reduces the system's tendency to "over-react" to the intensity fluctuations that might occur only for one or two laser shots when the sample is repositioned.

III. APPLICATIONS FOR PROTEIN CHEMISTRY

A. Intact Molecule Analysis

It was immediately apparent that application of MALDI to protein analysis would be important (Hillenkamp & Karas, 1990; Hillenkamp et al., 1991), after the initial round of discoveries that led to the development of a practical MALDI ion source. Unlike plasma desorption before it, MALDI was able to produce signals with good signal-to-noise ratios for most intact proteins. It was also relatively insensitive to the amino acid composition of a protein: highly basic or highly acidic proteins all produced spectra dominated by $z = +1$ or $+2$ ions. Its main competitor, electrospray ionization (ESI), generated much more complex spectra caused by distributions of ion charge that depended on the

number of basic residues contained in the protein. This ESI spectrum complexity made it very difficult to distinguish multiple protein species in the same mass spectrum, and required very long (sometimes even overnight) calculations to ''deconvolute'' the protein and peptide species present in a sample. In the simpler MALDI spectra, this task could easily be done by visual inspection. MALDI also was superior to electrospray in several practically important features: the ability to analyze mixtures of proteins and the ability to obtain useful spectra from samples that contained modest concentrations of buffer salts. These practical differences between the two methods led to MALDI being the method of choice to rapidly determine the molecular mass of intact proteins.

In addition to these positive features, MALDI instruments also had some limitations when used for this purpose. MALDI time-of-flight spectra of intact proteins with molecular masses greater than 20 kDa have rather modest mass resolution ($m/\Delta m$ in the range 100–1,000) compared to what can be achieved for peptides ($m/\Delta m > 10{,}000$). Therefore, the molecular masses that can be assigned to intact proteins are chemical masses, rather than the monoisotopic, all-^{12}C masses that are normally assigned to peptides. The limited resolution also reduces the confidence that can be placed in the mass determination: if the resolution is insufficient to distinguish an $(M+H)^+$ from an $(M+Na)^+$ (or $M+$ matrix) adduct peaks, then the mass determination is often biased by the contribution of these adducts to the calculated peak centroid.

This resolution-related problem required the design of specialized peak-assignment algorithms to distinguish between cases where all ^{12}C masses could be assigned and those where the spectrum was only sufficient to determine a chemical mass for a peak. A simple algorithm for performing this task is illustrated in Figure 5. Once the spectrum is in memory, the data are modified to artificially reduce the mass resolution of the spectrum for the mass range in which A_0 peaks can be resolved. This operation can be done by averaging an appropriate number of adjacent records, fitting to theoretical envelop distributions, or any other means that does not distort the resulting broadened peaks. It is a simple matter to identify all of the broadened peaks present. The obtained (m/z, I) pairs can be further filtered by applying a quality control metric to each peak, such as signal-to-noise ratio (s/n). With the chemical mass of each of the peaks stored, the original mass spectrum can be reloaded and the region near the chemical masses tested for resolved A_0 peaks. If they exist, then they can be stored with the chemical masses.

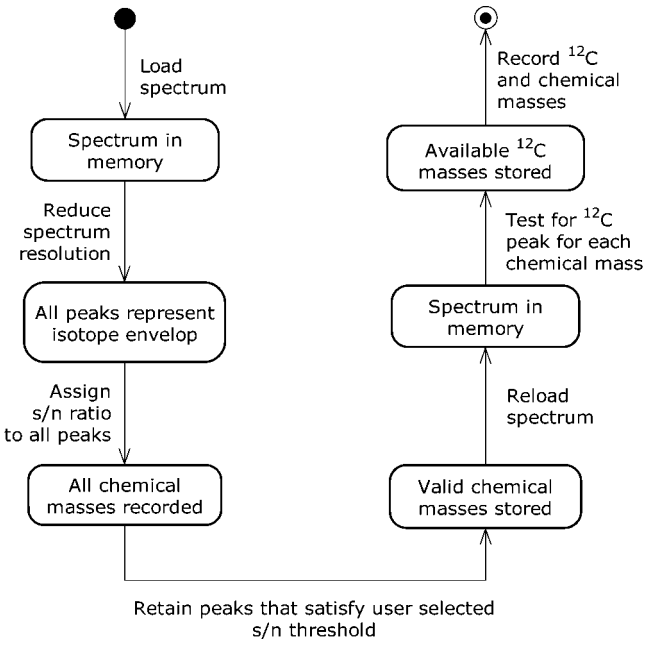

FIGURE 5. Simple scheme for finding all available ^{12}C and chemical mass peaks.

B. Protein Mixtures/Biological Fluids

The most important step in MALDI sample preparation is the growth of the protein-doped crystals of an appropriate matrix material, such as sinapic acid, alpha-cyano-4-hydroxy-cinnamic acid (Beavis & Chait, 1989a), or 2,5-dihydroxy-benzoic acid (Tsarbopoulos et al., 1994). This sample preparation is normally performed by drying a small aliquot of a solution that contained the proteins of interest and a nearly saturated solution of the matrix. As the solution dries, the solution becomes supersaturated with matrix; crystals of the material quickly form and trap protein molecules in the crystals as they grow. Once all of the solvent has evaporated, a thin layer of protein-doped crystals is left on the substrate, along with crystals composed of any other salts that were present in the original solution. By washing the substrate with a solvent that dissolves the salt crystals but leaves the matrix crystals intact, it is possible to remove the majority of nonmatrix crystals from the substrate, as well as to remove any salt crystals that might be adhered to the matrix crystals. So long as the solution did not contain any compounds that would interfere with the ability of the growing matrix crystals to incorporate proteins, the resulting doped crystals will contain a reasonable sampling of all of the proteins present in the original solution, even if the proteins were a relatively minor component of that solution.

The simple fact that MALDI's sample preparation method separates proteins from other components of a solution makes MALDI an ideal method to analyze the protein complement of biological fluids, such as milk, plasma, or urine (Beavis & Chait, 1990). Good-quality data on the intact molecular mass of these natural source proteins can be obtained with little, if any, additional sample preparation. Although this feature of MALDI was noticed very soon after the discovery of the cinnamic acid matrices, it was left largely unexploited for almost 10 years, because most MALDI research became focused on protein identification by analyzing peptides.

Interest in the analysis of biological fluids was rekindled by the application of MALDI to the discovery of protein biomarkers of disease (Petricoin et al., 2002). The experiment was simple:

1. prepare MALDI samples from the plasma of a cohort of normal individuals;
2. prepare MALDI samples from the plasma of a cohort of individuals that suffered from a particular disease;
3. obtain MALDI spectra from each sample set; and

4. perform statistical analysis on the peaks found in the two patient cohorts and determine which peaks are significantly different between the two groups.

Even though the experimental protocol is simple, this approach to biomarker discovery has been very controversial (Baggerly, Coombes, & Morris, 2005). Although it is possible to observe significant differences between the disease and normal cohorts, there is a high degree of variability in the signals observed within a cohort. Hierarchical clustering has been used to attempt to compensate for the variability within a cohort and to detect differences between cohorts. This approach has suffered from relatively naïve application of clustering methods, and has resulted in claims for the usefulness of this approach that have not been widely accepted by the biomarker community. Various technical problems in the execution of this method have also made the results unnecessarily difficult to interpret, such as the use of very low resolution time-of-flight mass spectrometers, poorly characterized proprietary affinity reagents, and a lack of adequate techniques to determine the identity of the molecular species that give rise to individual signals.

A much less controversial and far more useful technique to directly sample proteins in biological fluids is the use of ''capture'' columns that use high affinity antibody-coated beads to harvest a protein (or proteins) of interest from the fluid (Kiernan et al., 2002). This technique also uses a very simple protocol:

1. pass a biological fluid through a small bed of affinity capture beads confined in a pipette tip;
2. wash the column to remove any contaminants;
3. elute the analytes of interest in a small volume of MALDI matrix solution; and
4. dry the eluant and perform MALDI analysis.

The implementation of this approach has number of significant advantages over the simple method described above. The identity of the analyte proteins can be easily inferred from a combination of the capture affinity reagent specificity and protein precursor ion mass measurements. Because this method is compatible with the use of high-accuracy, high-resolution time-of-flight analyzers, many of the difficulties caused by peak congestion in the simpler process have been resolved.

C. Enzymatic Cleavage Time Course Experiments

As mentioned above, one of the true strengths of MALDI has been its ability to produce simultaneous signals from most of the components of a mixture of analytes without any extra effort or equipment. Given this capability, one of the early applications of MALDI to protein chemistry was to follow the dynamics of the digestion of a pure protein sample by a proteolytic enzyme (Severinov et al., 1994; Cohen et al., 1995). A protein that is properly folded into its native tertiary structure tends to be quite resistance to enzymatic proteolysis: most of the bonds that are vulnerable to a particular protease are hindered from fitting into the enzyme's active site either by being on the inside of the folded structure or by being held tightly in place by hydrogen bonds or salt bridges.

This hindrance to cleavage does not imply, however, that no cleavage will take place. Some peptide bonds are more susceptible to cleavage than others; for example, bonds that are not contained in an organized part of the folded structure, or bonds that are associated with relatively flexible portions of the structure. By coincidence, the use of X-ray crystallography techniques to study the three-dimensional structure of proteins requires that the protein sequences used to form crystals must have a minimum amount of solution-phase flexibility; flexible regions of a protein sequence will cause poor crystal formation and low-resolution protein structures. In the language of protein crystallographers, flexible regions of a sequence are commonly referred to as ''floppy domains;'' whereas tightly folded, inflexible regions are called ''core domains.'' The success of any protein crystallization experiment often depends on how well a crystallographer can predict which portions of a sequence will contain floppy domains, remove them from the recombinant protein, and leave only the core domains. Often, the core domains are synthesized and their structures solved individually.

The difficulty of attempting to *a priori* predict ''floppy'' versus ''core'' domains had led crystallographers to use proteolytic digestion as an experimental tool to attempt to distinguish disordered or flexible regions of a sequence. However, prior to MALDI, it was difficult to analyze the results: HPLC separation followed by either Edman sequencing or amino acid analysis was the only way to determine which bonds were being attacked first by the protease. These techniques were slow and expensive.

MALDI, by contrast, could determine the masses of all of the peptides present in a proteolysis time-course fraction with a single measurement within a few minutes. The only difficulty was that only the peptide's masses were available: no amino acid sequence information could be obtained—at least using the instruments available at the time. Therefore, algorithms had to be developed to match an observed set of peptide masses with the known sequence of the recombinant full-length protein being studied. Given the rather modest mass accuracies available, this task was complicated by the fact that, in a 30 kDa protein, there are approximately 270 amino acid residues. Assuming that a sequence-specific protease was used for the cleavage experiment (for example, trypsin), there would be on average 30 cleavage sites in the protein sequence and thousands of peptides that could possibly be generated. Due to the repetitive nature of many protein sequences and to the many redundancies in the possible peptide masses, it was often not possible to simply assign a single peptide sequence unambiguously to an observed mass. Instead, a small set of peptide amino acid sequences were often equally likely assignments to a particular mass.

In fortunate cases, this ambiguity was unimportant if the most telling features of a spectrum could be assigned. More commonly, the most potentially important peaks in a spectrum were those associated with this type of ambiguity. A set of algorithms was formulated based on observations of the emergence of structure-dependant patterns of proteolytic cleavage. These algorithms attempted to mimic the process by which an experienced protein chemist would solve simple problems that involved a choice between small sets of possible peptide sequences; given how they could be assembled onto the full protein sequences. For example, given a set of peptides and a known protein sequence, what subset of those peptides best answers the following questions with ''yes:''

1. Is the difference between the experimental and calculated masses a minimum?
2. Is the number of possible internal cleavage sites a minimum?
3. Is one of the peptide's termini the N- or C-terminus of the complete sequence?
4. Is the peptide's C-terminus shared with another observed peptide?
5. Is the peptide's N-terminus shared with another observed peptide?
6. In the peptide's N-terminus adjacent to the C-terminus of other observed peptides?
7. Is the peptide's C-terminus adjacent to the N-terminus of other observed peptides?

With the problem formulated in this manner, it was possible to utilize existing formal methods to solve the problem; for example, fuzzy logic or Bayesian probability (Zhang & Chait, 2000). Fortunately, this type of calculation can be solved in polynomial time (i.e., time to perform that calculation is no greater than a polynomial function of the number of proteins), so no special computer configurations or approximations were necessary to determine the most likely solutions to the problem.

Another common class of problems in primary amino acid sequence determination is the assignment of protein inter- and intra-molecular chemical cross-linkage sites. In this type of experiment, a cross-linked collection of purified molecules is treated with a amino acid-specific protease (e.g., trypsin), and the digestion allowed to proceed to completion. The masses of all of the resulting peptides can be determined in a single MALDI spectrum. Peptides that did not participate in a cross-linking reaction can be easily identified because they correspond to simple cleavage products of the known precursor molecules. The informatics task is to explain any remaining peptide masses based on the known chemistry of the cross-linking reagent and the precursor proteins' amino acid sequences (Fenyo, 1997; Trester-Zedlitz et al., 2003).

This problem cannot be completely solved in polynomial time. The problem to determine which peptides add together to form a molecule with approximately the same mass as one of the unexplained observed masses is equivalent to one of the famous "NP-complete" problems the "knapsack problem." In this problem, an analyst is given a knapsack and a collection of stones of various weights with assigned values, and is asked to determine which subset of these stones, when placed in the knapsack, will have a weight closest to, but not exceeding, an arbitrary weight limit. The only solution to this problem in general is to test every possible combination of stones and to determine which combination best fits the given criteria. This fact implies that the required number of calculations for N stones will be of the order $O(2^N)$, where O is the Landau notation used to describe the asymptotic behavior of functions. If one considers the possible peptide masses to be the stones, then a value that corresponds to a "goodness-of-fit" and the observed mass to be the weight limit; clearly, the problem of solving which subset of the available peptides can be cross-linked together to conform to an observation is formally the same problem. Although some efforts have been made to formulate approximate solutions to this problem, the solution of any particular problem might be practically intractable to calculate, particularly if the observed peptide mass is greater than 10 kDa.

D. Detecting Post-Translational Modifications

Proteins might be thought of as composed of polypeptides that have been translated from messenger RNA (mRNA) that was transcribed from a gene (i.e., chromosomal DNA). After a polypeptide has been created by translation of mRNA, its chemical structure is almost always modified by other enzymes produced by the cell for this purpose. Often, the polypeptide as translated will have limited (or no) biological activity: chemical modification is required in order to give the molecule its active properties. These modifications might be made immediately upon translation or subsequently, when the polypeptide has been moved into a compartment where its full activity is required. The activity of even active proteins might be modulated by the transitory modification of the protein's sequence to allowed the organism to "switch" the activity of an enzyme or transport molecule "on" or "off."

Any modification to a protein's chemical structure made following translation is commonly referred to as a "post-translational modification" (PTM). Many different types of PTMs have been discovered, and many different biological functions have been ascribed to these modifications. Table 1 lists several examples of these modifications that have proven to be particularly

TABLE 1. Examples of Post-Translational Modifications That can be Detected by Mass Spectrometry

Modification	Residue affected	Function
phosphorylation	serine, threonine	Structural modification, signaling
phosphorylation	tyrosine	Signaling
acetylation	N-terminus	Affect protein half-life in cell
methylation	lysine, arginine	Control chromosome remodeling
ubiquitinylation	Lysine	Mark protein for degradation
peptide bond cleavage	Any	Removal of pre- or pro-protein regions

important biologically, and which can be readily detected with mass spectrometric-based approaches.

The methods to detect these modifications and to locate the site of post-translational modification are simple variants of the methods that follow the consequences of enzymatic cleavage. For example, if a purified sample of a protein has been made available, then the molecular mass of the protein can be determined by MALDI. If the molecular mass determined experimentally is different from what would be expected from its known amino acid sequence, then a post-translational modification has occurred. To determine the site and type of modification, the protein sample can be digested with a amino acid-specific enzyme and the time course followed. Using the techniques described above, the peptide masses obtained by sampling the digest with MALDI can be assigned, and any peaks that cannot be assigned are marked for later analysis. Using this approach, it is often apparent by simple inspection which regions of amino acid sequence are present in the expected, unmodified form, and which sets of peaks can be explained by consistent mass shifts to the known sequence. By comparing these mass shifts with tables of known post-translational modifications (e.g., UniMod (Creasy & Cottrell, 2004)), a hypothesis about the site and type of modification can be formed. This hypothesis can be tested by performing the same experiment with a different proteolytic enzyme, and by looking for consistent mass shifts from the same region of protein sequence. If further confirmation is necessary, then tandem mass spectrometry can be used on peptides suspected of containing modified residues to confirm the site and mass shift associated with the hypothesized modification.

IV. APPLICATIONS FOR NUCLEIC ACID CHEMISTRY

A. Intact Molecule Analysis

The use of MALDI for the analysis of nucleic acids (Nordhoff et al., 1992) was demonstrated soon after its introduction for the analysis of proteins and peptides. However, the development of methods for the mass spectrometric analysis of nucleic acids has been slower and much less successful compared to the widely used methods for the analysis of proteins and peptides. One reason that the analysis of nucleic acids by mass spectrometry is inherently more difficult than protein and peptide analysis is because of the differences in the physical chemical properties of their respective building blocks and the different strength of their polymer bonds. Another reason is that any mass spectrometric method for nucleic acid analysis must compete with the extremely successful molecular biology methods that are available for DNA and RNA analysis.

The limitations of mass spectrometric analysis of nucleic acids include: gas phase fragmentation and depurination of nucleic acids caused by the internal energy transferred to the nucleic acids during the desorption and ionization process; and adduct formation with sodium and potassium due to the attraction to the negatively charge phosphate groups. A large effort has been put into method development to overcome these limitations. For example, matrices like 3-hydroxypicolinic acid (Wu, Steding, & Becker, 1993) minimize fragmentation by producing gas-phase nucleic acids with lower internal energy, and by adding ammonium ions (Pieles et al., 1993) to minimize sodium and potassium adduct formation. In spite of method-development efforts like these, the practical upper limit for analysis is about 50 bases.

One advantage in mass spectrometric analysis of nucleic acids over proteins and peptides is the very limited variation in desorption and ionization efficiencies between different sequences because they are composed of only four different nucleotide bases that have similar properties. In contrast, proteins and peptides are composed of 20 different amino acids with large differences in the properties of individual amino acids; accurate quantitation is possible only with internal standards. The similar properties of nucleic acids make quantitation much more amenable. Another advantage with nucleic acids is that PCR amplification can be used when targeted analysis is done to increase the amount of material that is available for MALDI-TOF analysis; one can analyze very small amounts of material in complex mixtures.

The data analysis for intact nucleic acid mass spectra is similar to that of intact protein mass spectra: The spectra are dominated by peaks corresponding to $z = +1$. The peaks are, however, much wider due to more adduct formation and fragmentation; the measurement of the intact nucleic acid mass at high accuracy is prevented. The combination of low mass accuracy and the fact that there are only four different building blocks of DNA: adenine (A, 135 Da), cytosine (C, 111 Da), guanine (G, 151 Da) and thymine (T, 126 Da) all within a narrow range of masses, allows many different nucleotide base compositions to fit within the measurements of a mass measurement. Therefore, mass measurement of intact nucleic acids can be used only to detect mutations in combination with molecular biology methods.

The advantage of using MALDI-TOF to detect nucleic acids is, however, that many nucleotides can be detected simultaneously; thus it is amenable to multiplexing. By using clever experimental designs that combine molecular biology methods with detection of nucleic acids with MALDI-TOF, it is possible to develop efficient methods for mutation detection and analysis. The most successful so far among these methods is high-throughput Single Nucleotide Polymorphisms (SNP) genotyping analysis. Other applications include elucidation of DNA methylation patterns, sequencing, and expression profiling.

The data analysis for all nucleic acid applications is performed in a similar manner (Fig. 6):

1. the mass spectra are analyzed by determining the mass and intensity of the peaks that correspond to the nucleotides;
2. the masses are used to identify which nucleotides (out of a small set determined by the experimental design) are present in the sample;
3. the intensities of the peaks are used as a measure the quantity of the nucleotides;
4. from the identity and quantity of the nucleotides and the experimental design, conclusions are drawn regarding the nucleotide sequences in the sample.

The different steps in the mass spectrum analysis workflow are background subtraction, peak detection, and quantitation. First, the low-frequency background is removed by fitting a

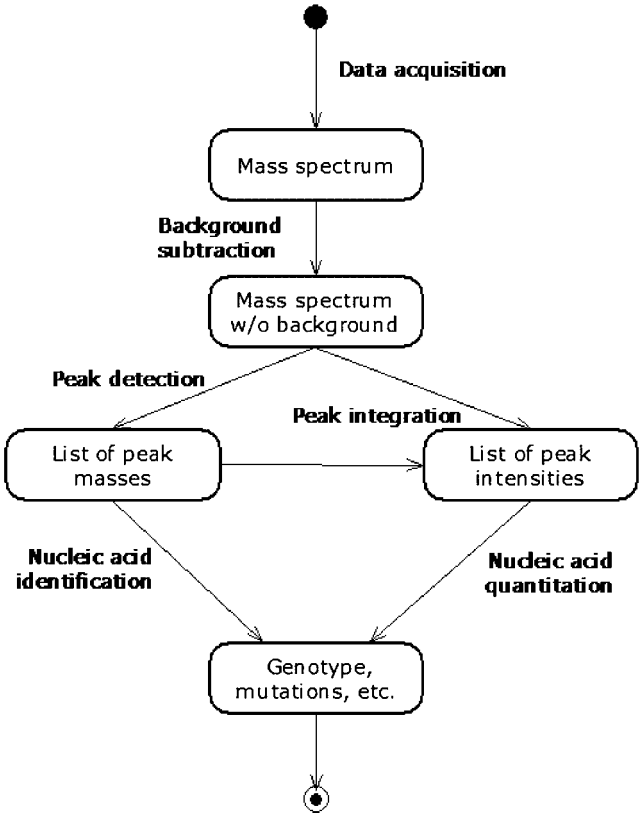

FIGURE 6. The steps in the analysis of mass spectra of nucleic acids are: background subtraction, peak detection, and quantitation. The mass and intensity of the peaks are calculated. The mass is used to identify the nucleic acid from a short list of possibilities determined by the molecular biology performed during the experiment. The intensity is taken as a measure of nucleic acid quantity. The identity and the quantity of the nucleic acids in the sample are used to determine nucleic acid properties, including genotype and mutations.

smooth curve to the regions of the mass spectrum where there are no peaks. This smoothing can, for example, be achieved by applying a very wide and strong smoothing function to the entire spectrum, which will result in a smooth function slightly higher than the background. Subsequently, points in the original spectrum that are far away from this smooth curve are removed (i.e., the peaks), and the smoothing procedure is repeated, this time without including the peaks, to produce a smooth function that will closely follow the background of the spectrum.

After the background has been removed, the masses of the nucleic acids that give rise to the peaks are determined by detecting the peak centroid locations. Usually, two parameters are used to detect the centroid of the peaks: the approximate peak width and the signal-to-noise ratio threshold. The detection is performed by scanning the entire mass spectrum for peaks of the specified width. The signal-to-noise ratio is calculated by taking the ratio of the root mean square (RMS) of the intensities over the peak and the RMS of the intensities in a region close by where there are no peaks. For most nucleic acid applications, the mass of the peaks of interest are known in advance; that knowledge can be used in a more directed approach where only the small changes in peak location caused by variations in the experimental conditions must be detected.

Following the mass determination, the quantity of the nucleic acids is measured by calculating the height or the area of the peak. Careful background subtraction is essential for accurate determination of the height and the area of peaks. The advantage of taking the height of the peak as the measure of quantity is the simplicity and robustness of its calculation; usually, the average height for a few points around the centroid is used. The peak height is a good measure of nucleic acid quantity if the with of the peak does not vary much between mass spectra. In contrast, when there is substantial variation in the peak width between mass spectra, the peak area is a better measurement of quantity. The difficulty in calculating the peak area is in deciding where the peak ends and the background starts; especially for peaks with long tails this determination can be very challenging. One way of circumventing this problem is to select a function and fit its parameters (e.g., centroid, width, skewness, etc.) to the peak and integrate the function. For a majority of nucleic acid applications of MALDI-TOF, the peak height is a robust measure of nucleic acid quantity.

B. Genotyping Single Nucleotide Polymorphisms

SNP genotyping (Haff & Smirnov, 1997; Little et al., 1997) analysis is currently the most successful application of MALDI-TOF to nucleic acid analysis. In the first step, PCR is used to amplify the templates of interest to provide sensitivity and specificity; that is, PCR allows analysis even when there is only a very small amount of the compound with nucleic acid sequence of interest in a very complex mixture. After template identification, primers complementary to sequences close to the SNP's are added. These primers are extended in the presence of three deoxynucleotides and one dideoxynucleotide. The incorporation of the dideoxynucleotide will terminate the primer extension; therefore, the length of the extended primer will depend on template sequence, and different SNP's will give rise to extended primer of varying length. The dideoxynucleotide is selected to maximize the mass difference of the extended primers. The mixture of extended primers is subsequently analyzed by MALDI-TOF. The generated mass spectra are analyzed by determining the mass and intensity of the peaks that correspond to the nucleotides. These masses are used to identify which of the different extended primers are present in the sample, and the intensities of the peaks are used as a measure their quantity. From the identity and quantity of the extended primers, the quantity of the different SNP's can be inferred.

High-throughput SNP genotyping analysis can be achieved by automating many steps of this process. The template amplification, primer extension, and spotting the samples on the MALDI plate can be automated with simple robotics. The MALDI-TOF analysis can also be fully automated with commercial mass spectrometers, and the data analysis is straightforward and does not require any sophisticated algorithms.

C. Other Nucleic Acid Applications

Other applications that are being developed include elucidation of DNA methylation patterns, sequencing, and expression profiling.

The initial attempts at analysis of DNA methylation patterns by MALDI-TOF show some promise (Tost et al., 2003). The methylation patterns are analyzed by converting the nonmethylated cytosines to uracil prior to amplification and by utilizing the multiplexing capabilities of MALDI-TOF to simultaneously determine the methylation state at many locations.

Sequencing with the dideoxy and exonuclease methods in combination with MALDI-TOF detection has been demonstrated (Hahner et al., 1997) but has found limited practical use. In contrast, the use of MALDI-TOF for mutation detection by resequencing holds some promise. For example, mutation-specific nucleic acid fragments for MALDI-TOF analysis can be generated with a combination of PCR, transcription into RNA, and cleavage with nucleotide sequence-specific RNases.

A few attempts have also been made to use MALDI-TOF for expression profiling (Ding & Cantor, 2003). It has been demonstrated that a very wide dynamic range can be achieved, and that detection down to a single molecule is possible. However, a great deal of method development will be necessary for it to successfully compete with microarray analysis.

For all these nucleic acid applications, the analysis of the MALDI-TOF mass spectra is basically identical. The mass and intensity of the peaks are calculated. The mass is used to identify the nucleic acid from a short list of possibilities determined by the molecular biology performed during the experiment. The intensity is taken as a measure of nucleic acid quantity. Finally, the identity and the quantity of the nucleic acids in the sample are used to determine nucleic acid properties, including genotype and mutations.

V. APPLICATIONS FOR PROTEOMICS

A. Peptide Mass Fingerprinting

One of the most persistent, oft-repeated justifications for research into bio-molecular mass spectrometry has been the hope that mass spectrometry-based technology could be used to directly sequence proteins and other biopolymers. Despite several decades of research, this goal has remained elusive. A combination of effects have frustrated these efforts, and range from sequence-dependent variations in the gas- and liquid-phase properties of polypeptides to the intractability of DNA to chemical sequencing methods. Although significant progress has been made and some very impressive applications of mass spectrometry to difficult sequencing problems have been published, there is still no general, easy-to-apply method to determine a high-quality polypeptide sequence without reference to some list of known peptide sequences. Determining a polymer's sequence from a mass spectrum alone is often referred to as *de novo* sequencing.

Fortunately, *de novo* sequencing methods are not the only possible application of mass spectrometry to the problem to determine the amino acid sequence of a protein. Henzel et al. (1993) at Genentech realized in 1993 that it was possible to use one of the strengths of solid-phase MALDI ion sources to indirectly determine the amino acid sequence of a protein. The particular strength that they exploited was the fact that MALDI could generate a representative set of ion signals from all of the peptides present in a complicated mixture of different peptide species. The alternative ion-source technology, electrospray ionization, would often selectively ionize only a few of the peptide species present in a mixture, even though it could ionize each of the peptides if they were present as pure compounds. In contrast, with very little effort it was possible to produce observable MALDI signals from most, if not all, of the peptides present in very complex mixtures. The signals for each peptide species varied considerably in intensity, and depended on solubility of the peptides and how they were incorporated into the matrix crystals during sample preparation; however, the signals were still simultaneously detectable.

Henzel et al. (1993) postulated that, if one had a pure preparation of a protein and treated the protein with a amino acid-specific protease, then the resulting mixture of peptides should have a characteristic pattern of molecular masses. If one took a list of all known protein sequences, then it should be straightforward to calculate the sets of theoretical peptide masses that corresponded to each of protein sequence, given the known amino acid-specificity of the chosen protease. The experimental set of masses, $\mathbf{E}$, could be compared against the theoretical mass set, $\mathbf{T_p}$, for each possible protein sequences, $\mathbf{p}$. Using a simple scoring system, it would be possible to determine which $\mathbf{T_p}$ is the best-fit with $\mathbf{E}$, with a score, $s_p = \mathbf{E} \otimes \mathbf{T_p}$. Assuming that some statistically valid criteria could be established, it should be possible to determine whether this score s_p is sufficient to establish a valid correlation between the experimental spectrum and the known protein sequence taken from the list.

This hypothesis was formulated in this manner because the research team at Genentech intuitively understood what was possible in terms of protein analytical chemistry at the time. The preparation of pure protein fractions from complex biological mixtures was by then a routine application of sodium-dodecyl sulfate (SDS) polyacrylamide gel electrophoresis (PAGE). Once separated, the proteins could be easily extracted from the polyacrylamide gel by electroblotting. Treating the blotted proteins with the inexpensive amino acid-specific protease trypsin was a well-established method to generate characteristic peptide mixtures for use in protein characterization by liquid chromatography. By combining these existing methods with their hypothesis, they developed the first practical system for what they simply referred to as "protein identification."

Figure 7 illustrates a generalize scheme to implement this hypothesis, based on the design of the commercial protein identification software ProFound (Zhang & Chait, 2000). As in any practical implementation of an algorithm, some steps unrelated to the primary task of protein identification were necessary. The most important required preprocessing steps were the removal of peaks deemed to be caused by the ^{13}C isotope envelope that was always associated with a peptide ion signal, and the removal of peaks that could be associated with common experimental artifacts, such as the presence of peptides from the auto-digestion of trypsin. The removal of these adventitious peaks was practically important, because only the all-^{12}C peak (also referred to as the "A_0" peak) was calculated for the theoretical spectra. Because the presence of a large number of irrelevant peaks in the experimental spectrum reduced the statistical significance of any sequence-to-spectrum correlation, any means to legitimately reduce the complexity of the

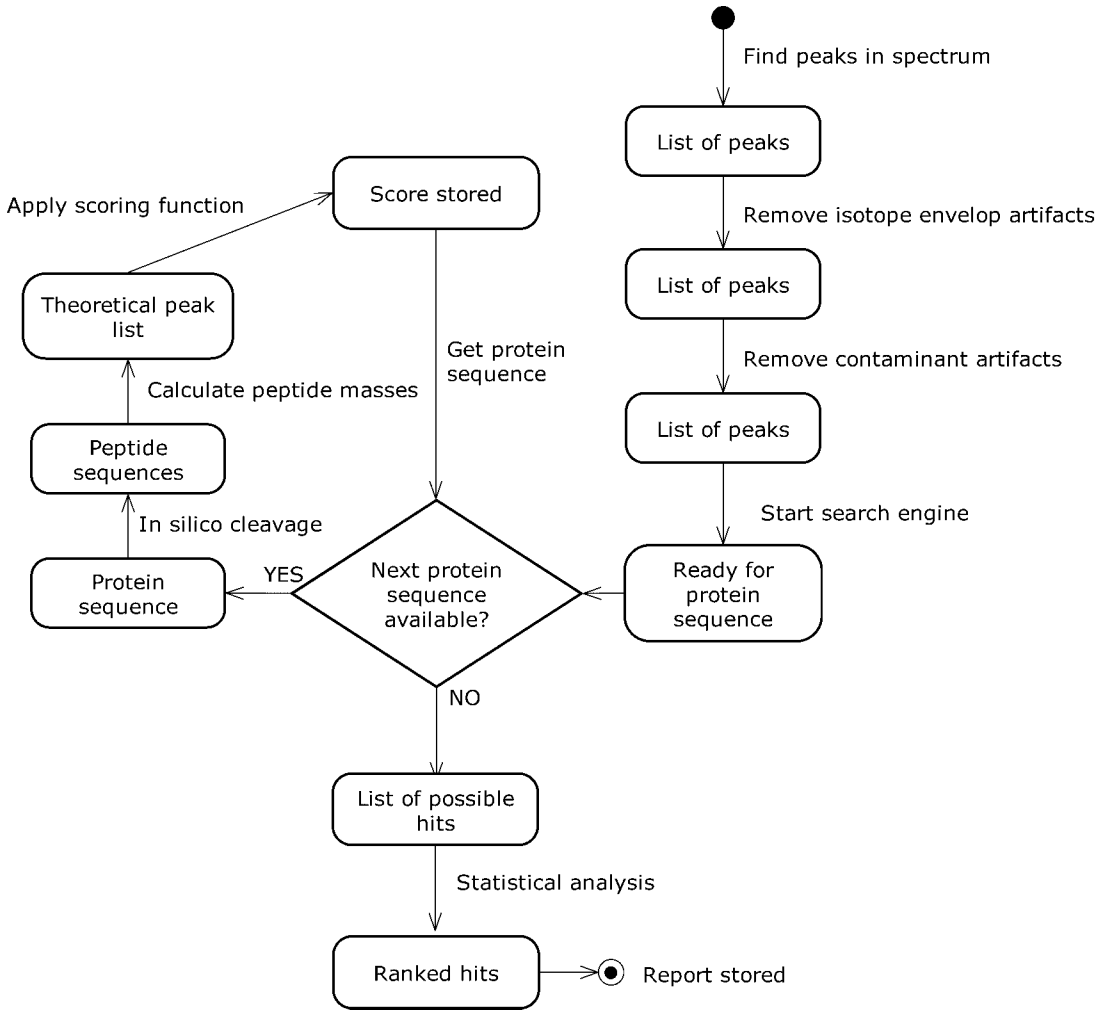

FIGURE 7. A state diagram of the peptide ion-mass mapping search process.

experimental spectrum normally improved the quality and sensitivity of the identification process.

The selection of the list of protein sequences to use for analysis was easy for Henzel et al. (1993): the combination of SWISS-PROT (Bairoch & Boeckmann, 1993), the PIR (Barker et al., 1993) and GenBank (Benson, Lipman, & Ostell, 1993) contained no more than 91,000 protein sequences. The current equivalent of this combination, NCBI's nonredundant database "nr," contains more than three million, or *ca.* 35×, as many proteins sequences. Naively, if one were to perform a search today, it should take approximately 35× more processor-clock cycles—a number that has been approximately compensated for by the increase in processor clock speeds. In practice, however, the presence of common experimental artifacts makes this simple analysis inaccurate. For example, any methionine residue can be oxidized either by endogenous enzymes, or during the sample preparation or storage. Therefore, for each methionine containing peptide, it is necessary to check for the presence of the unmodified and the oxidized form of the residue. If there are P nonmethionine-containing peptides to be tested for a particular protein and M methionine-containing peptides, then the number of calculations required for that protein, C, should be proportional to

$$C \sim 2^M$$

as each peptide must be checked with and without oxidation. Given this consideration, the total number of calculations for a set of proteins that contain N sequences, the total number of calculations, T_C, can be written as

$$T_C \sim \sum_{i=1}^{N} C_i = \sum_{i=1}^{N} 2^{M_i}$$

The highly nonlinear nature of T_C shows why the naïve calculation fails in practice: adding more sequences increases the number of calculations at a much higher-than-linear rate. Modifications where the value of M may be quite large (compare to methionine) are serine/threonine phosphorylation or asparagine/glutamine deamidation. Proteins that contain an unusually large number of potentially modified residues will obviously take a disproportionate time to calculate all of the possible modified peptide combinations, and can effectively "stall" a calculation.

This problem of dealing with potential modifications has led to a change to the use of lists of proteins that are a subset of the complete *nr* collection. For many model species, such as human, mouse or *E. coli*, there are completed genomes that allow searches to be performed against the translated gene models, rather than all of the sequences in *nr*. These genome-based sequences (the organism's proteome) can be calculated in reasonable time, and they represent all of the possible solutions to the protein identification problem, assuming that no adventious proteins from other organisms are present in a given sample.

B. Tandem Mass Spectrometry

The event-driven nature of MALDI ion sources initially limited the type of compatible mass analyzer to time-of-flight analyzers, which had limited ability to perform tandem mass spectrometry experiments. Therefore, Henzel-type protein identification became a natural application of these instruments. The limitations of this type of protein-identification experiment made it impractical to apply to mixtures of proteins. The Henzel-type identification scheme requires that a sample contain at most five distinct protein sequences: the best results are obtained when only a single protein sequence is present. Given the nature of most protein separation and isolation methods, obtaining a fraction that contained only a single protein sequence can be difficult. Even in the case of gel-band analysis, it is very common to find multiple protein species present in one spot. Most SDS–PAGE bands are contaminated with human cytokeratins 1, 2, 9, and 10, which are the most-prevalent proteins shed by the palm of an experimenter's hand. Whereas it is possible to limit this contamination by taking the appropriate precautions, these proteins (as well as those characteristic of human saliva) often are the dominant protein species present in bands generated by biological researchers. Therefore, it was desirable to develop instrumentation that could take advantage of the development of protein identification with tandem mass spectra.

Numerous experimental schemes were devised to take advantage of the spontaneous unimolecular decay of ions generated by the MALDI effect that occurred during the ions' passage through the mass analyzer. The analysis of these fragments was referred to as Post-Source Decay (PSD) measurements. The resulting spectra had much of the character of tandem mass spectra, because they were the result of ion fragmentation reactions. However, it was difficult to control the internal energy distribution of ions generated by the ion source; that difficulty led to a lack of reproducibility in these experiments. Subsequent development led to the production of mass analyzers with the capability of obtaining true tandem mass spectra that were compatible with the low duty-cycle, pulsed nature of the MALDI-generated ion currents (Medzihradszky et al., 2000; Shevchenko et al., 2000).

The development of true tandem mass spectrometers allowed the application of the other common method for protein identification: the use of collections of tandem mass spectra to determine which proteins were present in a particular biological sample. A tandem mass spectrum could be associated with a peptide amino acid sequence by comparing the experimental spectrum with the theoretical spectra generated from a list of potential peptides, based on known, sequence-specific peptide ion fragmentation reactions. A list of peptides can be derived from all of the protein sequences that could possibly be expressed by a particular organism. The comparison process generates a set of scores that indicate the similarity between any particular peptide sequence and the experimental mass spectrum. The peptide (or peptides) that are judged to be the most similar to the spectrum are associated with that spectrum, and the process is repeated for all of the spectra generated by the experiment. By combining the results of all of these spectrum-to-peptide correlations, a list of candidate proteins can be generated for the biological researcher. The idea of selecting a chemical structure based on an enumeration of theoretical mass spectra has a long history, and began with the research work of Djerassi and Lederberg (Duffield et al., 1969; Lederberg et al., 1969; Schroll, 1969) to identify organic compounds (Fig. 8). Its application to peptides was made possible by the sequence-specific bond cleavage rules described by Roepstroff and Folman (1984) and Biemann and Martin (1987). The method became popular for several reasons: there are practical software implementations of the idea; it is simple to automate the analysis of large data sets; and the scores can be interpreted statistically for large-scale applications (Keller et al., 2002; Fenyo & Beavis, 2003; Craig & Beavis, 2004; Geer et al., 2004).

This style of protein identification was developed for use with tandem mass spectrometers that used an electrospray ion source to generate precursor ions from reversed-phase high performance liquid chromatograph eluant. The ions generated by electrospray were predominantly doubly and triply charged ($z = 2+$ and $3+$). These ions frequently generate peptide tandem mass spectra that can be easily interpreted, given the peptide's amino acid sequence.

MALDI ion sources have been adapted for use with tandem mass spectrometers, primarily quadrupole-time-of-flight hybrid analyzers and TOF–TOF analyzers. MALDI normally generates intense singly charged ions ($z = 1+$) from peptides. The resulting tandem mass spectra can be somewhat more challenging to interpret than those generated from $z = 2+$ or $3+$ precursor ions. Figure 9 illustrates an example of the differences between the spectra generated from the $1+$ and $2+$ charge states from the same peptide. In general, $1+$ precursor ions generate fewer product fragment ions in a tandem mass spectrum than corresponding $2+$ ions. This observation can be rationalized by the fact that the protons that actually carry the positive charges tend to be localized at the N-terminal amine group and the C-terminal basic side chain that is characteristic of peptides generated by trypsin digestion of proteins. Therefore, when a doubly charged peptide fragments, the N-terminal and C-terminal fragments are both charged, and produce two charged fragment ions. A singly charged peptide can generate only a single charged fragment. The pulsed nature of the ion source and its low duty cycle also can result in tandem mass spectra that are composed of relatively few total ions, to produce spectra with discontinuous backgrounds that can appear to be "noisy."

Even with these limitations, product fragment ion spectra produced from MALDI-generated precursor ions can be used very effectively to identify peptides and proteins. In practice, data sets automatically generated with MALDI tend to have fewer spectra generated from the fragmentation of nonpeptide

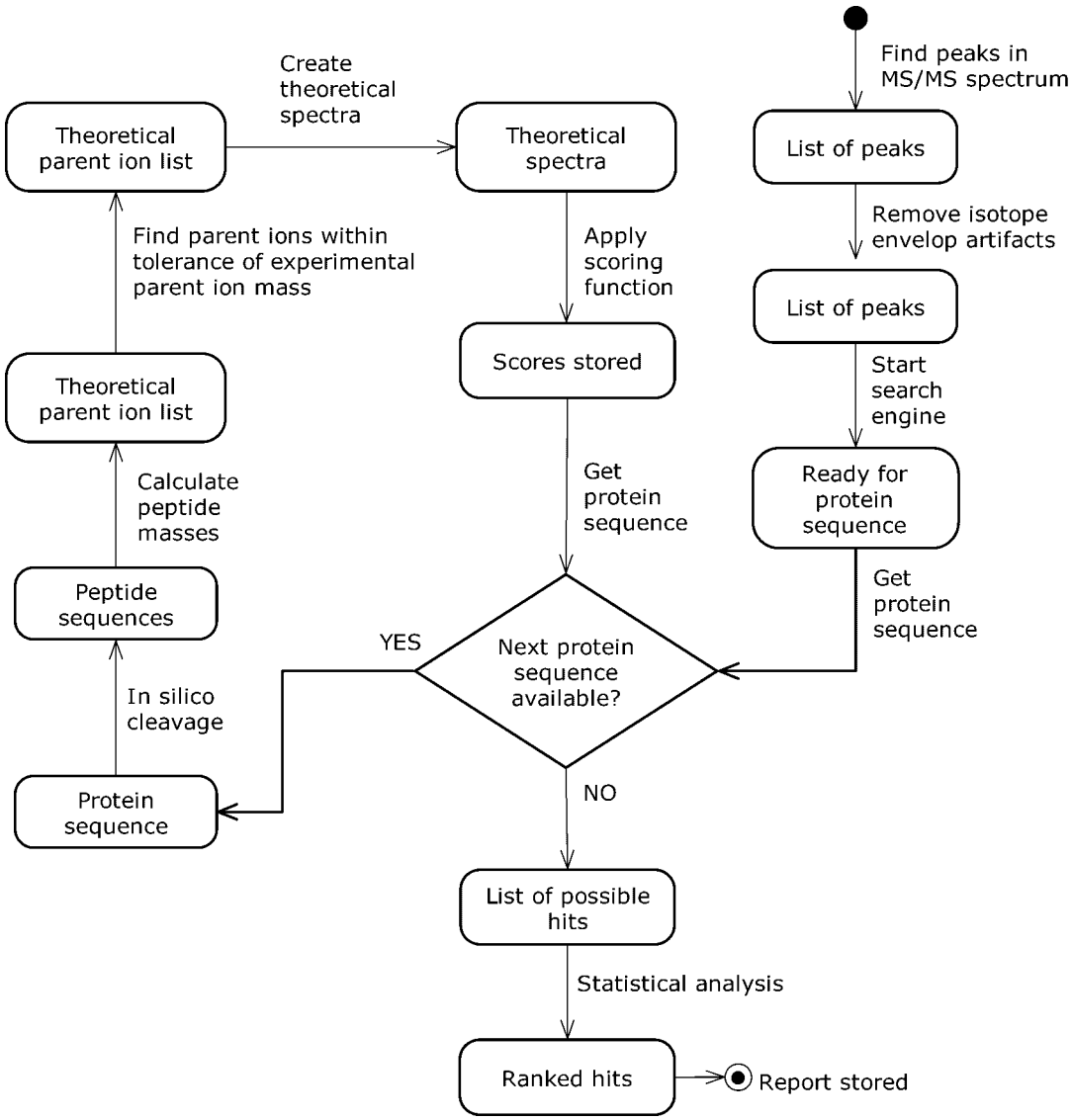

FIGURE 8. Lederberg–Djerassi style peptide fragmentation pattern matching process.

precursor ions, as compared to similar data sets generated by electrospray ion sources. Electrospray ion sources tend to produce chemical noise precursor ions that are difficult to automatically distinguish from peptide ions in the relatively short amount of time that the data acquisition system has to decide which peaks to fragment as chromatographic peaks elute. Although, MALDI ion sources also produce considerable chemical noise, there is no time limit: MALDI samples can be interrogated for as long as necessary to achieve a good measurement.

The lack of a time limit on MALDI MS/MS measurements has another considerable advantage for protein identification. In electrospray sources with a directly coupled HPLC, frequently more than one peptide elutes simultaneously and each peptide may generate several different charge states. Because measuring a tandem mass spectrum requires a considerable period of time, it is only possible to acquire mass spectra for a limited number of the precursor ions during elution; missing the opportunity to identify many peptides. MALDI samples can be interrogated systematically, so that all ions that might correspond to peptides can be analyzed. The data acquisition system also has the time to perform more sophisticated analysis of the results of one analysis before beginning the next analysis; that longer time produce a data set with a uniformly high overall spectrum quality.

C. Statistical Considerations

MALDI protein identifications that use the methods described in Sections V.A. and V.B. can be performed in a fully automated manner to generate thousands of complex data sets per day. In both techniques, the original information-handling software was created to allow the investigator to manually examine all of the results and to judge the quality of any identification based on their

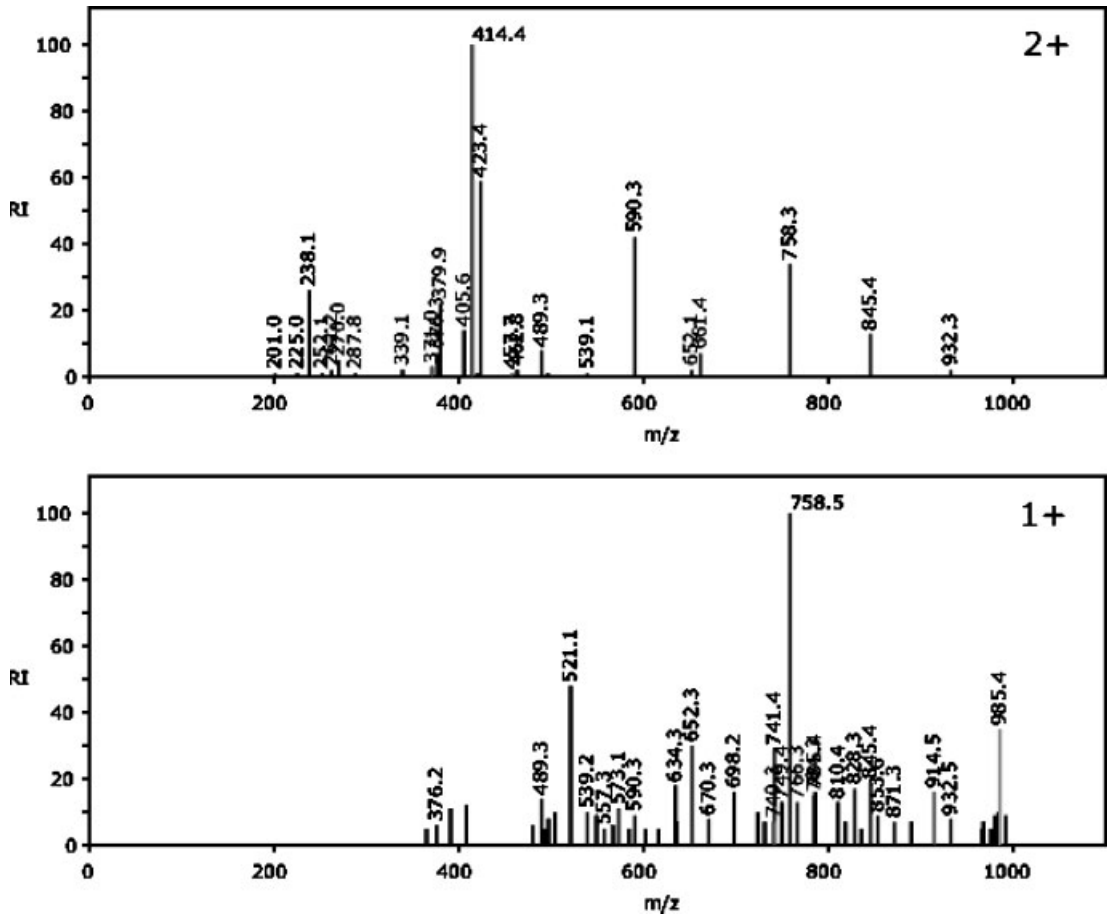

FIGURE 9. Demonstration of the effect of precursor ion charge state on the MS/MS spectrum generated from the peptide sequence LSSPATLNSR. The precursor ion charge is indicated in the upper right corner of each spectrum.

own expert knowledge. However, manual interpretation of the results is time-consuming and, therefore, prohibitive for large data sets. Different techniques for manual interpretation must be learned for each search algorithm because they each have a particular scoring scheme. Also, depending on the level of expertise and mood of the person doing the interpretation, it can also lead to large differences in the results. To overcome these limitations of manual interpretation, objective criteria that can be automated should be used to determine the quality of the results.

The basic challenge with protein identification by peptide mass fingerprinting and tandem mass spectrometry is that there is always random matching between the observed masses and the masses calculated from the protein sequences in the database (Eriksson, Chait, & Fenyo, 2000). This random matching can be minimized by maximizing mass accuracy and minimizing noise, but it cannot be removed completely. The consequence of random matching is that protein sequences that do not correspond to proteins present in the sample will partially match the experimental data and, therefore, will receive a score. Therefore, when the quality of the experimental data is poor, the protein sequence that receives the highest score might not be present in the sample, but might simply be a random and false identification.

Several methods to estimate the probability that a particular protein identification is false have been developed. Their basis is that a distribution of scores for false and random identification is estimated and used to test the significance of the results. The distribution of scores for false and random identification can be obtained by computer simulations (Eriksson, Chait, & Fenyo, 2000), using a model of the matching (Eriksson & Fenyo, 2004), fitting distributions for the score distribution of false and true identifications to a large set of results (Keller et al., 2002; Nesvizhskii et al., 2003) or by calculating expectation values based on the distribution of false identifications, that follows an extreme value distribution (Fenyo & Beavis, 2003).

Expectation values are calculated by collecting the score statistics during a search. The majority of sequences that partially match the experimental data are random, and the associated scores are distributed according to an extreme-value distribution. The expectation value of high-scoring results can be calculated by fitting an extreme value distribution to the observed distribution of scores and extrapolating. An expectation value assignment to an identification result should be interpreted as an

estimate of the number of false and random identification results at that specific score or at higher scores; that is, at low scores, it corresponds to the chance that it is a false and random result. The advantage of presenting the results with an associated expectation value instead of the score given by the particular search algorithm is that it allows the results to be judged without any detailed knowledge of the search algorithm used.

D. Quantitation

The quantitation of proteins and peptides is fundamental to understanding biological processes. It also has a great potential to yield more disease-specific drug targets to the discovery pipeline, and to allow the identification of biomarkers with higher diagnostic and prognostic value. Numerous well-known mass spectrometry-based quantitation methods have been applied to proteins and peptides during the last few years. All these methods attempt to overcome the fundamental issue that the mass spectrometric signal is strongly dependent on the properties of the peptide sequence. For example, the mass spectrum of a sample containing the same amount of two peptides with different amino acid sequences will most probably contain two peaks of very different intensity.

One way to overcome this issue is to compare the peaks that correspond to the same peptide in different samples. This method requires that the sample-handling and -preparation, as well as the mass spectrometric analysis, are done in a highly reproducible way (Villanueva et al., 2005).

Another way is to label the samples with different stable isotopes, and to mix them to minimize any variation in the sample handling and measurement. Here, the quantitation is done by comparing the peak intensities for the same peptide labeled with the different isotopes and therefore that originated from different samples. It is imperative to introduce the isotopic labeling at an early stage of the sample handling, and most stable isotopic labels are, therefore, designed to have very similar behavior during separation. Ideally, the isotopic label is introduced *in vivo* by metabolic labeling (Oda et al., 1999; Ong et al., 2002). *In vivo* metabolic labeling is most often used in cell cultures, because it is usually prohibitively expensive for animals. After the cells have been disrupted, the proteins can be labeled by Isotope Coded Affinity Tag (ICAT) (Gygi et al., 1999), quantification of the cysteine-containing peptides. During or after protein digestion, the C-terminus of each proteolytic peptide can be labeled with ^{18}O (Mirgorodskaya et al., 2000). After digestion, the iTRAQ (Ross et al., 2004) reagent can be used to label the N-terminus of each proteolytic peptide. The advantages of labeling the peptides after digestion are that no selection of peptides containing a certain amino acid is necessary, and that higher protein coverage can be achieved for less complex samples. Synthetic peptides with isotope-labeled amino acids can also be added for the quantitation of the corresponding peptides in the sample (Gerber et al., 2003).

For accurate quantitation, it is critical to correctly detect the peaks that correspond to a peptide, subtract the background, and accurately determine the limits to use for the integration. It is also important that the mixture of peptides is not very complex so that the isotopic distributions of different peptides do not overlap. When there are overlapping peaks, these need to be excluded from the quantitation to avoid any misleading results. In cases when the isotopes of the peptides are resolved, the isotopic distribution can be utilized to determine if there is interference from other peptides with similar mass. Criteria for accepting peaks can include: signal-to-noise ratio, mass tolerance, retention time tolerance, shape of isotope distribution, and mass difference between isotopes.

E. Data Repository Development and Standardization

A significant new strategy to understand the results of proteomics experiments is to collect a large number of peptide mass spectra obtained from proteomics experiments and to store them in a repository. For organic compounds this strategy has been very successful (Heller, 1972; Heller, 1999). When a new mass spectrum-to-peptide sequence correlation is postulated, the repository could be queried to return a list of the best previously observed mass spectra that have been associated with that sequence. In a comparison of the existing exemplar peptide ion fragmentation patterns with the newly observed pattern, the repository would provide some of the same functions as a library of spectra obtained from synthetic peptides, with the proviso that the sequence annotation would be based on spectrum-to-proteome matching, rather than on known peptide analytes. This sort of repository structure allows the system to remain relevant as new instrumentation becomes available. It also has the potential to provide additional confidence to particular assignments by having many redundant measurements of the same peptide sequence's fragmentation pattern under a variety of different experimental conditions; for example, different precursor ion charge states, fragment ion signal-to-noise ratios, or mass spectrometer configurations.

These repositories may be used to compare the patterns of peptides that have been observed for a particular protein sequence (often referred to as the observed ''coverage map'' of a sequence). This pattern of observed peptide ions is naturally a property of the protein sequence and the physical properties of the peptides. It is also a function of the analytical sample processing protocols, the mass spectrometer's ion source, and the fragmentation conditions for the peptides. This combination of characteristics makes it difficult to predict which of the theoretical peptides for a protein sequence will actually be observed. However, by comparing an observed peptide coverage map with the best previously observed coverage maps, it should be possible to determine whether the observed pattern is consistent with previous results. This type of comparison becomes particularly important when only one or two peptides from a particular protein are observed, where knowing that these few peptides consistently produce the strongest signals would add considerable confidence to their assignment.

Several database schemas have been proposed. These schemas are, to one degree or another, an extension or a simplification of the Minimum Information About a Proteomics Experiment (MIAPE) (Taylor et al., 2007) idea, for the purpose of validating observed protein coverage and peptide fragmentation data. The design goal of any schema should be to create a database that could be used on its own to provide answers to queries as well as to serve as an index to

experimental information stored in XML documents. The grafting of a specialized relational schema with the object structure of the XML document has considerably simplified the design process, and has allowed the creation of working, publicly available data repository for the bioinformatics analysis of proteomics data. Any of these systems should be complementary with raw data repositories currently either being deployed or designed (e.g., GPMDB (Craig, Cortens, & Beavis, 2004), Peptide Atlas (Desiere et al., 2006) or PRIDE (Hermjakob & Apweiler, 2006)).

Preliminary studies on the most developed of these repositories, GPMDB, indicate that coupling a protein identification system to a database and data-retrieval system is possible through the use a repository system. This system provides useful validation information about the assignment of mass spectra to peptide sequences in a simple manner, and uses a relatively compact relational database structure. The repository has proven to be useful in the planning of proteomics experiments, by providing insights into the peptides that are reasonably expected to produce signals and the other proteins that have been observed in concert with a desired protein species. It also has proven useful for teaching and demonstration purposes, by supplying a range of examples that can illustrate the appropriate use of the information derived from proteomics. The current prototype repository contains the results of over 4×10^4 complete LC MS/MS (or equivalent) analyses of proteins, that produced 1.3×10^7 confident peptide-to-spectrum identifications. All of these analyses have been donated by laboratories from all over the world, as well as the results of major proteomics projects, such as the Human Proteome Organization's Plasma Proteome Project (Omenn et al., 2005).

Examples of the broader, proteome-wide questions that can be investigated with a completed, populated repository of this type are as follows:

1. Are the patterns observed in protein coverage maps primarily a property of the primary sequence of the protein?
2. What is the variability of tandem mass spectrum fragmentation, and does attempting to predict this variability truly improve the confidence of identifications?
3. What are the characteristics of peptides that respond well to current proteomics techniques and, more importantly, what are the characteristics of peptides that persistently do not produce interpretable signals?
4. How many of the genes, predicted for a particular organism, have ever been observed on the translated protein level?
5. Given that a particular protein has been conclusively observed for a particular organism, is that sufficient evidence to require the retention of a particular gene model in a genome, even though improved gene-prediction methods suggest that the gene was originally called in error? Does a replacement gene have to be found to be consistent with proteomics observations?

Questions such as these can only be answered now on a case-by-case basis, by individual laboratories that generate data and analysis. With these repository systems in place, these questions can be answered quickly by biological and bioinformatics researchers, and new experiments can be planned more efficiently.

VI. CONCLUDING REMARKS

MALDI-MS has been successfully been applied to solve a wide range of biological problems. Arguably the main reason for this success is that MALDI is easy to use. The most common workflow is practiced in many biological laboratories worldwide: a gel band is excised; the proteins in the band are digested with a proteolytic enzyme; and the proteins are identified by peptide mapping using MALDI-MS. The informatics support for this workflow is maturing, and during the last few years the awareness of the importance of proper statistical analysis to obtain reliable results has increased (Eriksson, Chait, & Fenyo, 2000).

There has been a trend towards using tandem mass analyzers in combination with MALDI to allow for analysis of more complex mixtures. Many aspects of the informatics support for tandem MS are already available, but the current developments of large public databases of tandem mass spectra have the potential to increase the information that we can extract from MALDI experiments: these databases can be used validate results (Craig, Cortens, & Beavis, 2004) and identify peptides through library searches (Craig et al., 2006).

REFERENCES

Baggerly KA, Coombes KR, Morris JS. 2005. Bias, randomization, and ovarian proteomic data: A reply to "producers and consumers". Cancer Inform 1:9–14.

Bairoch A, Boeckmann B. 1993. The SWISS-PROT protein sequence data bank, recent developments. Nucleic Acids Res 21:3093–3096.

Barker WC, George DG, Mewes HW, Pfeiffer F, Tsugita A. 1993. The PIR-International databases. Nucleic Acids Res 21:3089–3092.

Beavis RC, Chait BT. 1989a. Cinnamic acid derivatives as matrices for ultraviolet laser desorption mass spectrometry of proteins. Rapid Commun Mass Spectrom 3:432–435.

Beavis RC, Chait BT. 1989b. Factors affecting the ultraviolet laser desorption of proteins. Rapid Commun Mass Spectrom 3:233–237.

Beavis RC, Chait BT. 1989c. Matrix-assisted laser-desorption mass spectrometry using 355 nm radiation. Rapid Commun Mass Spectrom 3:436–439.

Beavis RC, Chait BT. 1990. Rapid, sensitive analysis of protein mixtures by mass spectrometry. Proc Natl Acad Sci USA 87:6873–6877.

Benson D, Lipman DJ, Ostell J. 1993. GenBank. Nucleic Acids Res 21:2963–2965.

Biemann K, Martin SA. 1987. Mass spectrometric determination of the amino acid sequence of peptides and proteins. Mass Spectrom Rev 6: 1–76.

Cohen SL, Ferre-D'Amare AR, Burley SK, Chait BT. 1995. Probing the solution structure of the DNA-binding protein Max by a combination of proteolysis and mass spectrometry. Protein Sci 4:1088–1099.

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

Craig R, Cortens JP, Beavis RC. 2004. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res 3: 1234–1242.

Craig R, Cortens JC, Fenyo D, Beavis RC. 2006. Using annotated peptide mass spectrum libraries for protein identification. J Proteome Res 5:1843–1849.

Creasy DM, Cottrell JS. 2004. Unimod: Protein modifications for mass spectrometry. Proteomics 4:1534–1536.

Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R. 2006. The PeptideAtlas project. Nucleic Acids Res 34:D655–D658.

Ding C, Cantor CR. 2003. A high-throughput gene expression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight MS. Proc Natl Acad Sci USA 100:3059–3064.

Duffield AM, Robertson AV, Djerassi C, Buchanan BG, Sutherland GL, Feigenbaum EA, Lederberg J. 1969. Applications of artificial intelligence for chemical inference. II. Interpretation of low-resolution mass spectra of Ketones. J Am Chem Soc 91:2977–2981.

Eriksson J, Fenyo D. 2004. Probity: A protein identification algorithm with accurate assignment of the statistical significance of the results. J Proteome Res 3:32–36.

Eriksson J, Chait BT, Fenyo D. 2000. A statistical basis for testing the significance of mass spectrometric protein identification results. Anal Chem 72:999–1005.

Fenyo D. 1997. A software tool for the analysis of mass spectrometric disulfide mapping experiments. Comput Appl Biosci 13:617–618.

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

Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. 2004. Open mass spectrometry search algorithm. J Proteome Res 3:958–964.

Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. 2003. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proc Natl Acad Sci USA 100:6940–6945.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17:994–999.

Haff LA, Smirnov IP. 1997. Multiplex genotyping of PCR products with MassTag-labeled primers. Nucleic Acids Res 25:3749–3750.

Hahner S, Ludemann HC, Kirpekar F, Nordhoff E, Roepstorff P, Galla HJ, Hillenkamp F. 1997. Matrix-assisted laser desorption/ionization mass spectrometry (MALDI) of endonuclease digests of RNA. Nucleic Acids Res 25:1957–1964.

Heller S. 1972. Conversational mass spectral retrieval system and its use as an aid in structure determination. Analytical Chemistry 44:1951–1961.

Heller S. 1999. The history of the NIST/EPA/NIH mass spectral database. Today's Chemist at Work 8:45–50.

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

Hermjakob H, Apweiler R. 2006. The proteomics identifications database (PRIDE) and the ProteomExchange consortium: Making proteomics data accessible. Expert Rev Proteomics 3:1–3.

Hillenkamp F, Karas M. 1990. Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization. Methods Enzymol 193:280–295.

Hillenkamp F, Karas M, Beavis RC, Chait BT. 1991. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. Anal Chem 63:1193A–1203A.

Johnson RE, Sundqvist BU, Hedin A, Fenyo D. 1989. Sputtering by fast ions based on a sum of impulses. Phys Rev B Condens Matter 40:49–53.

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

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

Kiernan UA, Tubbs KA, Gruber K, Nedelkov D, Niederkofler EE, Williams P, Nelson RW. 2002. High-throughput protein characterization using mass spectrometric immunoassay. Anal Biochem 301:49–56.

Lederberg J, Sutherland GL, Buchanan BG, Feigenbaum EA, Robertson AV, Duffield AM, Djerassi C. 1969. Applications of artificial intelligence for chemical inference. I. The number of possible organic compounds. Acyclic structures containing C, H, O, and N. J Am Chem Soc 91:2973–2976.

Little DP, Braun A, Darnhofer-Demar B, Koster H. 1997. Identification of apolipoprotein E polymorphisms using temperature cycled primer oligo base extension and mass spectrometry. Eur J Clin Chem Clin Biochem 35:545–548.

Macfarlane RD, Torgerson DF. 1976. Californium-252 plasma desorption mass spectroscopy. Science 191:920–925.

Medzihradszky KF, Campbell JM, Baldwin MA, Falick AM, Juhasz P, Vestal ML, Burlingame AL. 2000. The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. Anal Chem 72:552–558.

Mirgorodskaya OA, Kozmin YP, Titov MI, Korner R, Sonksen CP, Roepstorff P. 2000. Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18)O-labeled internal standards. Rapid Commun Mass Spectrom 14:1226–1232.

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

Nordhoff E, Ingendoh A, Cramer R, Overberg A, Stahl B, Karas M, Hillenkamp F, Crain PF. 1992. Matrix-assisted laser desorption/ionization mass spectrometry of nucleic acids with wavelengths in the ultraviolet and infrared. Rapid Commun Mass Spectrom 6:771–776.

Oda Y, Huang K, Cross FR, Cowburn D, Chait BT. 1999. Accurate quantitation of protein expression and site-specific phosphorylation. Proc Natl Acad Sci USA 96:6591–6596.

Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, Hermjakob H, Apweiler R, Haab BB, Simpson RJ, Eddes JS, Kapp EA, Moritz RL, Chan DW, Rai AJ, Admon A, Aebersold R, Eng J, Hancock WS, Hefta SA, Meyer H, Paik YK, Yoo JS, Ping P, Pounds J, Adkins J, Qian X, Wang R, Wasinger V, Wu CY, Zhao X, Zeng R, Archakov A, Tsugita A, Beer I, Pandey A, Pisano M, Andrews P, Tammen H, Speicher DW, Hanash SM. 2005. 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 5:3226–3245.

Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics 1:376–386.

Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA. 2002. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 359: 572–577.

Pieles U, Zurcher W, Schar M, Moser HE. 1993. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: A powerful tool for the mass and sequence analysis of natural and modified oligonucleotides. Nucleic Acids Res 21:3191–3196.

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

Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ. 2004. Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 3:1154–1169.

Schroll G. 1969. Applications of artificial intelligence for chemical inference. III. Aliphatic ethers diagnosed by their low-resolution mass spectra and nuclear magnetic resonance data. J Am Chem Soc 91:2977–2981.

Severinov K, Fenyo D, Severinova E, Mustaev A, Chait BT, Goldfarb A, Darst SA. 1994. The sigma subunit conserved region 3 is part of "5′-face" of active center of *Escherichia coli* RNA polymerase. J Biol Chem 269: 20826–20828.

Shevchenko A, Loboda A, Shevchenko A, Ens W, Standing KG. 2000. MALDI quadrupole time-of-flight mass spectrometry: A powerful tool for proteomic research. Anal Chem 72:2132–2141.

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

Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK Jr, Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW, Dunn MJ, Heck AJ, Leitner A, Macht M, Mann M, Martens L, Neubert TA, Patterson SD, Ping P, Seymour SL, Souda P, Tsugita A, Vandekerckhove J, Vondriska TM, Whitelegge JP, Wilkins MR, Xenarios I, Yates JR III, Hermjakob H. 2007. The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol 25:887–893.

Tost J, Schatz P, Schuster M, Berlin K, Gut IG. 2003. Analysis and accurate quantification of CpG methylation by MALDI mass spectrometry. Nucleic Acids Res 31:e50.

Trester-Zedlitz M, Kamada K, Burley SK, Fenyo D, Chait BT, Muir TW. 2003. A modular cross-linking approach for exploring protein interactions. J Am Chem Soc 125:2416–2425.

Tsarbopoulos A, Karas M, Strupat K, Pramanik BN, Nagabhushan TL, Hillenkamp F. 1994. Comparative mapping of recombinant proteins and glycoproteins by plasma desorption and matrix-assisted laser desorption/ionization mass spectrometry. Anal Chem 66:2062–2070.

Villanueva J, Philip J, Chaparro CA, Li Y, Toledo-Crow R, DeNoyer L, Fleisher M, Robbins RJ, Tempst P. 2005. Correcting common errors in identifying cancer-specific serum peptide signatures. J Proteome Res 4:1060–1072.

Wu KJ, Steding A, Becker CH. 1993. Matrix-assisted laser desorption time-of-flight mass spectrometry of oligonucleotides using 3-hydroxypicolinic acid as an ultraviolet-sensitive matrix. Rapid Commun Mass Spectrom 7:142–146.

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

David Fenyo received his Ph.D. in Physics from Uppsala University for work on the mechanism for plasma desorption mass spectrometry. He has worked in academia and the private sector. His research interests include bioinformatics, proteomics, and mass spectrometry. Currently he is Senior Research Associate at The Rockefeller University, New York, New York.

Ronald C. Beavis was born in Winnipeg, Canada. He graduated from the University of Manitoba with B.Sc. degrees in Physics and Zoology. He then joined Ken Standing's group, earning his Ph.D. working on the development of time-of-flight mass spectrometers for biological applications. He has worked in academia and the private sector. Currently, he is the Canada Research Chair in Experimental Bioinformatics at the University of British Columbia in Vancouver, Canada.