Data-Controlled Automation of Liquid Chromatography/Tandem Mass Spectrometry Analysis of Peptide Mixtures

Douglas C. Stahl, Kristine M. Swiderek, Michael T. Davis, and Terry D. Lee
Beckman Research Institute of the City of Hope, Duarte, California, USA

The structural characterization of proteins and peptides isolated in minute quantities requires the most efficient use of available sample. A mass spectrometer data system was programmed to continuously evaluate incoming liquid chromatography/mass spectrometry data against a user-defined array of information. The resulting conclusions were used to automatically set and modify acquisition parameters in real time to collect collision-induced dissociation spectra for selected ions (tandem mass spectrometry). This approach has provided a mechanism to target specific subsets of masses in a complex mixture and/or to discriminate selectively against masses that are known or not of interest. Masses of contaminants or peptide masses derived from known proteins can be automatically recorded and removed from further consideration for collision-induced dissociation analysis. Once recorded, these "libraries" of masses can be used across multiple analyses. This technique directs the mass spectrometer data system to focus on the analysis of masses significant to the user, even if their signal intensities are well below the intensities of contaminating masses. When combined with a database search program to correlate tandem mass spectra to known protein sequences, the identity of the protein can be established unequivocally by using less than 100 fmol of sample. *(J Am Soc Mass Spectrom 1996, 7, 532–540)*

It is axiomatic in the biological sciences that understanding the function of a protein or peptide requires determination of its structure. Unfortunately, proteins and peptides of biological interest are often purified in minute quantities and/or as components of complex mixtures. For this reason, an emphasis has been placed on increasing the sensitivity of mass spectrometry, chromatography, and Edman sequence analysis for primary structure elucidation. The development of methodologies that directly couple liquid chromatography separations to tandem mass spectral analysis (LC/MS/MS) [1–6] has significantly reduced the sample losses that inevitably occur during multistep characterization procedures. The characterization of a complex mixture typically requires multiple liquid chromatography/mass spectrometry (LC/MS) runs and consumes larger quantities of sample than are required for a single component. A thorough interpretation of the collected data for each run is usually necessary to determine how to proceed with the next analysis. This postacquisition evaluation of data often yields information that could have been used to optimize the analysis if it had been available in real time. We report the development of automated procedures that utilize incoming or previously collected data to control data acquisition. Such "data-controlled" procedures greatly increase the efficiency of LC/MS/MS analyses of peptides in complex mixtures by reducing the amount of sample and time required to perform the analysis.

Experimental Methods

Mass Spectrometry

All mass spectrometry was performed on a Finnigan-MAT (San Jose, CA) TSQ-700 triple quadrupole mass spectrometer equipped with a Finnigan-MAT electrospray ion source modified for microelectrospray as previously described [7]. The electrospray source was coupled to a gradient capillary high-performance liquid chromatography (HPLC) system built by the authors [8]. Both the mass spectrometer and the HPLC system were controlled by a Digital Equipment Corp. (Merrimack, NH) DECStation 5000/240 computer running Finnigan ICIS data system software version 7.2. Programs for data acquisition and instrument control were developed by using Finnigan Instrument Control Language (ICL) version 7.27. Copies of the programs are available from the authors.

Address reprint requests to Dr. Terry D. Lee, Division of Immunology, Beckman Research Institute of the City of Hope, 1450 East Duarte Road, Duarte, CA 91010.

1044 0305/96/$15.00
PII S1044-0305(96)00057-8

Received October 17, 1995
Revised January 17, 1996
Accepted January 22, 1996

Precursor mass spectra were collected by scanning the first quadrupole analyzer from *m/z* 500 to 2000 in 2.5 s. Collision-induced decomposition (CID) product ion spectra were collected by setting the first quadrupole analyzer to pass the precursor ion (2–6-u window) and scanning the third quadrupole analyzer over the range of *m/z* 50–2000 in 3.4 s. For both precursor spectra and CID product ion spectra, the argon collision gas in the second octupole collision cell was kept constant at a pressure of 2–3 mtorr. Precursor scanning through a pressurized collision cell attenuated the signal by a factor of 2–3. The collision energy was 17 eV for all data shown. Although the collision gas pressure and collision energies are not optimum for all peptides, empirically, these values yield useful CID spectra for the majority of peptides analyzed. The resolution of the final quadrupole analyzer was set to 1.5–2.5 u.

Automated Data Collection

Precursor ions were selected automatically by the computer based on a user-defined set of criteria. After each scan, the mass and signal-to-noise ratio (SNR) of the base peak were determined. SNR was defined as the intensity of a given mass divided by the standard deviation of the intensity of all signals in the spectrum. If the SNR was below a user-defined threshold, no masses were selected for CID analysis and the next scan was acquired (Figure 1). If the SNR prerequisite was met, the base peak mass was compared with a user-defined list of masses. The list is a floating point array that can be both read and updated between scans. The user-defined array described in this work is available through the ULIST view of the Finnigan TSQ data system. ICL programs define the mass list as either a list of masses to select for CID analysis or a list of masses to reject. If the base peak mass is rejected, the next most intense peak in the spectrum is considered. This process continues until a mass is found that meets the criteria for CID analysis or until the SNR of the peak under consideration is below the minimum value. The real time analysis of each incoming spectrum typically requires 0.5–1.5 s and is proportional to the number of masses contained in the user list.

For the work presented in this report, two different criteria were used to determine how many CID spectra to collect. For the LC/MS runs that analyze the synthetic protein digest and the serotransferrin protein digest, CID scans were performed as long as the total ion chromatogram (TIC) of the last product ion spectrum exceeded the TIC of the last precursor ion spectrum or until eight CID spectra were acquired for that mass. At low sample levels, these criteria were not sufficient to distinquish between weak but useful tandem mass spectra and spectra that were largely noise. Recently, we found that spectral reproducibility is a more useful criterion for determination of how long tandem mass spectrometry data collection should continue. As long as one of the three most abundant ions is the same mass as the base peak in the previous scan, the spectra are considered "similar" and tandem mass spectrometry data collection continues for a user-defined maximum number of scans (generally four to eight).

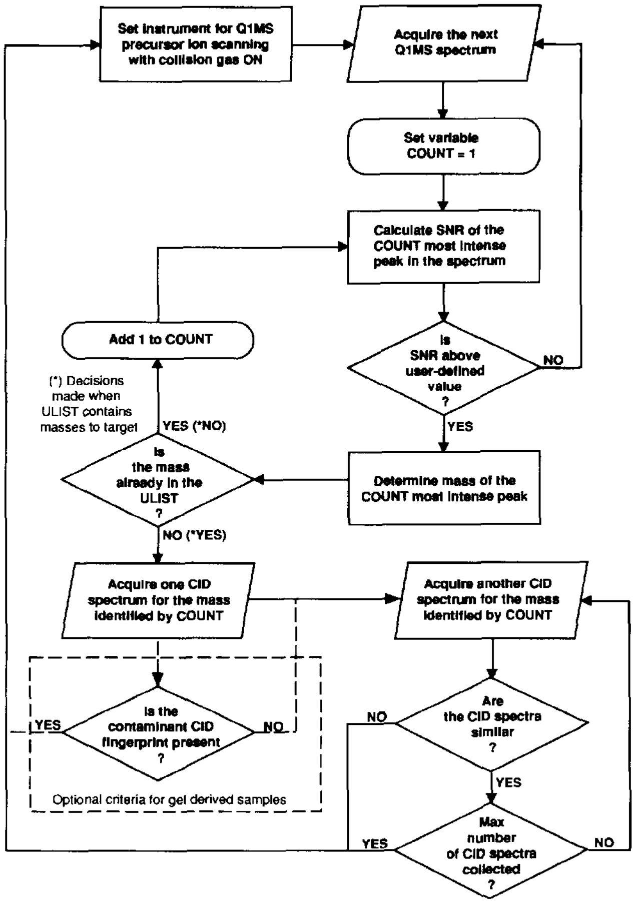
Set instrument for Q1MS precursor ion scanning with collision gas ON
Acquire the next Q1MS spectrum
Set variable COUNT = 1
Calculate SNR of the COUNT most intense peak in the spectrum
Add 1 to COUNT
Is SNR above user-defined value ?
NO
YES
(*) Decisions made when ULIST contains masses to target
YES (*NO)
Is the mass already in the ULIST ?
Determine mass of the COUNT most intense peak
NO (*YES)
Acquire one CID spectrum for the mass identified by COUNT
Acquire another CID spectrum for the mass identified by COUNT
YES
Is the contaminant CID fingerprint present ?
NO
NO
Are the CID spectra similar ?
YES
Optional criteria for gel derived samples
YES
Max number of CID spectra collected ?
NO

Figure 1. Flow chart for the computer program used for data-controlled LC/MS/MS analyses of peptide mixtures.

For the analysis of the protein samples derived from in-gel digestions, the program was modified to analyze the first collected CID spectrum for the presence of a characteristic contaminant "fingerprint." If the most abundant ion in the mass range below *m/z* 300 was 88, 133, or 175, CID data collection was stopped and the instrument was reset to collect precursor ion spectra.

Sample Preparation

The endo-Lys C digestion of horse heart cytochrome *c* was performed in 100-mM ammonium bicarbonate, pH 7.8 at an *E*:*S* ratio of 1:100 and a substrate concentration of 20 pmol/μL. The reaction was incubated overnight at 37 °C. The digest mixture was quantitated by amino acid analysis and brought to a final concentration of 5 pmol/μL in 5% acetic acid for storage.

The synthetic polypeptide used in this work was synthesized (Peptide Synthesis Core Facility, City of Hope) for an unrelated project. The sample was digested with trypsin under the same conditions as described in the preceding text. The mixture was not quantitated by amino acid analysis.

The unknown protein sample that was matched to the sequence for human serotransferrin precursor (SWISS-PROT: locus TRFE_HUMAN, accession P02787) was isolated by Dr. Steve Akman (City of Hope). The unknown protein sample that was matched to the sequence for human endoplasmin precursor (SWISS-PROT: locus ENPL_HUMAN, accession P14625) was isolated by Dr. Jennifer M. Turley (National Cancer Institute, Frederick, MD). In both cases, the investigators had separated the protein by polyacrylamide gel electrophoresis [sodium dodecylsulfate–polyacrylamide gel electrophoresis (SDS-PAGE)] and the protein bands were stained with Coomassie-blue. The protein bands were excised and digested in situ with trypsin in accordance with the protocol described by Hellman et al. [9]. The in-gel digestion was carried out in the presence of the detergent Tween-20 at a concentration of 0.1% as described in the protocol.

High-Performance Liquid Chromatography Separations

On-line separations were performed via a microscale HPLC system built by the authors [8] by using columns and procedures previously described [10]. All separations were done by using a packed capillary column 6–8 cm long and 100 μm i.d. Samples were analyzed by using linear gradients from 2 to 92% buffer B [A, 0.1% trifluoroacetic acid (TFA) in water; B, 90% acetonitrile, 0.07% TFA in water, v/v] at 100–200 nL/min. Postcolumn UV detection was performed at 200 nm via an Applied Biosystems 759A UV/VIS spectrophotometer (Applied Biosystems, Inc., Foster City, CA) equipped with a capillary flow-cell holder.

Off-line separation of the peptide mixture obtained from the in-gel digestion of endoplasmin was achieved via a C18 microcapillary column (0.53 mm i.d., 20 cm long) packed with 5 μm Vydac support (Separations Group, Hesperia, CA) [11]. A linear gradient from 2 to 90% buffer B in 60 min at 20 μL/min was used to elute the peptides. Absorbance was monitored at a wavelength of 214 nm via a SPD-6A UV/Vis spectrophotometer (Shimadzu, Columbia, MD) equipped with a modified microflow cell.

Microsequence Analysis

Microsequence analysis of peptides by automated Edman degradation was carried out via the Hewlett-Packard G1005A protein sequencing system (Hewlett-Packard, Palo Alto, CA). The sequencing system was operated by using standard reagents, solvents, and programs (routine 3.0) as supplied by the manufacturer. The total collected HPLC fraction was loaded onto the hydrophobic half of the biphasic cartridge after 1:10 dilution with 2% TFA via the sample loading station.

Tandem Mass Spectrometry Data Correlation with Protein Databases

Tandem mass spectra were correlated with the OWL nonredundant composite protein sequence database version 26.0 using the SEQUEST database searching program [6, 12].

Results and Discussion

Criteria to Change Scan Modes

Data-controlled approaches to LC/MS/MS analyses require real time analysis of each incoming spectrum. High-resolution separations of peptide mixtures place severe constraints on the amount of time available to analyze individual spectra. Typically, peaks from the liquid chromatography column are 10–30 s wide. With a typical scan cycle time of 3 s for the quadrupole analyzer, a total of 3–10 scans are available to collect data for each component. Thus, the time spent to analyze a spectrum prior to making a decision must be kept to a minimum. The simplest approach is to select ions for CID analysis based on an intensity or signal-to-noise ratio (SNR) value for the most intense ion in the spectrum. However, most ions of interest are not the most intense ion in the spectrum during the early part of an eluting peak when the decision should be made to switch modes. Also, many useless tandem mass spectra would be collected for background ions or ions whose structures are already known. A list of masses (designated ULIST in Figure 1) to which ions in the spectrum can be compared greatly increases the depth of analysis that can be applied to a spectrum. If the sample has been analyzed previously, the list can be used to designate ions for which fragment ion spectra are to be collected. Because the spectra are analyzed and ions are selected in real time, the need to have reproducible retention times for the chromatography is eliminated. Alternatively, the list can be used to exclude ions from consideration. Background ions and ions for which the structures are known are ignored. Ions for which tandem mass spectra are collected are generally added to the list automatically. This prevents computer selection of that ion again if it is still intense when the precursor mass spectrometry mode is resumed. Also, a list is generated that can be used to exclude ions from consideration in a subsequent run. Analysis times between precursor scans are typically 0.5–1.5 s dependent on the length of the list (up to 500 masses). If the list of masses is used to exclude ions from tandem mass spectrometry analysis, there exists the possibility that an ion of interest will not be considered because it has the same mass as a known ion in the list. A more detailed analysis of the spectra that utilizes elution time or order could be incorporated into the programs, but with the penalty that less time would be spent on actual collection of data. This would increase the probability that closely eluting components would be missed.

The criteria to switch back to precursor mass spectrometry mode from tandem mass spectrometry mode can be just as problematic. While the instrument collects CID spectra, it is blind to other components that elute from the column. Thus, it is important to spend no more time in the tandem mass spectrometry mode than is necessary. Generally, components present in the greatest amount yield the broadest peaks (15–30 s) and the highest quality fragment ion spectra. When a maximum tandem mass spectrometry scan number is set, no more scans are collected than are needed to ensure good quality spectra (usually four to eight). By the same token, it is senseless to collect fragment ion spectra on a precursor ion that is no longer present. For much of the work presented subsequently, decreasing signal intensity in the collected tandem mass spectrum was taken as an indication that data collection should be terminated. This method works well for peaks that are detected on their leading edge, but works poorly for ions detected after the peak maximum has passed. At lower sample levels there was a tendency to collect tandem mass spectra on transient masses and stay in tandem mass spectrometry mode for the maximum number of scans because each spectrum was random noise with a fairly constant intensity. A careful study of the differences between good CID spectra and bad CID spectra revealed that for good CID spectra, individual scans were similar to each other in appearance in that they had the same dominant ions present. Poor spectra on the other hand had a random pattern of intense ions. The criterion used in the program represented by the logic flow chart (Figure 1) was to compare the three most abundant ions in each collected tandem mass spectrometry scan with the base peak from the previous scan. If a match was found, then CID data collection continued. By this means, bad tandem mass spectrometry data are limited to two scans. The time needed to analyze tandem mass spectrometry scans and switch modes is negligible (< 0.1 s).

Performance Evaluation

For several years, our lab has used the mixture of peptides generated from the endo Lys C digestion of equine cytochrome *c* as a standard to check the performance of chromatography and mass spectrometry systems [10]. The mixture is generated easily and contains peptides that have a range of molecular weights and HPLC retention times. Decreasing amounts of this mixture were used to determine the extent of structural information that could be obtained automatically in a single LC/MS/MS experiment and to evaluate the sensitivity limitations of the method. CID analyses were performed on all masses above a SNR threshold of 8.0. For each mass selected, CID scans were performed as long as the intensity of the total ion chromatogram (TIC) of the last product ion spectrum exceeded the TIC intensity of the last precursor ion spectrum, or until five CID spectra were acquired for that mass.

After each automated LC/MS/MS run, the collected product ion spectra were correlated with the OWL protein sequence database by using the SEQUEST database search program [12]. The database search program provides an objective criteria to judge the quality of the collected tandem mass spectra. Results for analyses done at sample levels of 50, 100, 200, and 1000 fmol are summarized in Figure 2. Peptides for which CID spectra were collected are indicated by a horizontal bar under the amino acid sequence. If the correct sequence for the peptide was ranked number 1 by the SEQUEST search program, this was indicated by a solid black bar. If the correct sequence was not the first choice of the search program, this was indicated by a cross-hatched bar. Cytochrome *c* has one heme

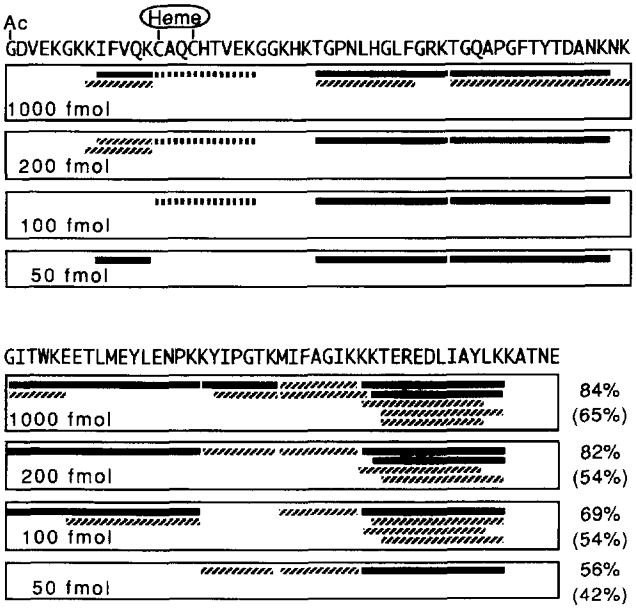
Ac
Heme
GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNK
1000 fmol
200 fmol
100 fmol
50 fmol
GITWKEETLMEYLENPKKYIPGTKMIFAGIKKKTEREDLIAYLKKATNE
1000 fmol
84%
(65%)
200 fmol
82%
(54%)
100 fmol
69%
(54%)
50 fmol
56%
(42%)

Figure 2. Results of the analysis of a peptide mixture generated from the endo-Lys C digestion of cytochrome *c* at sample levels of 1000, 200, 100, and 50 fmol. The sequence is given by using the standard single letter designation for amino acid residues. The Ac at the beginning of the sequence indicates that the N-terminus of the protein is blocked with an acetyl group. Solid bars indicate peptides whose CID spectra were matched to the correct sequence in the OWL database by the SEQUEST search program. Peptides whose sequences were not the first choice of the search program are indicated with a cross-hatched line. The peptide to which the heme is covalently attached could not possibly be matched and is indicated with a dotted bar. Numbers to the right of each data set indicate the percentage of the total sequence for which tandem mass spectra were acquired. Numbers in parentheses indicate the percentage of the total sequence that was ranked first by the SEQUEST search program.

covalently attached to the two Cys residues. Consequently, for the search parameters used, this peptide would not be found in the database and it is indicated by a dotted bar.

At higher sample levels (> 200 fmol) peptides selected for tandem mass spectrometry analysis cover 82% of the cytochrome *c* sequence. If care is taken to load the sample in a minimum volume of nearly 100% aqueous solvent and start the gradient at 0% acetonitrile, the N- and C-terminal peptides are observed. However, under normal conditions, these hydrophilic peptides elute very early in the gradient as broad peaks and are not observed. The other peptides not detected are short (1–3 amino acid residues) and hydrophilic, and are not retained on the column. In addition to the expected products of an endo Lys C digestion (cleavage after Lys residues), a number of peptides are observed due to incomplete cleavage, and in two instances, additional cleavages after Tyr and Phe.

For the experiment that used 1 pmol of sample, 18 peptides (84% of the sequence) were selected for tandem mass spectrometry analysis by the computer. The database search program matched the correct sequence to the CID spectrum for seven of the peptides (65% of the sequence). The search program ranked an additional seven peptides as second or third in its list of possible matches. At the 200-fmol level, only 12 of the cytochrome *c* peptides were selected for tandem mass spectrometry analysis. This still represents 82% of the cytochrome *c* sequence because peptides selected in the 1-pmol run and not selected in the 200-fmol run covered the same regions of the protein (Figure 2). The quality of some of the tandem mass spectra decreased. The correct sequence was ranked first by the search program for five of the peptides (54% of the sequence). Still fewer peptides (69% of the sequence) were selected for tandem mass spectrometry analysis at the 100-fmol level. The correct sequence was ranked first for only four peptides. However, the portion of the sequence covered (54%) was the same as for the 200-fmol run. When the digest mixture was analyzed at the 50-fmol level with a SNR threshold of 8, only four peptide ions (39% of the sequence) were selected for CID analysis, and the correct sequence was ranked first by the database search program for only two of them (24% of sequence; this run is not included in Figure 2).

With a second run that used a SNR threshold level of 6.5, six cytochrome *c* peptides were selected for tandem mass spectrometry analysis (56% of the sequence) and the correct sequence was matched for four of them (42% of the sequence; Figure 2). Using the lower SNR threshold, a large number (21) of other ions were selected for CID analysis. None of these other spectra was consistently matched to any protein in the database; they are probably low level contaminants, both peptide and nonpeptide, that elute from the HPLC system. The 50-fmol level represents the practical limit for our instrumentation in its present form. At this level, there are still enough quality tandem mass spectra to obtain a good match by using the SEQUEST search program without an excessive amount of interference from noise and low level contaminants.

Applications

Characterization of a synthetic protein. The data-controlled approach to LC/MS/MS analyses is particularly well suited to determine the modifications of proteins that have a known sequence. A list of masses for all of the expected digest peptides can be generated by the user and the mass spectrometer can be programed to collect CID spectra only for ions not found in the list. An example of this type of analysis is the characterization of a synthetic protein that had a mass 64 u higher than expected (expected mass = 10,342 u). A tryptic digestion was performed and analyzed in a single LC/MS/MS run. Results are summarized in Figure 3. Of the seven expected tryptic peptides (designated T_1–T_7), three were observed (T_1, T_3, and T_4) and only normal spectra were collected. Fragment T_5 is a small hydrophilic peptide (sequence NDAR) that probably was not retained on the column. A total of 13 ions (numbered 1–13 in Figure 3a) were selected for tandem mass spectrometry analysis.

Ions 2, 3, 4, 7, and 10 were of low intensity and yielded tandem mass spectra that had only one or two scans. No useful sequence data were obtained from these spectra and no assignments could be made with confidence based on the observed mass value alone. Ion 1 was assigned as the iron adduct of peptide T_3. The mass of ion 5 was consistent with the disulfide dimer of peptide T_6. Ion 6 gave a reasonably good tandem mass spectrum from which a partial sequence could be derived. However, the sequence was not related to any part of the synthetic protein and was assumed to be an impurity in the sample. The mass of ion 8 was 64 u higher than that expected for peptide T_7 and the tandem mass spectrum (Figure 3b) clearly indicated that Tyr had been substituted for Val at position 8 of the sequence. The fragment ion spectra collected for ions 9 (Figure 3c) and 11 (spectrum not shown) clearly indicated that these peptides resulted from an additional cleavage at a chymotryptic site. The mass-to-charge ratio of ion 12 (1364) corresponded to the 5^+ charge state of the disulfide dimer of T_2 complexed with iron (MW 6814). The 4^+, 6^+, and 7^+ charge states also were observed. As might be expected, the tandem mass spectrum of the 5^+ charge state gave very little useful information. The portion of the sequence that corresponded to T_2 also was observed in combination with T_1 (the result of incomplete digestion). Finally, the mass-to-charge ratio of ion 13 (1523) corresponded to the 7^+ charge state of a peptide with a molecular weight of 10,657 (five other charge states also were observed). This mass is higher than the mass of the starting protein and was assumed to be

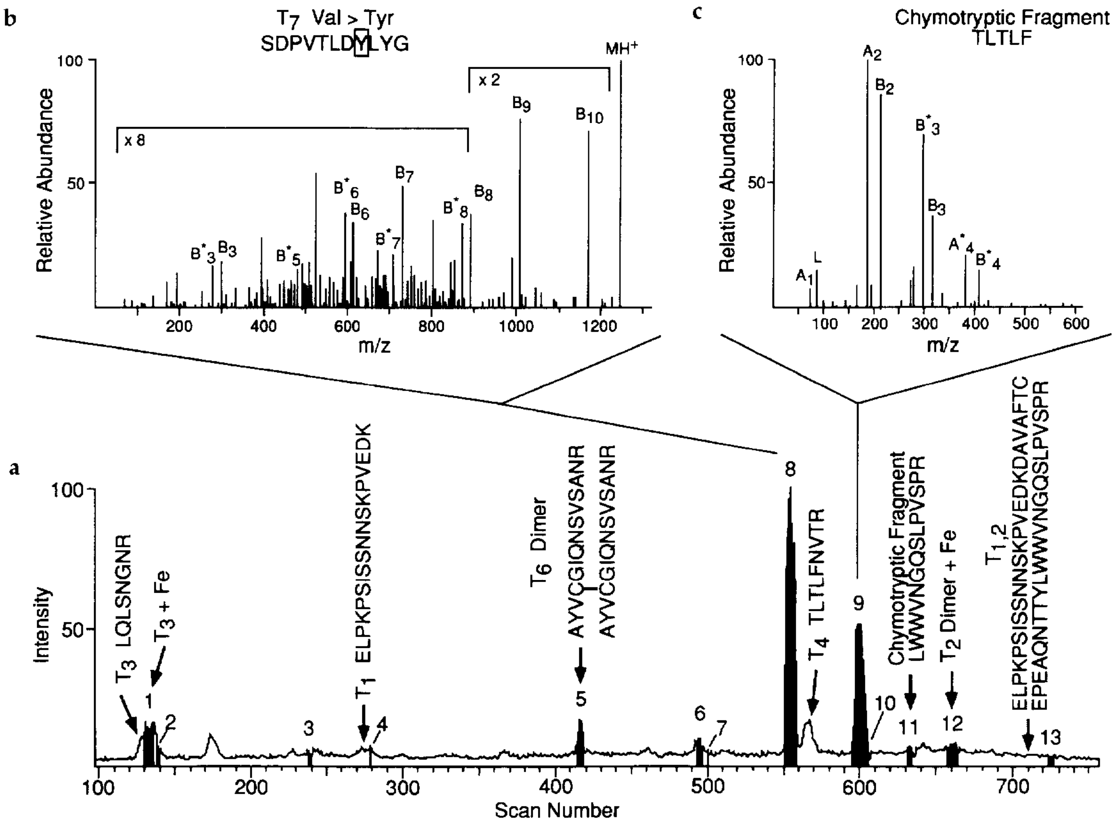

Figure 3. LC/MS/MS analysis of the tryptic digest of a synthetic protein. (**a**) Total ion chromatogram (TIC). Black bands indicate tandem mass spectra and are numbered 1–13. Peaks in the chromatogram assigned to specific tryptic peptides are marked with a tryptic peptide number (T_n) based on the order in which they occur in the sequence. The sequence of each peptide by using the single letter amino acid code also is given. Those sequences designated dimers result from disulfide bond formation between two Cys residues. (**b**) Tandem mass spectrum (peak 8) of the T_7 peptide used to determine that a Tyr had been substituted for a Val (indicated by a box around the Y in the sequence) during the synthesis. (**c**) Tandem mass spectrum for peak 9 that confirms peptide was a chymotryptic fragment.

due to an impurity, possibly a by-product of the synthesis in which some of the protecting groups were not removed.

This experiment is a good example of the problems encountered when trying to characterize "real life" protein digest mixtures by LC/MS/MS. There are always more peptides found than expected either because of foreign protein contamination or lack of specificity by the enzyme. Many of these components will be present in amounts insufficient to yield good fragment ion spectra. Even the expected peptides may have unexpected masses because of oxidation of Cys and Met residues or adduct formation with trace amounts of metal ions. Although the complete characterization of any digest mixture is always tedious, it is often possible to rapidly acquire enough information to solve a given problem. In the example just cited, a single LC/MS run was sufficient to confirm the major portion of the sequence and to determine the exact location of the error in the synthesis. Of equal importance, the necessary information was obtained by using about 1 pmol of sample.

Elimination of contaminant interference. Another problem frequently encountered is the contamination of a peptide sample with nonpeptide components. A typical example is the analysis of a protein band excised from a single SDS-PAGE gel, digested in situ, and extracted from the supporting gel. In this instance, a preliminary LC/MS analysis by using 2% of the sample was performed. The collected spectra were analyzed manually to generate a list of candidate masses for CID analysis. In a second analysis, the mass spectrometer was programed to collect CID spectra for all of the candidate masses. As a result, 26 tandem mass spectra were automatically acquired. Analysis of the spectra revealed that only 12 of the 26 CID spectra were derived from peptides. The 14 nonpeptide CID spectra were for a set of contaminant compounds derived from the polyacrylamide gel.

The contaminant compounds had different molecular weights and HPLC retention times. Consequently, there was considerable interference throughout the course of the LC/MS run. However, CID spectra for all contaminants were found to contain a characteristic "fingerprint" of ions in the low mass region of the spectrum (Figure 4). Thus, the possibility existed to quickly identify a contaminant ion from information in the first tandem mass spectrometry scan, and spend less time collecting CID spectra of nonpeptide compo-

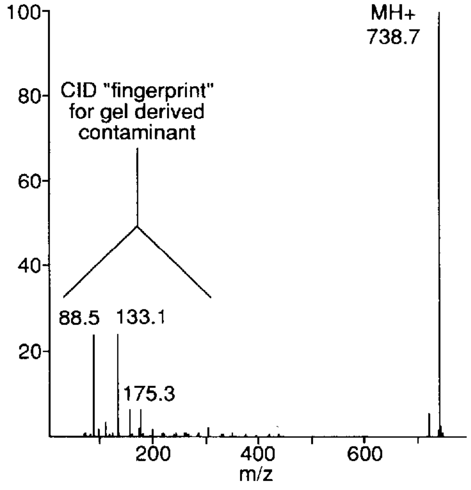

Figure 4. Tandem mass spectrum of one of the contaminant ions (m/z 738.7) derived from the polyacrylamide gel. Ions at m/z 88, 133, and 175 characterize a "fingerprint" common to members of this class of compounds.

nents. An additional ICL procedure was written that returned the mass spectrometer to the precursor scan mode whenever the base peak in the mass range below m/z 300 in the first tandem mass spectrometry scan was 88, 133, or 175. Otherwise, the normal criteria would be used to terminate tandem mass spectra collection. This procedure proved to be an effective means to detect contaminant ions without a significant increase in the time needed for the computer to analyze the spectrum. The actual time spent to analyze a contaminant ion is reduced from 18 to 4 s in the situation when five tandem mass spectra are collected.

In theory, a list of masses of known gel contaminants could be used to avoid collection of CID spectra on these components. However, given the number of these components in the mass range of m/z 500–900, there is a high probability of having peptide masses with the same values that also would be excluded. Use of HPLC retention times as a further constraint would require very reproducible chromatography, which for the most part is impractical.

A second LC/MS/MS analysis was performed by using the same amount of the protein digest sample. The ions for which CID spectra were obtained during the first LC/MS/MS run were specifically excluded as candidates for tandem mass spectrometry analysis. Otherwise, all masses above the SNR threshold were selected for CID analysis. When the contaminant fingerprint was detected, the CID analysis was terminated immediately and the parent mass was appended to a user list for rejection in all subsequent scans. A total of 35 contaminant ions were detected during the course of the run. CID spectra were collected for an additional 16 peptide ions.

The SEQUEST database search program was used to correlate the peptide CID spectra with the sequences of known proteins. Five of the spectra from the first LC/MS/MS run and four of the peptides from the second LC/MS/MS run were found to match peptides derived from human serotransferrin precursor (Figure 5). An additional four spectra had protonated molecular ion masses that matched portions of the serotransferrin sequence, but the quality of the tandem mass spectrometry data was insufficient to be confident of the sequence. All other peptide CID spectra were presumed to be derived from contaminants, disulfide dimers, or other posttranslationally modified components. (The Asn residues at positions 432 and 630 are known to be glycosylated [13].) Serotransferrin precursor is a relatively high molecular weight protein (77 ku) with 40 Cys residues. It is not surprising that many of the calculated tryptic peptides were not detected and many of the observed peptide ions could not be assigned readily, given the large number of possible disulfide-bonded dimers that could exist. Note that 13 of the 27 tryptic fragments that would not contain Cys were identified.

Mass spectrometry method compared to Edman microsequencing. The power of automated LC/MS/MS and

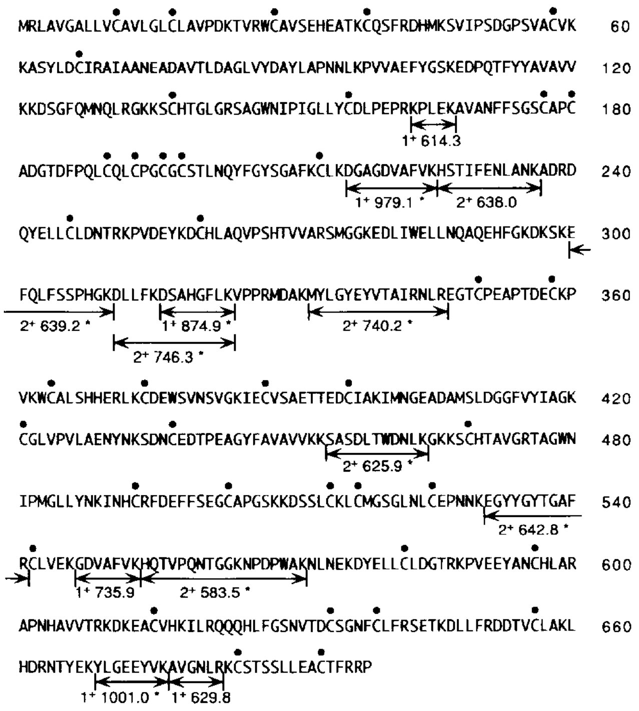
MRLAVGALLVCAVLGLCLAVPDKTVRWCAVSEHEATKCQSFRDHMKSVIPSDGPSVACVK 60
KASYLDCIRAIAANEADAVTLDAGLVYDAYLAPNNLKPVVAEFYGSKEDPQTFYYAVAVV 120
KKDSGFQMNQLRGKKSCHTGLGRSAGWNIPIGLLYCDLPEPRKPLEKAVANFFSGSCAPC 180
1+ 614.3
ADGTDFPQLCQLCPGCGCSTLNQYFGYSGAFKCLKDGAGDVAFVKHSTIFENLANKADRD 240
1+ 979.1 *
2+ 638.0
QYELLCLDNTRKPVDEYKDCHLAQVPSHTVVARSMGGKEDLIWELLNQAQEHFGKDKSKE 300
FQLFSSPHGKDLLFKDSAHGFLKVPPRMDAKMYLGYEYVTAIRNLREGTCPEAPTDECKP 360
2+ 639.2 *
1+ 874.9 *
2+ 740.2 *
2+ 746.3 *
VKWCALSHHERLKCDEWSVNSVGKIECVSAETTEDCIAKIMNGEADAMSLDGGFVYIAGK 420
CGLVPVLAENYNKSDNCEDTPEAGYFAVAVVKKSASDLTWDNLKGKKSCHTAVGRTAGWN 480
2+ 625.9 *
IPMGLLYNKINHCRFDEFFSEGCAPGSKKDSSLCKLCMGSGLNLCEPNNKEGYYGYTGAF 540
2+ 642.8 *
RCLVEKGDVAFVKHQTVPQNTGGKNPDPWAKNLNEKDYELLCLDGTRKPVEEYANCHLAR 600
1+ 735.9
2+ 583.5 *
APNHAVVTRKDKEACVHKILRQQQHLFGSNVTDCSGNFCLFRSETKDLLFRDDTVCLAKL 660
HDRNTYEKYLGEEYVKAVGNLRKCSTSSLLEACTFRRP
1+ 1001.0 *
1+ 629.8

Figure 5. Results from the LC/MS/MS analysis of the in-gel digestion of human serotransferrin precursor. The amino acid sequence is given by using single letter code. Numbers to the right are the residue numbers of the amino acids at the end of each line. For clarity, Cys residues are marked with a black dot. Peptides identified by mass spectrometry are designated by arrows under the corresponding length of sequence. Both the charge states and the observed mass-to-charge ratio values are given underneath the arrow. An asterisk to the right of the mass-to-charge ratio value indicates that the CID spectrum of the peptide matched human serotransferrin precursor in the OWL database by using the SEQUEST database search program.

database matching is further illustrated by the analysis of another protein isolated by SDS-PAGE. In this instance, 90% of the in-gel digest mixture was chromatographed and fractions were collected for sequence analysis by automated Edman degradation. One third of the remaining 10% of the sample was analyzed by LC/MS/MS by using data-controlled procedures that selected ions with an SNR greater than 8 and rejected gel-derived contaminant ions. A single LC/MS/MS run provided a wealth of information (Figure 6). A total of 26 peptide CID spectra were collected. Of these, 15 were matched to human endoplasmin precursor (which accounts for 26% of the sequence) by the SEQUEST database search program (13 as the first choice). Five of the other peptides had molecular weights consistent with logical endoplasmin tryptic fragments, but yielded CID spectra of poor quality. Three of the peptides were identified as trypsin autodigestion products. Only three of the peptide spectra remained unassigned.

Of the fractions analyzed by automated Edman degradation, only two yielded an unambiguous sequence. A search of the database for these two sequences yielded a best match to the endoplasmin protein sequence (15 out of a possible 15 residues).

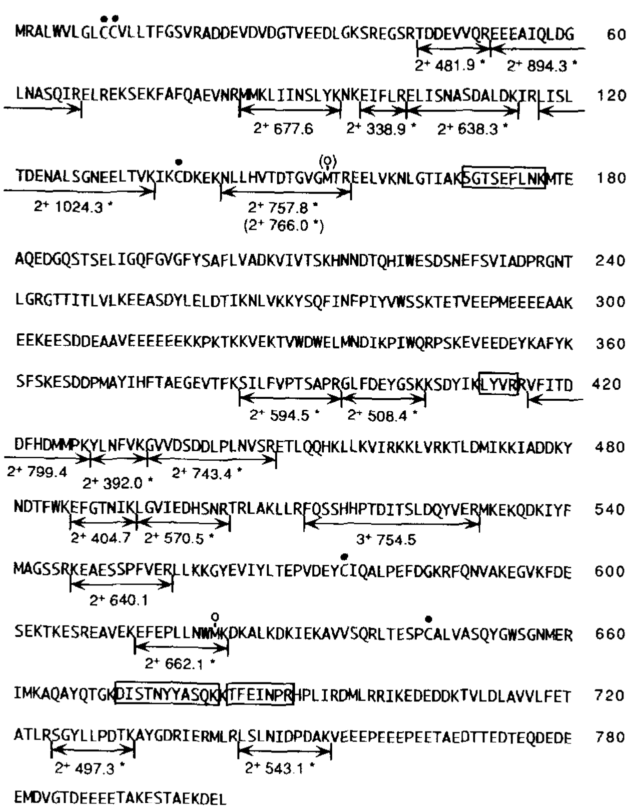
MRALWVLGLCCVLLTFGSVRADDEVDVDGTVEEDLGKSREGSRTDDEVVQREEEAIQLDG 60
2+ 481.9 *
2+ 894.3 *
LNASQIRELREKSEKFAFQAEVNRMMKLIINSLYKNKEIFLRELISNASDALDKIRLISL 120
2+ 677.6
2+ 338.9 *
2+ 638.3 *
TDENALSGNEELTVKIKCDKEKNLLHVTDTGVGMTREELVKNLGTIAKSGTSEFLNKMTE 180
(o)
2+ 1024.3 *
2+ 757.8 *
(2+ 766.0 *)
AQEDGQSTSELIGQFGVGFYSAFLVADKVIVTSKHNNDTQHIWESDSNEFSVIADPRGNT 240
LGRGTTITLVLKEEASDYLELDTIKNLVKKYSQFINFPIYVWSSKTETVEEPMEEEEAAK 300
EEKEESDDEAAVEEEEEEKKPKTKKVEKTVWDWELMNDIKPIWQRPSKEVEEDEYKAFYK 360
SFSKESDDPMAYIHFTAEGEVTFKSILFVPTSAPRGLFDEYGSKKSDYIKLYVRRVFITD 420
2+ 594.5 *
2+ 508.4 *
DFHDMMPKYLNFVKGVVDSDDLPLNVSRETLQQHKLLKVIRKKLVRKTLDMIKKIADDKY 480
2+ 799.4
2+ 392.0 *
2+ 743.4 *
NDTFWKEFGTNIKLGVIEDHSNRTRLAKLLRFQSSHHPTDITSLDQYVERMKEKQDKIYF 540
2+ 404.7
2+ 570.5 *
3+ 754.5
MAGSSRKEAESSPFVERLLKKGYEVIYLTEPVDEYCIQALPEFDGKRFQNVAKEGVKFDE 600
2+ 640.1
SEKTKESREAVEKEFEPLLNWMKDKALKDKIEKAVVSQRLTESPCALVASQYGWSGNMER 660
o
2+ 662.1 *
IMKAQAYQTGKDISTNYYASQKKTFEINPRHPLIRDMLRRIKEDEDDKTVLDLAVVLFET 720
ATLRSGYLLPDTKAYGDRIERMLRLSLNIDPDAKVEEEPEEEPEETAEDTTEDTEQDEDE 780
2+ 497.3 *
2+ 543.1 *
EMDVGTDEEEETAKESTAEKDEL

Figure 6. Results from the LC/MS/MS analysis of the in-gel digestion of endoplasmin precursor. Sequence annotation is the same as Figure 5. Rectangular boxes around portions of the sequence indicate peptides that were sequenced by Edman degradation. Two peptides detected with oxidized Met residues are annotated with an "o" above the M in the sequence.

Another two could be assigned with confidence once it was known that the peptides were derived from endoplasmin. Many of the collected fractions contained multiple peptides that made the sequence data ambiguous. With some knowledge of the identity of the protein or molecular weight information for the sample components, it is often possible to obtain useful sequence data for samples that have two or three principal components. Many of the fractions analyzed by automated Edman degradation contained levels of sample too low to yield good sequence data on our instruments. From initial yields of the microsequencing runs, the amount of sample obtained from the in-gel digestion was determined to be approximately 3 pmol. Thus, the levels of peptides in the LC/MS/MS analysis were 100 fmol or less. Mass spectrometry combined with database searching is clearly the method of choice to identify known proteins isolated by gel electrophoresis. Much less sample is required, the time needed to complete the analysis is hours rather than days, confirmation of a greater portion of the protein sequence is obtained, and complex mixtures are analyzed more easily. Microsequencing can still play a useful role in characterization of the N- and C-termini of an intact protein and can provide long stretches of unambiguous sequence needed to construct probes used to sequence the DNA that codes for an unknown protein.

Conclusions

The ability of a software interface to efficiently "self-direct" the acquisition of mass spectral information and streamline the real time acquisition of CID spectra has been demonstrated. By using signal-to-noise ratio as the criterion for mass selection, peptide ions can be isolated automatically and selected for tandem mass spectrometry analysis in real time. In an automatic LC/MS/MS analysis of a complex peptide digest mixture, partial sequence information for peptides was obtained by using less than 100 fmol of sample. Continuous evaluation of incoming data against a list of expected masses made it possible to isolate mass discrepancies with respect to expected structures, to minimize analysis time and the amount of sample consumed, and to reduce the volume of data that must be analyzed. This technique can be applied to the rapid confirmation of a site directed mutagenesis or to screen for posttranslational modifications. Conceivably, the structural analysis of an unknown protein could be performed in the presence of a known, and even more abundant, contaminating protein.

Data-controlled tandem analysis of all masses above a threshold SNR value can be improved when combined with the automatic detection and elimination of sample-related contaminant masses. This method can be used to discriminate against nonprotein contaminants associated with in-gel digestions. This approach logically can be extended to the analysis of other complex peptide mixtures such as cell lysates.

The power of data-controlled LC/MS/MS analysis combined with the SEQUEST database search program clearly has been demonstrated [12]. By using the fragment ion information in the CID spectra, the accuracy of the sequence assignment is improved greatly over that achieved by using molecular mass information alone. When combined with the microelectrospray interface, the methodology provides rapid confirmation of large portions of protein sequence at sample levels below 100 fmol.

Acknowledgments

This work was supported in part by the Public Health Service (NIH CA33572 and NIH RR06217). The synthetic protein was provided by Bruce Kaplan. We particularly thank John R. Yates for the SEQUEST database search program. A preliminary report of this work was published in the Proceedings of the 43rd ASMS Conference on Mass Spectrometry and Allied Topics, Atlanta, Georgia, May 21–26, 1995.

References

1. Huang, E. C.; Henion, J. D. *J. Am. Soc. Mass Spectrom.* **1990,** *1,* 158–165.
2. Kassel, D. B.; Shushan, B.; Sakuma, T.; Salzmann, J. P. *Anal. Chem.* **1994,** *66,* 236–243.
3. Stahl, D. C.; Martino, P. A.; Swiderek, K. M.; Davis, M. T.; Lee, T. D. *Proceedings of the 40th ASMS Conference on Mass Spectrometry and Allied Topics*; Washington, DC, 1992.
4. Davis, M. T.; Stahl, D. C.; Swiderek, K. M.; Lee, T. D. *Methods: A Companion to Methods in Enzymology* **1994,** *6,* 304–314.
5. Hunt, D. F.; Henderson, R. A.; Shabanowitz, J.; Sakaguchi, K.; Michel, H.; Sevilir, N.; Cox, A. L.; Appella, E.; Engelhard, V. H. *Science* **1992,** *255,* 1261–1263.
6. Yates, J. R. I.; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995,** *34,* 1426–1436.
7. Davis, M. T.; Stahl, D. C.; Hefta, S. A.; Lee, T. D. *Anal. Chem.* **1995,** *67,* 4549–4556.
8. Davis, M. T.; Stahl, D. C.; Lee, T. D. *J. Am. Soc. Mass Spectrom.* **1995,** *6,* 571–577.
9. Hellman, U.; Wernstedt, C.; Gonez, J.; Heldin, C.-H. *Anal. Biochem.* **1995,** *224,* 451–455.
10. Davis, M. T.; Lee, T. D. *Protein Sci.* **1992,** *3,* 935–944.
11. Swiderek, K. M.; Lee, T. D.; Shively, J. E. In *Methods in Enzymology,* Vol. 271; Karger, B. L.; Hancock, W. S., Eds.; Spectrum Publisher Services, York, PA, 1996.
12. Eng, J. K.; McCormack, A. L.; Yates, J. R. *J. Amer. Soc. Mass Spectrom.* **1994,** *5,* 976–989.
13. MacGillivary, R. T. A.; Mendez, E.; Shewale, J. G.; Sinha, S. K.; Lineback-Zins, J.; Brew, K. *J. Biol. Chem.* **1983,** *258,* 3543–3553.