Ladder Sequencing of a Peptide Using MALDI-TOF Mass Spectrometry

Owen A. Moe,* Walter A. Patton, Yun K. Kwon, and Mollie G. Kedney

Department of Chemistry, Lebanon Valley College, Annville, PA 17003, moe@lvc.edu

Received July 20, 2004. Accepted September 2, 2004.

Abstract: Matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry, an important new mass spectral technique for the analysis of biomolecules, has revolutionized experimental approaches in the biochemical sciences. We describe an experiment that demonstrates C-terminal ladder sequencing of a peptide, one of a myriad of new applications of MALDI in the area of protein chemistry. In this experiment students use carboxypeptidase Y to generate C-terminal cleavage products (ladders) of a tetradecapeptide, followed by MALDI analyses to determine the masses of the ladder components. From mass differences between successive peptide products, students determine the first seven amino acids from the C-terminal end of the peptide. They then resolve ambiguities in the sequence and identify the source of the peptide using the online BLAST search/alignment algorithm.

Introduction

Innovative mass spectral methods are transforming the study of biological and chemical macromolecules [1]. Leading the way is one especially versatile technique called matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry [2]. The MALDI technique embeds macromolecules in an organic-acid matrix that absorbs radiation from a N₂ laser at 337 nm, initiating a solid-to-gas phase change in a microvolume of the absorbing sample, desorbing the intact macromolecules, and promoting ionization. An electric-field pulse accelerates the ionized macromolecules into the drift path of a time-of-flight mass analyzer for determination of molecular mass. Desorption by MALDI is gentle, producing little or no fragmentation and yielding mainly singly charged ions [1, 2].

An essential tool in biochemical research, especially in the emerging area of proteomics, MALDI-TOF quickly and accurately determines masses of proteins, nucleic acids, and other polymers well in excess of 200,000 amu and, just as importantly, resolves and analyzes complex mixtures of smaller peptides and oligonucleotides. The biochemical applications of MALDI are many and varied: protein identification, ladder sequencing, protein modification analysis, bacterial identification, DNA sequencing, and disulfide bond localization [1–7]. The costs of low-end benchtop instrumentation are continually dropping, making such instrumentation accessible to the undergraduate laboratory.

MALDI-TOF mass spectrometry, because of its ability to provide an accurate mass profile of a collection of peptides, is an ideal tool for ladder sequencing [6, 7]. For a peptide of N amino acids, either C- or N-terminal proteolysis can generate a population of products that differ successively by one amino acid residue, yielding a “ladder” of the N, N − 1, N − 2, N − 3, etc. peptides. Enzymes commonly used for C-terminal sequencing are carboxypeptidase Y (CpY) and carboxypeptidase C (CpC) [7]. For ladder sequencing, the mass differences between successive rungs of the ladder are used to identify the amino acids released. Our Bruker Daltronics MALDI-TOF mass spectrometer, representative of other commercial linear instruments, achieves a mass accuracy of ca. 100 ppm [8], allowing the determination of peptides masses ≤ 2000 amu to within ±0.2 amu.

The monoisotopic mass of a peptide is defined as the mass of the species that contains only the most abundant isotopes (i.e., $^{12}$C, $^{1}$H, $^{16}$O, $^{14}$N, $^{32}$S, etc.), making it generally the lightest form in an isotopic cluster. For peptides < 2000 amu, the monoisotopic forms of the protonated molecular ion (M + H)⁺ peaks are generally the most abundant and, therefore, the most accurately measured in the cluster of isotope peaks. We work in the <2000 amu mass range and therefore employ monoisotopic masses in this experiment.

When an amino acid is hydrolyzed by an exopeptidase, the loss of mass is equal to the amino acid residue mass (i.e., the mass of the amino acid minus the mass of water). Table 1, which provides the monoisotopic residue masses for the twenty common amino acids, demonstrates that most amino acids can be unambiguously identified from mass differences in a generated peptide ladder. Ambiguities exist, however, for the isomeric pair leucine/isoleucine and for lysine/glutamine.

The experiment described below, designed for a low-end linear MALDI-TOF mass spectrometer, allows students to carry out C-terminal sequencing of an unknown tetradecapeptide through the first seven residues in less than two hours of laboratory time. The heptapeptide sequence so determined is analyzed using the BLAST (Basic Local Alignment Search Tool) sequence database search algorithm [9] to resolve sequence ambiguities and to identify the source of the peptide unknown.

Experimental

Proteolysis. Proteolysis buffer, 0.20 M ammonium acetate at pH 5.0, is prepared by titrating acetic acid with conc. NH₄OH. We prepare a 12.5 µg·µL⁻¹ peptide stock solution by dissolving 1 mg of renin substrate tetradecapeptide (Sigma R-8129) in 80 µL of glass-distilled water and store it at −80 °C. We prepare a 1.0 µg·µL⁻¹ stock solution of carboxypeptidase Y, CpY, by dissolving 1 mg CpY, (Sigma C-3888) in 1.00 mL of glass-distilled water and store it at −80 °C. We prepare an ACTH peptide standard at 2465.20 amu (monoisotopic M + H) by dissolving 10 nmol (24.7 µg) of ACTH 18–39 (Sigma A-8346) in 50 µL of MALDI-grade 0.10% aqueous
trifluoroacetic acid (TFA) (Sigma T-3443). We prepare a MALDI matrix solution by adding 500 µL of MALDI-grade acetonitrile (Sigma A-8596), 400 µL of glass-distilled water, and 100 µL of MALDI-grade 1.0% TFA (Sigma T-3693) to 20 mg of α-cyano-4-hydroxycinnamic acid, CHCA, (Aldrich 47,687-0), mixing to achieve a saturated solution, and centrifuging to remove any suspended matrix. The CHCA solution may be stored at room temperature in the dark for up to one week.

For the proteolysis reaction we combine in a 0.50-mL plastic microcentrifuge tube 14 µL of glass-distilled water, 3µL of 0.20 M ammonium acetate at pH 5.0, and 2 µL of 12.5 µg µL peptide stock solution. At times of 30 s, 2 min, 4 min, 8 min, and 20 min, we initiate the proteolysis reaction by adding 1 µL of 1 µg µL peptide stock solution. After pre-incubation of this solution at 35 °C for 3 to 5 min, we initiate the proteolysis reaction by adding 1 µL of 1 µg µL peptide stock solution. At times of 30 s, 2 min, 4 min, 8 min, and 20 min, we remove 1-µL aliquots from the reaction mixture and add them to separate 0.50-mL centrifuge tubes, each containing 30 µL of saturated CHCA solution. The acetonitrile in the CHCA solution immediately denatures CpY and stops the proteolytic reaction. We then add 1 µL of ACTH standard to each of the five timed peptide/CHCA solutions.

After vortex mixing, we apply in duplicate 1-µL aliquots of the peptide/CHCA solutions from each of the reaction times to a MALDI target plate and allow the ten resulting spots to air dry at room temperature for 10 min. In a sixth 0.50-mL centrifuge tube, we prepare a composite mixture of reaction times by combining 3 µL of the original timed solutions shows an overlay of individual timed runs in Figure 1a. At 8 min in Figure 1a, we see the original N-peptide and the N – 1 through N – 4 products of C-terminal cleavage. At 8 min, we see the N – 4 through N – 6 products, and at 20 min we see the N – 6 and N – 7 products. Figure 1b, achieved by mixing the equal aliquots of the original timed solutions shows a somewhat different ratio of peak intensities than does the overlay of individual timed runs in Figure 1a.

Sodium ions in the buffer salts that accompany the commercial peptide and CpY samples can replace some protons from the ladder peptides, producing sodium ion adduct peaks at 22, 44, and 66 amu greater than the molecular ion peak. Figure 1 shows that the sodium ion peaks are largely removed by the on-plate wash procedure. The monoisotopic masses expected theoretically from the known sequence and experimentally measured for the peptide ladder (student data) are summarized in Table 2. Also shown in Table 2 are the mass differences for each set of successive peptides and the amino acids that correspond to those mass differences. The student data in Table 2 yields the following seven amino acid sequence: FH(L/I)(L/I)VYS. Ambiguity in this sequence exists due to the leucine and isoleucine isomers.

Further sequencing is made difficult by the fact that the proline residue on the N-terminal side of F yields, in this particular peptide, a slow rate of proteolysis. The amino acids on the other side of the proline produce more rapid proteolysis rates, greatly reducing the size of their ladder components. Furthermore, when working below 600 amu, peaks from the CHCA matrix begin to interfere with identification of the peptide ladder peaks. We therefore elect to terminate the experiment after the elucidation of the first seven amino acids. It is often possible to resolve the ambiguities of ladder-based sequencing and to determine the source of the partially

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Symbol</th>
<th>Residue Mass, amu (monoisotopic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine</td>
<td>G</td>
<td>57.02</td>
</tr>
<tr>
<td>Alanine</td>
<td>A</td>
<td>71.04</td>
</tr>
<tr>
<td>Serine</td>
<td>S</td>
<td>87.03</td>
</tr>
<tr>
<td>Proline</td>
<td>P</td>
<td>97.05</td>
</tr>
<tr>
<td>Valine</td>
<td>V</td>
<td>99.07</td>
</tr>
<tr>
<td>Threonine</td>
<td>T</td>
<td>101.05</td>
</tr>
<tr>
<td>Cysteine</td>
<td>C</td>
<td>103.01</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>I</td>
<td>113.08</td>
</tr>
<tr>
<td>Leucine</td>
<td>L</td>
<td>113.08</td>
</tr>
<tr>
<td>Asparagine</td>
<td>N</td>
<td>114.04</td>
</tr>
<tr>
<td>Aspartic Acid</td>
<td>D</td>
<td>115.03</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Q</td>
<td>128.06</td>
</tr>
<tr>
<td>Lysine</td>
<td>K</td>
<td>128.09</td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td>E</td>
<td>129.04</td>
</tr>
<tr>
<td>Methionine</td>
<td>M</td>
<td>131.04</td>
</tr>
<tr>
<td>Histidine</td>
<td>H</td>
<td>137.06</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>F</td>
<td>147.07</td>
</tr>
<tr>
<td>Arginine</td>
<td>R</td>
<td>156.10</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Y</td>
<td>163.06</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>W</td>
<td>186.08</td>
</tr>
</tbody>
</table>

**Table 1. Monoisotopic Residue Masses of Amino Acids**
Table 2. Representative Student Data from MALDI Peptide Sequencing

<table>
<thead>
<tr>
<th>(M + H) Peptide Mass, amu</th>
<th>(M + H) Peptide Mass, amu</th>
<th>Δ Mass, amu Experimental</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theoretical</td>
<td>Experimental</td>
<td>Averages of all measurable peaks from spectra of individual and composite proteolysis times. Calibration was at 379.1 and 2465.2 amu.</td>
<td></td>
</tr>
<tr>
<td>1758.9</td>
<td>1759.2</td>
<td>87.3</td>
<td>S</td>
</tr>
<tr>
<td>1671.9</td>
<td>1671.9</td>
<td>163.1</td>
<td>Y</td>
</tr>
<tr>
<td>1508.8</td>
<td>1508.8</td>
<td>99.3</td>
<td>V</td>
</tr>
<tr>
<td>1409.7</td>
<td>1409.5</td>
<td>112.8</td>
<td>L/I</td>
</tr>
<tr>
<td>1296.6</td>
<td>1296.7</td>
<td>113.2</td>
<td>L/I</td>
</tr>
<tr>
<td>1183.5</td>
<td>1183.5</td>
<td>136.9</td>
<td>H</td>
</tr>
<tr>
<td>1046.4</td>
<td>1046.6</td>
<td>147.0</td>
<td>F</td>
</tr>
<tr>
<td>899.3</td>
<td>899.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

sequences, which give one match each to all seven amino acids, but in both cases the matches are to portions of hypothetical proteins (i.e., postulated by analyzing for possible open-reading frames in genomic sequence data).

Only the BLAST analysis of sequence FHLLVYS, for which the top five alignments are shown in Figure 2, generates exact matches in known protein sequences for all seven amino acids. Exact matches of FHLLVYS are obtained for portions of canine angiotensinogen (both a 15-amino-acid fragment of the angiotensinogen precursor and the 181-amino-acid full protein) and equine angiotensinogen (a 14-amino-acid fragment). The next best hits can match only six of the seven amino acids. We ask our students to carry out exact monoisotopic mass analyses of canine and equine angiotensinogen peptides that incorporate FHLLVYS at their C-terminal ends (online links to relevant peptide sequences are provided by BLAST). The students find that the mass of the unknown peptide, 1758.9 amu, matches exactly the monoisotopic mass (M + 1) of the sequence, DRVYIHPFHLLVYS, which is the known N-terminal sequence of canine angiotensinogen precursor tetradecapeptide [13] and, as well, the known N-terminal sequence of canine angiotensinogen precursor tetradecapeptide [14]. Thus, from partial C-terminal sequencing combined with database searching and mass analyses, it is possible for students to assign an unambiguous full sequence to the unknown peptide and to deduce its possible origin.

It should be noted that CpY, while it will hydrolyze all amino acids at pH 5, moves relatively slowly through glycine, amino acids C-terminal to a penultimate glycine, and aspartate [7]. A mixture of CpY and CpP is often used to achieve more consistent rates through all amino acids [7]. We have selected a peptide that works well with CpY alone, simplifying the experiment.

We consider this experiment to be reliable and rugged, especially if the CpY and peptide solutions are made in advance and tested by the instructor. We recommend protease and peptide storage at –80 °C, but –20 °C can be used. Students may need some guidance in working with the very small volumes involved in this experiment, as great care must be taken in pipetting to setup the enzymatic reactions as well as spotting the MALDI target.

After completing the experiment, the student teams give brief presentations that include the rationale for ladder sequencing, the mass spectral data collected, and their interpretation of that data. Of particular note in these presentations is the student interpretation of ambiguity in the assignment of amino acid residues and their discussion of the BLAST search in resolving that ambiguity. Their goal is to
Query Sequence - FHLLVYS

Alignment #1

gi|14188313|pep|A60834 angiotensin I precursor - dog (fragment)
Length = 16 Score = 26.5 bits (55), Expect = 70
Identities = 7/7 (100%), Positives = 7/7 (100%)
Query:  FHLLVYS
PHLLVYS
Sbjct:  0 PHLLVYS 14

Alignment #2

gi|5925267|gb|BA83741.1 angiotensinogen [Canis familiaris]
Length = 191 Score = 26.5 bits (55), Expect = 70
Identities = 7/7 (100%), Positives = 7/7 (100%)
Query:  FHLLVYS
PHLLVYS
Sbjct:  5 PHLLVYS 11

Alignment #3

gi|133879|sp|P01016|ANGOT_HORSE Angiotensinogen [Contains: Angiotensin I (Ang I); Angiotensin II (Ang II); Angiotensin III (Ang III) Des-Asp[1]-angiotensin II]
gi|98117|gi|A6129O angiotensin precursor - horse (fragment)
Length = 14 Score = 26.5 bits (55), Expect = 70
Identities = 7/7 (100%), Positives = 7/7 (100%)
Query:  FHLLVYS
PHLLVYS
Sbjct:  9 PHLLVYS 14

Alignment #4

gi|7837361|ref|NP 496286.1 putative endoplasmic reticulum protein, with at least 4 transmembrane domains (Chaetopterus variopedatus)
gi|7495661|gi|T15249 hypothetical protein C14A4.7 - Caenorhabditis elegans
gi|3374234|emb|CAB9911.1 Hypothetical protein C14A4.7a - Caenorhabditis elegans
Length = 194 Score = 24.4 bits (50), Expect = 306
Identities = 6/7 (85%), Positives = 7/7 (100%)
Query:  FHLLVYS
PHLLVYS
Sbjct:  34 PHLLVYS 40

Alignment #5

gi|39422249|emb|CAB4598.1 Hypothetical protein C14A4.7b - Caenorhabditis elegans
Length = 143 Score = 24.4 bits (50), Expect = 306
Identities = 6/7 (85%), Positives = 7/7 (100%)
Query:  FHLLVYS
PHLLVYS
Sbjct:  34 PHLLVYS 40

Figure 2. BLAST alignment data. Shown are the top five hits with our query sequence using the BLAST algorithm for short sequences with nearly exact matches. The results shown were generated on March 17, 2004. The identity of each matching protein is indicated along with its database accession annotations. Functionally conserved residues are indicated by a plus (+) in the consensus line and identical residues are shown in bold. The formatting here differs only slightly from that returned by the BLAST server.

This experiment provides some appreciation of the versatility inherent in MALDI, a technique that can translate its ability to determine accurate masses into a simple and rapid method for sequencing polypeptides. This experiment also illustrates the powerful synergy between experimental peptide data and online sequence information, a synergy that is mediated by Web-based search algorithms. Thus, students who generally would not have sufficient laboratory time to work through the complex and time-consuming process of Edman sequencing can, with this experiment, fully sequence a tetradecapeptide during a single laboratory period.

Acknowledgment. We gratefully acknowledge support from CCLI-A&I grant number 0126950 from the National Science Foundation.

References and Notes