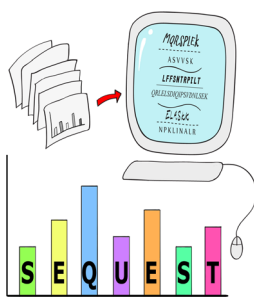


Pivotal Role of Computers and Software in Mass Spectrometry – *SEQUEST* and 20 Years of Tandem MS Database Searching

John R. Yates III

Chemical Physiology and Molecular and Cellular Neurobiology, The Scripps Research Institute, 10550 North Torrey Pines Road, SR302B, La Jolla, CA 92037, USA



Abstract. Advances in computer technology and software have driven developments in mass spectrometry over the last 50 years. Computers and software have been impactful in three areas: the automation of difficult calculations to aid interpretation, the collection of data and control of instruments, and data interpretation. As the power of computers has grown, so too has the utility and impact on mass spectrometers and their capabilities. This has been particularly evident in the use of tandem mass spectrometry data to search protein and nucleotide sequence databases to identify peptide and protein sequences. This capability has driven the development of many new approaches to study biological systems, including the use of “bottom-up shotgun proteomics” to directly analyze protein mixtures.

Keywords: SEQUEST, Database searching, Tandem mass spectrometry, Computers, Proteomics

Received: 27 March 2015/Revised: 17 June 2015/Accepted: 20 June 2015/Published Online: 19 August 2015

Technological advancement depends on an intricate balance between the inventions and innovations of different fields. For example, Charles Babbage had the idea to create programmable computers in the early 1800s, but the necessary electronics technology had not yet been invented for him to fully realize his idea. Babbage’s ideas were just far ahead of his time and his design was shown to be sound when a team built a mechanical calculator from his original plans in 1989. Evolution of technology depends on the innovations of many fields and, often, on the timing of those developments. In addition, unexpected advances in technology in one field can create disruptive leaps in other fields. Advances in mass spectrometry have come about this way. It was invented over 100 years ago and since then has benefited from advances in many fields, notably computerized metal machining and vacuum technology, but it has been advances in electronics and computers that have driven the creation of the powerful mass spectrometers we have today.

Many advances in mass spectrometry technology over the last 100 years can be directly linked to innovations in the electronics industry. Mass spectrometers depend on the use of electric and magnetic fields for the separation of ions and, thus,

the capabilities of instruments have increased with the robustness, accuracy, and precision of electronics. The invention and commercialization of solid-state electronics coincided with the emergence of quadrupole theory and eventually helped drive the development and use of quadrupole mass spectrometers by providing the dependable and precise electronics required for the electric fields used to separate ions in quadrupoles. Throughout the 1960s, innovations in the electronics industry, particularly at Fairchild Semiconductor and Texas Instruments, resulted in improved performance and decreased costs of integrated circuits. Eventually the advances in both price and performance allowed more affordable computers to be built in academic environments, and commercialization of these innovations through new companies such as Digital Equipment Corporation (DEC) led to the development of real time interactive minicomputers. MIT was one of the centers of the 1960s “computing revolution” and, as a result, the Biemann mass spectrometry laboratory was well poised to exploit computers to aid in the interpretation of mass spectra.

The integration of computer technology with mass spectrometry was most impactful in three areas. The first area was the use of computers to perform difficult or tedious calculations to aid in the interpretation of mass spectra (Figure 1a). Early examples included the determination of the originating m/z values forming metastable ions and the calculation of accurate mass data [1, 2]. When gas chromatographs began to be

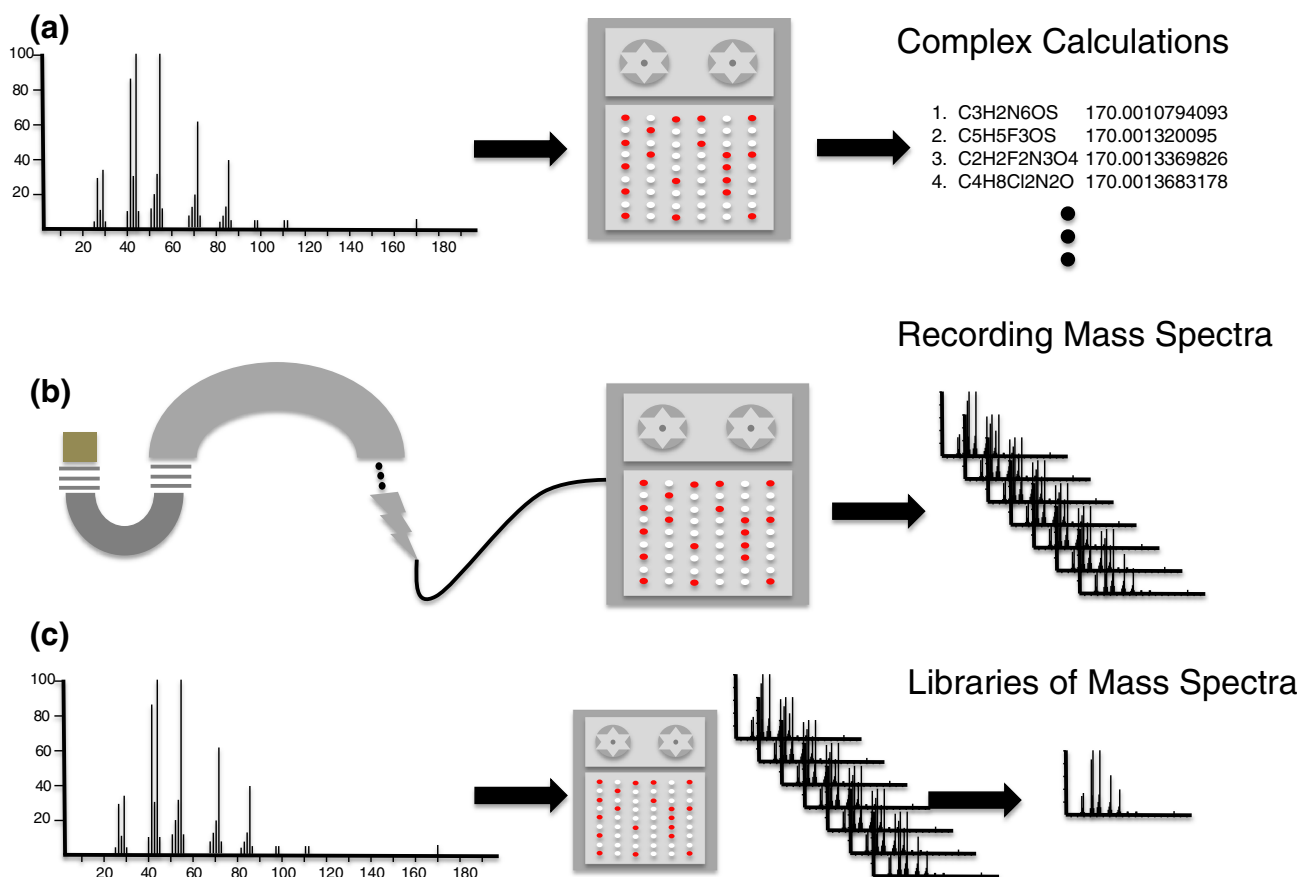


Figure 1. The interfacing of computers with mass spectrometry began along three fronts. **(a)** Computers were used to calculate values for metastable ions or to calculate the chemical formulas associated with specific molecular weights. **(b)** Computers were interfaced with mass spectrometers to directly collect mass spectra. A need to develop more efficient data collection strategies began with on-line gas chromatography. **(c)** To speed the interpretation of data, laboratories began collecting interpreted mass spectra to form libraries and to develop methods to search those libraries with mass spectra

connected to mass spectrometers, an increase in the scanning speed of instruments became necessary [3–5]. The development of faster scanning mass spectrometers meant an increase in the number of spectra being collected and, thereby, an increase in the need to speed the calculation of chemical formulas from accurate mass data for the increasing number of molecules analyzed [6–8]. These new capabilities of separation and mass analysis drove the need to collect data directly into computers rather than by photographic plate or faraday cup [9]. Digital collection of data would be far more efficient and the data would already be in the computer for processing.

The need to collect data coming off the mass spectrometer directly into a computer drove a second stage of the integration with computers (Figure 1b). A dramatic increase in the amount of data collected using GCMS instead of a solids probe-based technique created a data analysis bottleneck and the need to not only collect the data directly into the computer but also to create algorithms to process the data. Initial GCMS analyses used a Mattuch-Herzog design to project a complete mass spectrum onto photographic plate. As many as 30 spectra could be collected onto one plate [4]. Mass spectral data from the photographic plate could be read by a comparator-densitometer

and then fed into a computer for processing. An advantage to this method over the use of a scanning magnetic sector was the ability to collect high resolution data [10]. Although photographic plates could collect as many as 30 spectra before a new plate had to be inserted into the mass spectrometer, there was still a drive to read data directly into computers to accommodate increasingly complicated samples, and this process required conversion of ion signals into electrical signals [8, 9]. Adding an electron multiplier to detect the ions and then converting the analog signal to a digital form with storage on a magnetic tape, which could be fed directly into a computer, achieved this goal. Continuous recording of data created new obstacles to overcome. Hites and Biemann automated the collection and calibration of mass spectra using computers, which paved the way for better acquisition of GCMS data [9]. As these processes became more sophisticated, new problems arose, such as the loss of fine detail of the molecules eluting in the mass spectrometer when GCMS data was plotted as a total ion chromatogram. The Biller-Biemann algorithm solved this problem by plotting the base peak ion for every scan, which restored fine detail to the chromatogram [11]. These algorithms enhanced the efficiency of data processing and put

MS data directly into a form that could be further processed by computer algorithms. However, in the 1960s, manual collection and interpretation of the mass spectra was still the norm for all but a few laboratories, and it was not until the middle of the 1970s that a commercial data system from INCOS Inc. became available for most mass spectrometers.

As mass spectrometers evolved, there was a concurrent drive to understand the gas-phase ion chemistry of molecules to aid in the interpretation of spectra. Throughout the 1960s there was a focus on deciphering fragmentation mechanisms of ions using 70-eV electron ionization. To understand how molecular bonds fragmented during ionization, many studies were carried out using specifically prepared synthetic molecules with strategically placed stable isotopes. Which fragment ion got the stable isotope, as represented by an increase in the expected m/z value, helped determine the mechanisms of bond cleavage. The details of these processes have been described in classic books on mass spectrometry [12–14]. As the details of fragmentation revealed themselves, efforts turned to making the process of spectral interpretation more efficient and, consequently, a third stage of algorithm development emerged that focused on increasing the speed of interpretation of mass spectra [15] (Figure 1c). Two efforts undertaken during this period are most notable. The first was the creation of mass spectral libraries and the algorithms used to match spectra as an aid in interpretation of mass spectra [16–19]. This was perhaps one of the earliest attempts to “crowd source” a problem since a scientist’s interpretation of a particular spectrum to identify the molecule it represented was subsequently shared throughout the mass spectrometry field through a library. Several groups were prominent in the development of algorithms to search spectral libraries. McLafferty’s and Biemann’s groups laid the early foundations for library searching, but their efforts were constrained by the limited capabilities of computers as well as by the availability of computers more generally [16–18, 20, 21]. This situation was remedied, as mentioned previously, with the development of a commercial data system. The community-wide effort to collect EI spectra of small molecules for inclusion in libraries would make this approach quite powerful in time. Such libraries are still widely used for the analysis of mass spectra, and the library concept has become even more relevant as metabolomics has surged in popularity [22–25].

The second strategy to improve data interpretation efficiency was to develop more “intelligent” approaches to data analysis. In the mid to late 1950s, the field of artificial intelligence (AI) was born. By the mid-1960s, the field had some surprising success and interest in the programming methods expanded. Edward Feigenbaum, Bruce Buchanan, Joshua Lederberg, and Carl Djerassi teamed up to develop an approach to use AI to interpret mass spectra of small molecules [26]. A goal was to use a heuristic approach to interpret mass spectra in much the same way a human would, and this idea was encoded in a software program called *Dendral*. Although the program never achieved complete independence of interpretation, it was able to eliminate the most implausible structures and narrow the possible structures to be considered. This narrowing of the

possible structures was sufficient to allow a non-expert to complete the interpretation process without too much difficulty. Despite this futuristic vision of automating de novo interpretation of organic molecule mass spectra, this concept never advanced much further than *Dendral* for many decades. Most interpretation of organic molecule mass spectra is still performed through library searching or by manual interpretation, although the increasing popularity of metabolomics has stimulated new interest in developing de novo interpretation algorithms [27, 28]. However, de novo interpretation remains a significant challenge, as the even-electron ionization methods of today (ESI and APCI) often yield limited fragmentation information, even when coupled with CID [27].

As early as 1980 it was understood that since the structure of peptides is well ordered and that interpretation of their mass spectra should be automatable [29]. Biemann was able to automate with good accuracy the interpretation of the EIMS of polyamino alcohol derivatives of di- to tetra-amino acid peptides [30]. Because it was necessary to modify the peptides to increase volatility for GCMS analysis, the size of peptide amenable to this process was quite limited. The arrival of fast atom bombardment (FAB) as a means to ionize underivatized peptides allowed the analysis of much larger peptides [31–33]. FAB ionization caused sufficient fragmentation to determine the amino acid sequence of peptides [34] (Figure 2a). There were attempts to develop computer algorithms to interpret these spectra, but the spectra were noisy from the background generated by FAB, and unless the peptide was pure they would likely be contaminated with fragments of several peptides [35]. When FAB was used on a tandem mass spectrometer for peptide sequencing, it provided the advantages of being able to select the peptide ion to eliminate contamination from other peptide ions and to reduce noise from the ionization method, as well as the use of a collision cell to produce more efficient and predictable fragmentation [32, 33]. Even though the reduction in background noise and fragmentation from other peptides resulted in cleaner spectra, automated de novo interpretation was not successful in creating large-scale analyses of peptides [35–37]. A limitation still plaguing algorithms was the quality of the mass spectral data, as it was neither high resolution nor high mass accuracy and de novo interpretation greatly benefits from both. While automation of data interpretation did not succeed to a great extent, the ability to manually interpret tandem mass spectra of peptides was still possible, allowing the use of MS to solve significant protein chemistry problems [38–45]. Since a robust method to integrate liquid separations with tandem mass spectrometry was yet to be invented, the one sample at a time nature of FAB-MS/MS did not create an unmanageable data analysis bottleneck [42].

When separation methods are integrated with mass spectrometers, a very powerful and large-scale method to detect the molecules being separated is created. Connecting liquid chromatography with mass spectrometers was of long-standing interest in the field. Some rather inventive methods were developed, from the moving belt interface to thermospray, but it was not until the development of electrospray ionization (ESI)

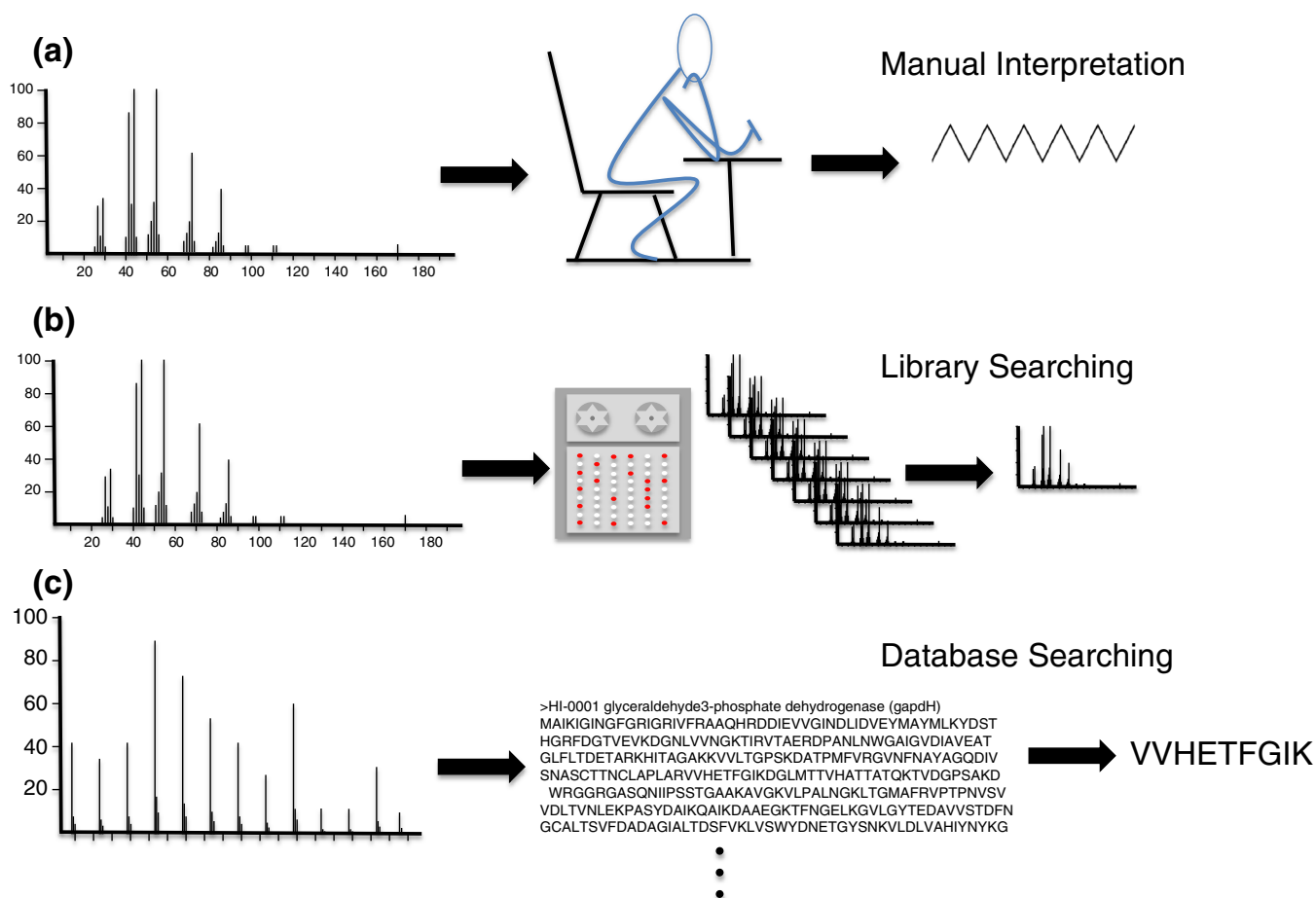


Figure 2. The interpretation of mass spectra has benefited enormously from computers. **(a)** Initial interpretation of mass spectra was done manually with the interpreted mass spectra being added to libraries. **(b)** The creation of mass spectral library stored on computers allowed dissemination to laboratories so all could benefit from the interpretation efforts of individuals. This was an early example of “crowd sourcing” a scientific activity. **(c)** A new strategy was developed that used protein sequences created by human genome project to interpret tandem mass spectra of peptides. An unusual feature of this approach was that models of the expected of tandem mass spectra were created “on the fly” from the sequences present in the database to interpret the spectra

that the problem was finally solved [46–48]. The essential element of ESI is placing a high voltage on the tip of the LC outlet to form a spray of fine droplets that is outside the vacuum system. Previous methods tried to introduce the liquid in the high vacuum of the ion source, an approach that is fraught with complications. ESI solved the interface problem for LC and it created a robust and powerful method to introduce biomolecules directly into the mass spectrometer. At the same time ESI was emerging, vigorous discussions on sequencing the human genome were taking place [49, 50]. Recent advances in the technology to sequence DNA drove an optimism that tackling the sequencing of the human genome as a world-wide, coordinated effort would yield huge benefits for human biology, medicine, and mankind. So important was the invention of DNA sequencing that Alan Malcolm argued this development spelled the end of protein sequencing, and he was substantially correct [51]. Automation of the DNA sequencing process drove a consensus of leading scientists, government officials, and politicians to decide to sequence not only the human genome

but also a set of model organisms to provide experimental systems in which to understand the functions of human genes. While the Genome projects were heralded for their vision, it was believed that the data would be understood through a combination of bioinformatics and genetics. In the report of the National Academy of Sciences recommending the project, there was little mention of the role of protein biochemistry in understanding the function of gene products, nor was there any indication that protein biochemistry might benefit by sequencing genomes [50]. Despite early optimism, so far it has been the simpler Mendelian diseases and traits that have been readily understood through genome sequencing. More complex, and multi-genetic, diseases have been slow to yield their secrets through large-scale genomics, suggesting the diseases are quite complicated in their etiology [52, 53].

Until this time, the principal method for the analysis of mass spectra had remained the library search (Figure 2b). While peptide tandem mass spectra might have seemed like an easy case for de novo interpretation, as the fragmentation patterns

are relatively straightforward, the number of possible combinations can escalate very quickly. A 5-residue peptide can have (20^5) 3,200,000 different sequence combinations, but the potential combinations can be limited by knowing the peptide's molecular weight. As molecular weight measurement becomes more accurate, the number of possible combinations of amino acids decreases. Despite efforts to automate the process, the combination of low mass resolution and poor accuracy made accurate de novo interpretation difficult to achieve [36, 37, 54, 55]. As genome-sequencing projects got underway, new strategies for mass spectral data interpretation emerged. Genome sequencing projects produced DNA sequences that could be translated into the amino acid sequences that evolved in a biological system. This information would then be used to further sharply limit the sequences that needed to be considered when trying to interpret tandem mass spectra.

The idea for creating an algorithm to directly search tandem mass spectra through sequence databases emerged quite simply from an effort to sequence peptides derived from class II MHC proteins [56, 57]. These peptides come from a well-studied family of immunologic proteins with a high level of sequence conservation. To minimize the time spent performing de novo sequencing on a tandem mass spectrum for a sequence that was already known, we would read off a stretch of five amino acids and then send the sequence for a BLAST search using the NCBI e-mail server. If the sequence matched to a protein, it was possible to see if the surrounding sequence of the initial 5-residue sequence fit the tandem mass spectrum. While waiting for a BLAST search to return, it occurred to me that we should simply send off the tandem mass spectrum to the database to see if a sequence would match the tandem mass spectrum. In thinking about how to achieve this, it was apparent that three immediate problems had to be solved. The first issue was how to get local access to the protein sequence database, which was stored at the National Center for Biotechnology Information (NCBI). At the time, the World Wide Web was in its infancy, so accessing sites via the Internet was not always straightforward. Assuming we could acquire a protein sequence database from the NCBI, the second obstacle was figuring out how to access the computerized tandem mass spectral data. Data files were kept in a proprietary format to maintain compliance with Federal regulations and, consequently, it was not easy to access an entire LC run's worth of tandem mass spectra. Finally, once we were able to access tandem mass spectra, we needed to figure out how to match them to sequences.

The problem of getting a database was solved quickly since increasingly they were becoming available via ftp-servers. Extracting data from the proprietary data formats of the Finnigan mass spectrometer we were using was not as straightforward. Finnigan had made a program available to us to extract data over a set of scans, but it was cumbersome to use because you needed to know the exact scan numbers for the tandem mass spectrum you wanted to extract. What was needed was a way to automatically extract all the tandem mass spectra from a file. Even though in pre-data dependent acquisition (DDA), the number of spectra collected was not large

and encompassed only those m/z values entered into the computer during the run, we still needed a way to extract all spectra at once. As a first step, we set about trying to decipher the proprietary format of the Finnigan file system and within a short time we could read the files. As a result, Finnigan gave us access to their software libraries that more faithfully accessed their file formats. Access to Finnigan's software libraries allowed us extract all MSMS files from an LC run, and this capability proved to be pivotal to fully automate the process. By granting access to their software libraries, Finnigan initiated a new era in software cooperation between academic laboratories and instrument manufacturers that continues to this day.

Once we could extract tandem mass spectra from the proprietary files, a strategy to match spectra to the sequences in the database was the next challenge. At this point, there was a significant amount of literature on spectral library searching, and some very powerful mathematical techniques had been developed to evaluate matches. Our problem was that we would not have a mass spectrum that represented the amino acid sequences extracted from the sequence databases. As fragment ions are easy to predict from amino acid sequence, a straightforward strategy for sequence matching is to predict the expected fragment ions and simply count the number of peaks the sequence shares with the experimental spectrum. We developed a scoring function based on this model, which became known as the *SEQUEST* preliminary score or Sp [58]. It calculates a score based on the number of shared peaks as a function of the expected peaks and the immonium ions present in spectrum when the amino acids that generate those immonium ions are present. This approach was tested and it worked pretty well, but was not as accurate as hoped when spectral quality was not high.

To improve scoring, we capitalized on an approach that had been employed for library searching. In the original paper, we referred to our method as a "pseudo library" search because by matching a tandem mass spectrum with a predicted one reconstructed from the amino acid sequence, we could take advantage of methods already developed for library searching. At the time it was difficult to predict the intensity of fragment ions, so a method was fashioned to recreate spectra that would minimize the impact of fragment ion intensity. The reconstructed spectrum made all the predicted b and y fragment ions the same intensity of 50%. Neutral losses of water and ammonia, a common occurrence in triple quadrupole spectra, were set at 10%. To make the experimental spectrum comply with this reconstruction, we divided the spectrum into 10 even windows across the mass range and then normalized ions within each window to 50%. We tried a few different methods to compare the experimental and reconstructed spectra, but none was satisfactory until we found a paper by Kevin Owens describing the use of a cross-correlation function as part of a mass spectral library searching approach [59]. When we implemented this approach it proved to be incredibly sensitive and accurate. However, the cross-correlation function uses a fast Fourier transform (FFT) in its calculation, making it a computationally intensive approach, particularly so in the early 1990s.

Computers have gotten faster and better implementations of the FFT calculation have been developed [60–62], while others have used the computationally less intensive “poor man’s cross-correlation” or dot product to compare spectra [21].

What was immediately clear from the early tests of the search program was that the interpretation of spectra was now a simplified- and possibly a solved-problem when databases became complete for an organism (Figure 2c). As shown in the 1994 JASMS paper, the search program also allowed the analysis of peptides obtained from intentionally digested protein mixtures. However, this data was collected by manually typing in the m/z values for the peptides, a process that required performing an LC/MS run to first measure m/z values for peptides and subsequently setting up MS/MS experiments for each m/z value of interest recorded to be analyzed in a second LC/MS analysis. This was not as cumbersome as it might sound, given that multiple windows could be open in Finnigan’s Interactive Chemical Information System (ICIS) MS control software, making it possible to cycle through the windows and trigger the MS/MS experiment for each m/z value. Clearly, this process was ripe for an infusion of new technology and it turned out it was already present in the Finnigan ICIS data system.

When the Finnigan MAT TSQ70 was designed, it contained an on-board computer to control the operation of the instrument. This computer was distinct from the computer workstation used to acquire and process the data. Also built into the system was an instrument control language (ICL). ICL was a scripting language that could be used to write little programs to operate the instrument in automated fashion, such as an autotune program. The real power of ICL was realized by Terry Lee’s laboratory through the creation of a data dependent acquisition (DDA) script [63]. This script made it possible to collect automated MS/MS over the course of an LC run. When the Lee paper came out, we were working on a DDA script of our own, so we quickly capitalized on this method and published a DDA script that used a neutral loss scan to identify the loss of phosphoric acid from phosphopeptides to trigger an MSMS of the phosphopeptides [64]. This method was published in combination with the demonstration of the *SEQUEST* algorithm as a way to identify differential modifications to peptides. By simplifying data interpretation, the analysis of post-translational modifications by mass spectrometry has become a straightforward endeavor. Because PTMs regulate biological processes, the ability to identify them in large-scale data created a paradigm shift in biology as it was now possible to analyze the state of modifications in cells as a function of different states or conditions [65].

By the mid-1990s, The Institute for Genome Research (TIGR) and academic laboratories were increasingly depositing cDNA sequences in the databases [66, 67]. In fact, since the deposition of cDNA sequence data was outstripping the deposition of whole genome DNA sequences, it made sense to develop an approach to search this data. Searching cDNA sequence data required converting the DNA sequences to protein sequences, but doing so in six frames, as the mRNA

sequencing methods sequenced randomly from either end of the transcript (e.g., 3’ end or 5’ end) [68]. This approach could be used to search genomic as well as cDNA sequences, and the upshot was that you could identify sequences within open reading frames using tandem mass spectra [69]. cDNAs represent transcribed sequences, which are most certainly proteins, but this searching concept enabled what would eventually be called “proteogenomics” as a way to identify ORFs within genomic data [70–72].

Tandem mass spectra from triple quadrupole mass spectrometers were used to develop *SEQUEST*. A question at the time was whether the approach was generalizable to other types of tandem mass spectra. MALDI-TOF mass spectrometers were capable of generating post source decay spectra where peptides fragment after acceleration into the flight tube [73]. Very often, PSD spectra have enough fragmentation to generate sequence, and when they were tested for database searching it worked quite well [74]. Although tandem double focusing magnetic sector mass spectrometers were being used less often with the development of ESI, this instrument produced tandem mass spectra with high energy collisions and lots of amino acid sidechain cleavages [75]. By adapting the sequence ions considered in the search, *SEQUEST* could be used to search tandem mass spectra. What was interesting was that a peptide containing a leucine (Leu) residue was differentiated from the same sequence in the database where the leucine residue was replaced by isoleucine (Ile). Differentiation was possible because w-ions are created in the high-energy CID spectra that are different depending on whether the sidechain is Leu or Ile [76]. The tandem mass spectra produced in ion traps also has a very different appearance from those produced in TSQs, including the lack of immonium ions in the low mass range of the spectra, but these spectra could still be used to search with *SEQUEST* [77]. A big advantage of the LCQ ion trap was a much more sophisticated DDA with the ability to exclude ions once a tandem mass spectrum was collected (dynamic exclusion), which greatly improved the efficiency of data collection.

The combination of DDA with the ability to rapidly and accurately match spectra to sequences in the database enabled revolutionary new approaches to protein analysis. The 1994 JASMS paper demonstrated the ability to collect tandem mass spectra of digested protein mixtures obtained from a yeast cell lysate [58], and although DDA was not used in the data collection, it was shown that the tandem mass spectra of peptides could be matched back to their proteins from a purposely-digested protein mixture. The invention of DDA made the process much more efficient and opened the possibility to develop a new approach to analyze protein mixtures. The development of faster scanning instruments together with more effective DDA methods has resulted in larger and larger data sets necessitating methods to speed up searches like computing clusters and methods to filter and process search results [78–82].

Understanding biology requires determining the function of proteins. Some of the information used to understand the function of proteins includes expression in response to perturbations, physical interactions, modifications, and localization. After demonstrating in 1994 that we could identify proteins in mixtures, we set out to determine if this new approach to protein analysis could be used to measure these types of parameters.

Proteins performing specific functions are frequently sequestered in specific compartments in the cell. We determined the proteins localized to the periplasmic space of *E. coli* by combining a specific method to enrich the proteins with separation of the intact proteins using ion exchange LC. Fractions of proteins were collected, digested, and then identified using LC/MS/MS. DDA was used to collect the tandem mass spectra, resulting in the identification of 80 proteins [83]. This study demonstrated that digesting the mixtures and analyzing the resulting peptides by LC/MS/MS with database searching could identify proteins. In effect, this approach could be used to identify proteins localized to a subcellular compartment of the cell, an approach that has been used to identify the components of many of the cell's compartments [84]. This strategy is very dependent on the efficacy of the initial enrichment method, which is why robust cell biological methods are important for these types of studies. The idea of purposely digesting a mixture of proteins for mass spectrometry analysis was called "shotgun proteomics."

Protein–protein interactions are a powerful tool to understand the role of proteins in cellular processes. It is based on a "guilt by association" phenomenon that assumes if proteins are interacting with each other, they must all have some role in the biological process. If you know something about the function of one of the proteins and you use this protein as a "bait" to find its interactors, the interactors are highly likely to be involved in the same biological process. As protein complexes frequently contain a small number of proteins (10s of proteins) it should be an ideal problem for direct analysis of the proteins or shotgun proteomics. We used shotgun proteomics to analyze protein complexes enriched by three different methods [85]. The methods consisted of immunoprecipitation with an antibody to a protein, affinity enrichment of binding proteins using a bound protein bait, and enrichment of a protein complex using a nondenaturing separation method. The enriched proteins were digested with trypsin and analyzed by LC/MS/MS with database searching to identify the proteins. This approach has proven to be very powerful, and numerous large-scale studies have been made to identify the protein networks of organisms [86–88].

The most complicated measurements in cells are those of protein expression. Cells contain roughly 10–15,000 expressed gene products even before considering the numerous proteoforms that could exist within the cell [89]. Complicating these measurements further is the abundance range

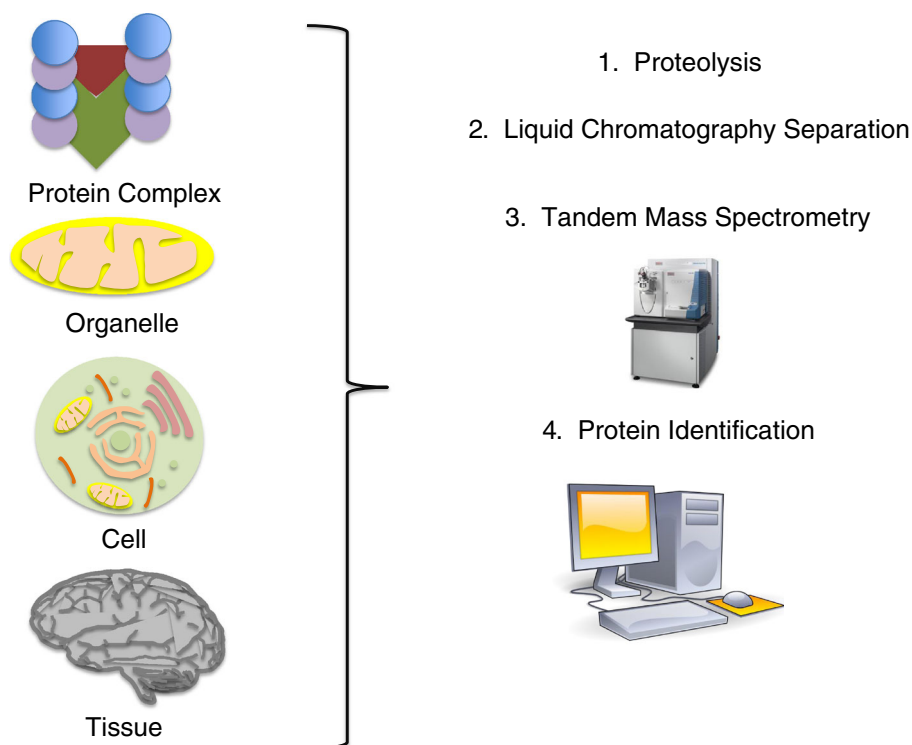


Figure 3. Interpretation of tandem mass spectra using protein sequence databases enabled the creation of new analytical paradigms such as "shotgun proteomics." Almost any cellular material or tissue can be digested with proteases, the peptides separated by liquid chromatography into a tandem mass spectrometer, and the tandem mass spectra used to identify the proteins present. This process improves with better separation technologies and methods, faster scanning mass spectrometers, and faster computers

of proteins in the cell that can exceed 10^6 , a range that is a challenge for most analytical techniques. A shotgun proteomic approach to identify proteins of the cell faces many analytical challenges as it measures an already complex mixture of proteins that is digested into peptides. While this challenges separation methods, the benefit is that it only requires a couple of peptides for a firm identification of the protein. Additionally, the chemistry of peptides is more uniform than that the chemistry of proteins and, consequently, tandem mass spectrometry methods to sequence peptides are robust and mature. Furthermore, specific features of the cell are difficult to analyze, such as membrane proteins, as they are hydrophobic and water insoluble, but shotgun methods simplify their analysis because only a few peptides need to be sequenced from the lipid-bound protein to identify it [90]. The use of the shotgun method for whole cell and tissue analysis has grown dramatically over the years and it continues to improve in terms of the numbers of proteins identified and sequence coverage of those identified proteins [91–93] (Figure 3).

The co-evolution of mass spectrometers with the computing capacity has driven innovation in mass spectrometry and increased the impact of mass spectrometers in many fields. These advances have come about not simply as an improvement in data processing but also as a result of increased operational capabilities of mass spectrometers to collect data and to combine the collection of different data types [94, 95]. The power of computer-driven data processing is the creation of more accurate and large-scale data analysis approaches that enable new experimental paradigms. The ability to search tandem mass spectra of peptides through sequence databases enabled an approach to identify intentionally digested proteins in complex mixtures, upending the established paradigm for protein discovery. Search is only the first step in the discovery process, and search results need to be filtered and assembled back into protein sequences to be more useful [78, 80–82]. Beyond search and filtering is the need to quantify results and, consequently, software tools to quantify using label-free methods, stable isotope labels, or covalent tags such as TMT were developed [93]. Many of these software tools have been combined into pipelines to streamline data processing. All of these tools and methods have made it possible to answer many biological questions, to explore the mechanisms of diseases, and to search for biomarkers of disease. A question going forward is: Can new, innovative software tools be envisioned to address as yet unanswered biological questions using mass spectrometry?

Acknowledgments

The author thanks Claire Delahunty for reading drafts of this manuscript, and members of the Yates laboratory past and present for the exciting journey. Funding has been provided

by National Institutes of Health grants P41 GM103533, R01 MH067880, 1 R01 MH100175, and HHSN268201000035C.

References

1. Biemann, K., McMurray, W.: Computer-aided interpretation of high resolution mass spectra. *Tetrahedron Lett.* **6**, 647–653 (1965)
2. Mancuso, N.R., Tsunakawa, S., Biemann, K.: Computer-aided correlation of metastable peaks in high-resolution mass spectrometry. *Anal. Chem.* **38**, 1775–1779 (1966)
3. Watson, J.T., Biemann, K.: High-resolution mass spectra of compounds emerging from gas chromatograph. *Anal. Chem.* **36**, 1135–1137 (1964)
4. Watson, J.T., Biemann, K.: Direct recording of high resolution mass spectra of gas chromatographic effluents. *Anal. Chem.* **37**, 844–859 (1965)
5. Amy, J.W., Chait, E.M., Baitinger, W.E., McLafferty, F.W.: A general technique for collecting gas chromatographic fractions for introduction into mass spectrometer. *Anal. Chem.* **37**, 1265–1266 (1965)
6. Biemann, K., Cone, C., Webster, B.R.: Computer-aided interpretation of high-resolution mass spectra. 2. amino acid sequence of peptides. *J. Am. Chem. Soc.* **88**, 2597–2598 (1966)
7. Biemann, K., Cone, C., Webster, B.R., Arsenault, G.P.: Determination of the amino acid sequence in oligopeptides by computer interpretation of their high-resolution mass spectra. *J. Am. Chem. Soc.* **88**, 5598–5606 (1966)
8. Venkataraghavan, R., McLafferty, F.W., Amy, J.W.: Automatic reduction of high-resolution mass spectral data - computer techniques for improved mass-measuring accuracy and resolution. *Anal. Chem.* **39**, 178–185 (1967)
9. Hites, R.A., Biemann, K.: A computer-compatible digital data acquisition system for fast-scanning single-focusing mass spectrometers. *Anal. Chem.* **39**, 965–970 (1967)
10. Eneroth, P., Hellstroem, K., Ryhage, R.: Identification and quantification of neutral fecal steroids by gas-liquid chromatography and mass spectrometry: studies of human excretion during two dietary regimens. *J. Lipid. Res.* **5**, 245–262 (1964)
11. Biller, J.E., Biemann, K.: Reconstructed mass-spectra—novel approach for utilization of gas chromatograph-mass-spectrometer data. *Anal. Lett.* **7**, 515–528 (1974)
12. McLafferty, F.W.: *Interpretation of mass spectra*. W. A. Benjamin, Inc., New York, NY (1966)
13. Biemann, K.: *Mass spectrometry: organic chemical applications*. McGraw-Hill, New York (1962)
14. Beynon, J.H.L.: *Mass spectrometry and its applications to organic chemistry*. Elsevier Publishing Co., Amsterdam (1960)
15. McLafferty, F.W., Gohlke, R.S.: Mass spectrometric analysis—spectral data file utilizing machine filing and manual searching. *Anal. Chem.* **31**, 1160–1163 (1959)
16. Hertz, H.S., Evans, D.A., Biemann, K.: User-oriented computer-searchable library of mass spectrometric literature references. *Org. Mass Spectrom* **4**, 453–460 (1970)
17. Hertz, H.S., Hites, R.A., Biemann, K.: Identification of mass spectra by computer-searching a file of known spectra. *Anal. Chem.* **43**, 681–691 (1971)
18. Kwok, K.S., Venkatar, R., McLafferty, F.W.: Computer-aided interpretation of mass-spectra. 3. Self-training interpretive and retrieval system. *J. Am. Chem. Soc.* **95**, 4185–4194 (1973)
19. McLafferty, F.W., Hertel, R.H., Villwock, R.D.: Computer identification of mass-spectra. 6. Probability-based matching of mass-spectra—rapid identification of specific compounds in mixtures. *Org. Mass Spectrom* **9**, 690–702 (1974)
20. Stein, S.E.: Estimating probabilities of correct identification from results of mass spectral library searches. *J. Am. Soc. Mass Spectrom* **5**, 316–323 (1994)
21. Stein, S.E., Scott, D.R.: Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom* **5**, 859–866 (1994)
22. Smith, C.A., O'Maille, G., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R., Siuzdak, G.: METLIN: a metabolite mass spectral database. *Ther. Drug Monit.* **27**, 747–751 (2005)
23. Zhu, Z.J., Schultz, A.W., Wang, J., Johnson, C.H., Yannone, S.M., Patti, G.J., Siuzdak, G.: Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the METLIN database. *Nat. Protoc.* **8**, 451–460 (2013)
24. Kind, T., Wohlgemuth, G., Lee do, Y., Lu, Y., Palazoglu, M., Shahbaz, S., Fiehn, O.: FiehnLib—mass spectral and retention index libraries for

- metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal. Chem.* **81**, 10038–10048 (2009)
25. Kumari, S., Stevens, D., Kind, T., Denkert, C., Fiehn, O.: Applying in-silico retention index and mass spectra matching for identification of unknown metabolites in accurate mass GC-TOF mass spectrometry. *Anal. Chem.* **83**, 5895–5902 (2011)
26. Smith, D.H., Buchanan, B.G., Engelmores, R.S., Duffield, A.M., Yeo, A., Feigenbaum, E.A., Lederberg, J., Djerassi, C.: Applications of artificial intelligence for chemical inference. 8. An approach to the computer interpretation of the high resolution mass spectra of complex molecules. Structure elucidation of estrogenic steroids. *J. Am. Chem. Soc.* **94**, 5962–5971 (1972)
27. Ma, Y., Kind, T., Yang, D., Leon, C., Fiehn, O.: MS2Analyzer: A software for small molecule substructure annotations from accurate tandem mass spectra. *Anal. Chem.* **86**, 10724–10731 (2014)
28. Kind, T., Fiehn, O.: Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal Rev* **2**, 23–60 (2010)
29. Herlihy, W.C., Royal, N.J., Biemann, K., Putney, S.D., Schimmel, P.R.: Mass spectra of partial protein hydrolysates as a multiple phase check for long polypeptides deduced from DNA sequences: NH₂-terminal segment of alanine tRNA synthetase. *Proc. Natl. Acad. Sci. U.S.A.* **77**, 6531–6535 (1980)
30. Nau, H., Biemann, K.: Amino acid sequencing by gas chromatography-mass spectrometry using perfluoro-dideuteroalkylated peptide derivatives. A Gas chromatographic retention indices. *Anal. Biochem.* **73**, 139–153 (1976)
31. Barber, M., Bordoli, R.S., Sedgwick, R.D., Tyler, A.N.: Fast atom bombardment of solids as an ion source in mass spectroscopy. *Nature* **293**, 270–275 (1981)
32. Hunt, D.F., Bone, W.M., Shabanowitz, J., Rhodes, J., Ballard, J.M.: Sequence analysis of oligopeptides by secondary ion/collision activated dissociation mass spectrometry. *Anal. Chem.* **53**, 1704–1708 (1981)
33. Hunt, D.F., Buko, A.M., Ballard, J.M., Shabanowitz, J., Giordani, A.B.: Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer. *Biomed. Mass Spectrom* **8**, 397–408 (1981)
34. Morris, H.R., Panico, M., Barber, M., Bordoli, R.S., Sedgwick, R.D., Tyler, A.: Fast atom bombardment: a new mass spectrometric method for peptide sequence analysis. *Biochem. Biophys. Res. Commun.* **101**, 623–631 (1981)
35. Ishikawa, K., Niwa, Y.: Computer-aided peptide sequencing by fast-atom-bombardment mass-spectrometry. *Biomed. Environ. Mass* **13**, 373–380 (1986)
36. Johnson, R.S., Biemann, K.: Computer-Program (SEQPEP) to aid in the interpretation of high-energy collision tandem mass-spectra of peptides. *Biomed. Environ. Mass* **18**, 945–957 (1989)
37. Scoble, H.A., Biller, J.E., Biemann, K.: A graphics display-oriented strategy for the amino-acid sequencing of peptides by tandem mass-spectrometry. *Fresenius Z. Anal. Chem.* **327**, 239–245 (1987)
38. Scoble, H.A., Whitaker, J.N., Biemann, K.: Analysis of the primary sequence of human myelin basic-protein peptide-1-44 and peptide-90-170 by fast-atom-bombardment mass-spectrometry. *J. Neurochem.* **47**, 614–616 (1986)
39. Biemann, K., Scoble, H.A.: Characterization by tandem mass-spectrometry of structural modifications in proteins. *Science* **237**, 992–998 (1987)
40. Hargrove, J.L., Scoble, H.A., Biemann, K.: What features of tyrosine aminotransferase determine its rapid rate of turnover in vivo. *FASEB J* **2**, A565–A565 (1988)
41. Ferone, R., Singer, S.C., Hunt, D.F.: In vitro synthesis of alpha-carboxyl-linked folylpolyglutamates by an enzyme preparation from *Escherichia coli*. *J. Biol. Chem.* **261**, 16363–16371 (1986)
42. Hunt, D.F., Yates III, J.R., Shabanowitz, J., Winston, S., Hauer, C.R.: Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **83**, 6233–6237 (1986)
43. Hunt, D.F., Yates III, J.R., Shabanowitz, J., Zhu, N.Z., Zirino, T., Averill, B.A., Daurat-Larroque, S.T., Shewale, J.G., Roberts, R.M., Brew, K.: Sequence homology in the metalloproteins; purple acid phosphatase from beef spleen and uteroferrin from porcine uterus. *Biochem. Biophys. Res. Commun.* **144**, 1154 (1987). III, 1160
44. LeVine III, H., Hunt, D.F., Zhu, N.Z., Shabanowitz, J.: Amino acid sequence analysis of the neuronal type II calmodulin-dependent protein kinase by tandem mass spectrometry. *Biochem. Biophys. Res. Commun.* **148**, 1104–1109 (1987)
45. Michel, H., Hunt, D.F., Shabanowitz, J., Bennett, J.: Tandem mass spectrometry reveals that three photosystem II proteins of spinach chloroplasts contain N-acetyl-O-phosphothreonine at their NH₂ termini. *J. Biol. Chem.* **263**, 1123–1130 (1988)
46. Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., Whitehouse, C.M.: Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64–71 (1989)
47. Patrick, A.: Combined liquid chromatography mass spectrometry. Part I Coupling by means of a moving belt interface. *Mass Spectrom Rev.* **8**, 35–55 (1989)
48. Blakley, C.R., Carmody, J.C., Vestal, M.L.: Combined liquid chromatograph/mass spectrometer for involatile biological samples. *Clin. Chem.* **26**, 1467–1473 (1980)
49. Editorial: Why sequence the human genome? *Nature* **331**, 465 (1988)
50. Mapping and sequencing the human genome. Report of the National Academy of Sciences (1988)
51. Malcolm, A.D.: The decline and fall of protein chemistry? *Nature* **275**, 90–91 (1978)
52. Escudero, I., Johnstone, M.: Genetics of schizophrenia. *Curr. Psychiatry Rep.* **16**, 502 (2014)
53. Need, A.C., Goldstein, D.B.: Schizophrenia genetics comes of age. *Neuron* **83**, 760–763 (2014)
54. Taylor, J.A., Walsh, K.A., Johnson, R.S.: Sherpa: a Macintosh-based expert system for the interpretation of electrospray ionization LC/MS and MS/MS data from protein digests. *Rapid. Commun. Mass Spectrom* **10**, 679–687 (1996)
55. Taylor, J.A., Johnson, R.S.: Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal. Chem.* **73**, 2594–2604 (2001)
56. Hunt, D.F., Michel, H., Dickenson, T.A., Shabanowitz, J., Cox, A.L., Sakaguchi, K., Appella, E., Grey, H.M., Sette, A.: Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-A. *Science* **256**, 1817–1820 (1992)
57. Hunt, D.F., Henderson, R.A., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., Cox, A.L., Appella, E., Engelhard, V.H.: Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **255**, 1261–1263 (1992)
58. Eng, J., McCormack, A.L., Yates, J.R.: An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom* **5**, 976–989 (1994)
59. Owens, K.G.: Application of correlation analysis techniques to mass spectral data. *Appl Spectrosc. Rev.* **27**, 1–49 (1992)
60. Venable, J.D., Xu, T., Cociorva, D., Yates, J.R.: Cross-correlation algorithm for calculation of peptide molecular weight from tandem mass spectra. *Anal. Chem.* **78**, 1921–1929 (2006)
61. Sadygov, R.G., Liu, H., Yates, J.R.: Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal. Chem.* **76**, 1664–1671 (2004)
62. Eng, J.K., Fischer, B., Grossmann, J., Maccoss, M.J.: A fast SEQUEST cross correlation algorithm. *J. Proteome. Res.* **7**, 4598–4602 (2008)
63. Davis, M.T., Stahl, D.C., Swiderek, K.M., Lee, T.D.: Capillary liquid chromatography/mass spectrometry for peptide and protein characterization methods. *Methods* **6**, 304–314 (1994)
64. Yates III, J.R., Eng, J.K., McCormack, A.L., Schieltz, D.: Method to correlate tandem mass of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* **67**, 1426–1436 (1995)
65. Ficarro, S.B., McClelland, M.L., Stukenberg, P.T., Burke, D.J., Ross, M.M., Shabanowitz, J., Hunt, D.F., White, F.M.: Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **20**, 301–305 (2002)
66. Adams, M.D., Soares, M.B., Kerlavage, A.R., Fields, C., Venter, J.C.: Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat. Genet.* **4**, 373–380 (1993)
67. Korenberg, J.R., Chen, X.N., Adams, M.D., Venter, J.C.: Toward a cDNA map of the human genome. *Genomics* **29**, 364–370 (1995)
68. Yates III, J.R., Eng, J.K., McCormack, A.L.: Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal. Chem.* **67**, 3202–3210 (1995)
69. Carney, J.P., Maser, R.S., Olivares, H., Davis, E.M., Lebeau, M., Yates III, J.R., Hays, L., Morgan, W.F., Petrini, J.H.J.: The hMre11/hRad50 protein complex and Nijmegen breakage syndrome: Linkage of DNA repair to the cellular DNA damage response. *Cell* **93**, 477–486 (1998)
70. Carlton, J.M., Angiuoli, S.V., Suh, B.B., Kooij, T.W., Pertea, M., Silva, J.C., Ermolaeva, M.D., Allen, J.E., Selengut, J.D., Koo, H.L., Peterson, J.D., Pop, M., Kosack, D.S., Shumway, M.F., Bidwell, S.L., Shallom, S.J., van Aken, S.E., Riedmuller, S.B., Feldblum, T.V., Cho, J.K.,

- Quackenbush, J., Sedegah, M., Shoaibi, A., Cummings, L.M., Florens, L., Yates III, J.R., Raine, J.D., Sinden, R.E., Harris, M.A., Cunningham, D.A., Preiser, P.R., Bergman, L.W., Vaidya, A.B., van Lin, L.H., Janse, C.J., Waters, A.P., Smith, H.O., White, O.R., Salzberg, S.L., Venter, J.C., Fraser, C.M., Hoffman, S.L., Gardner, M.J., Carucci, D.J.: Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* **419**, 512–519 (2002)
71. Jaffe, J.D., Berg, H.C., Church, G.M.: Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **4**, 59–77 (2004)
72. Banfield, J.F., Verberkmoes, N.C., Hettich, R.L., Thelen, M.P.: Proteogenomic approaches for the molecular characterization of natural microbial communities. *OMICS* **9**, 301–333 (2005)
73. Spengler, B., Kirsch, D., Kaufmann, R., Jaeger, E.: Peptide sequencing by matrix-assisted laser-desorption mass spectrometry. *Rapid. Commun. Mass Spectrom* **6**, 105–108 (1992)
74. Griffin, P.R., MacCoss, M.J., Eng, J.K., Blevins, R.A., Aaronson, J.S., Yates III, J.R.: Direct database searching with MALDI-PSD spectra of peptides. *Rapid. Commun. Mass Spectrom* **9**, 1546–1551 (1995)
75. Yates, J.R., Eng, J.K., Clauser, K.R., Burlingame, A.L.: Search of sequence databases with uninterpreted high-energy collision-induced dissociation spectra of peptides. *J. Am. Soc. Mass Spectrom* **7**, 1089–1098 (1996)
76. Johnson, R.S., Martin, S.A., Biemann, K., Stults, J.T., Watson, J.T.: Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine. *Anal. Chem.* **59**, 2621–2625 (1987)
77. Link, A.J., Hays, L.G., Carmack, E.B., Yates III, J.R.: Identifying the major components of *Haemophilus influenzae* type-strain NCTC 8143. *Electrophoresis* **18**, 1314–1334 (1997)
78. Tabb, D.L., McDonald, W.H., Yates III, J.R.: DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome. Res.* **1**, 21–26 (2002)
79. Sadygov, R.G., Eng, J., Durr, E., Saraf, A., McDonald, H., MacCoss, M.J., Yates III, J.R.: Code developments to improve the efficiency of automated MS/MS spectra interpretation. *J. Proteome. Res.* **1**, 211–215 (2002)
80. Nesvizhskii, A.I., Keller, A., Kolker, E., Aebersold, R.: A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75**, 4646–4658 (2003)
81. Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J., Gygi, S.P.: Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome. Res.* **2**, 43–50 (2003)
82. Kall, L., Canterbury, J.D., Weston, J., Noble, W.S., MacCoss, M.J.: Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007)
83. Link, A.J., Carmack, E., Yates III, J.R.: A strategy for the identification of proteins localized to subcellular spaces: application to *E. coli* periplasmic proteins. *Int. J. Mass Spectrom Ion Process* **160**, 303–316 (1997)
84. Yates III, J.R., Gilchrist, A., Howell, K.E., Bergeron, J.J.: Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell Biol.* **6**, 702–714 (2005)
85. McCormack, A.L., Schieltz, D.M., Goode, B., Yang, S., Barnes, G., Drubin, D., Yates III, J.R.: Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal. Chem.* **69**, 767–776 (1997)
86. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreau, M., Musk, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthies, J., Hendrickson, R.C., Gleason, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D., Tyers, M.: Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002)
87. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelman, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., Superti-Furga, G.: Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002)
88. Hazbun, T.R., Malmstrom, L., Anderson, S., Graczyk, B.J., Fox, B., Riffle, M., Sundin, B.A., Aranda, J.D., McDonald, W.H., Chiu, C.H., Snyderman, B.E., Bradley, P., Muller, E.G.D., Fields, S., Baker, D., Yates, J.R., Davis, T.N.: Assigning function to yeast proteins by integration of technologies. *Mol. Cell* **12**, 1353–1365 (2003)
89. Ramskold, D., Wang, E.T., Burge, C.B., Sandberg, R.: An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS. Comput. Biol.* **5**, e1000598 (2009)
90. Wu, C.C., MacCoss, M.J., Howell, K.E., Yates III, J.R.: A method for the comprehensive proteomic analysis of membrane proteins. *Nat. Biotechnol.* **21**, 532–538 (2003)
91. Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M., Yates III, J.R.: Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**, 676–682 (1999)
92. Washburn, M.P., Wolters, D., Yates III, J.R.: Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001)
93. Zhang, Y.Y., Fonslow, B.R., Shan, B., Baek, M.C., Yates, J.R.: Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **113**, 2343–2394 (2013)
94. Vincent, C.E., Potts, G.K., Ulbrich, A., Westphall, M.S., Atwood III, J.A., Coon, J.J., Weatherly, D.B.: Segmentation of precursor mass range using “tiling” approach increases peptide identifications for MS1-based label-free quantification. *Anal. Chem.* **85**, 2825–2832 (2013)
95. Bailey, D.J., McDevitt, M.T., Westphall, M.S., Pagliarini, D.J., Coon, J.J.: Intelligent data acquisition blends targeted and discovery methods. *J. Proteome. Res.* **13**, 2152–2161 (2014)