

Biological Mass Spectrometry: Proteomics

Table of Contents

Introduction	2
Information on how to use the module	3
Section 1. Proteins and Proteomics	5
Section 1A. What is a protein?	5
Section 1B. What is proteomics?	10
Section 2. Calculating the Molecular Weight of a Protein from its Electrospray Ionization Mass Spectrum (ESI-MS)	15
Section 2A. Electrospray and Quadrupole Ion Trap Instrumentation	15
Section 2B. ESI-MS Data	17
Section 3. Peptide Mass Mapping for Protein Identification	21
Section 3A. MALDI-TOF Mass Spectrometry	22
Section 3B. Time of Flight (TOF) Mass Analyzer	26
Section 3C. Peptide Mass Mapping for Protein Identification	34
Section 3D. MASCOT Database Search	39
Section 4: Sequencing Peptides from Tandem MS (MS-MS) Data	56
Section 4A. Tandem MS	56
Section 4B. CID of Peptides and <i>De Novo Sequencing</i>	57

Introduction

This module is an introduction to biological mass spectrometry with an emphasis on proteomics, the study of proteins, particularly their structures and functions. The first section provides key background information on proteins and proteomics. The second section describes the use of electrospray ionization to determine the molecular weight of a protein using an ion trap mass analyzer. The third section shows how to identify an unknown protein using peptide mass fingerprinting with matrix-assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF). The fourth section describes shotgun proteomics and tandem MS experiments for *de novo* peptide sequencing.

Each section includes some brief explanatory text with accompanying **reading questions** that emphasize student comprehension. The text and accompanying reading questions are intended to be completed as a pre-class assignment so that students have the background information needed to work on more difficult questions during the class period. Alternatively, the reading questions could be completed at the start of the class period. Each section also includes **discussion questions** and activities that are intended for in-class work in small groups. The instructor's manual gives more detailed suggestions for how different portions of the module might be used and the amount of class time needed to complete each section.

Information on how to use this module

What prior knowledge do students need before beginning this module?

It is helpful if students have had at least one class period on the basics of mass spectrometry before starting this module. For example, students should understand that analytes must be ionized for mass spectrometric analysis so that the molecules may be separated by m/z . No prior knowledge of proteomics, specific ion source, or mass analyzers is needed because these topics are covered in the module.

The information in Section 1 is a necessary prerequisite to Sections 2-4, but Sections 2-4 are independent of each other. An individual instructor could complete one or two of these sections without the others and in any order. Students who have taken biochemistry may not need Section 1A, which is an introduction to amino acids, the peptide bond, and proteins.

How much time is required in and out of class to complete each section?

Section 1: Proteins and proteomics

Class Period 1: (Section 1A What is a protein?) Reading questions should be completed prior to class. Discussion questions can be answered in small groups during class.

Class Period 2: (Section 1 B Proteomics) Reading questions should be completed prior to class. It is best to summarize and discuss the answers to this set of reading questions at the start of class. During class, distribute the following comparative proteomics paper and have students work in groups to answer the discussion questions.

Comparative proteomics of oral cancer cell lines: identification of cancer associated proteins, Karsani *et al*, Proteome Science, 2014, **12**:3. doi:10.1186/1477-5956-12-3 (Open access journal)

Section 2: Calculating the molecular weight of a protein from its electrospray ionization mass spectrum

Class Period 1: Reading and video questions should be completed prior to class. Discussion questions can be answered in small groups during class. (If all questions are completed during the class period, then two class sessions are needed.)

Section 3: Identification of a protein: Peptide mass mapping

Class Period 1 and 2: Two full class periods are suggested for the reading and discussion of MALDI and TOF-MS (Sections 3A-B).

Class Period 3: Section 3C Reading questions should be completed prior to class. Discussion questions can be answered in small groups during class.

Class Periods 4 and 5: Section 3D Reading questions should be completed prior to class. Students will need a computer to access the maxtrix science website and mascot database. It is best if students download an excel file of the mass list and then they can simply copy and paste the m/z values into the search query. Students will explore how changing search parameters affects the protein score (statistical significance that a protein has been correctly identified). The varying protein scores must be interpreted.

A series of homework problems accompanies Section 3D. These are meant to be assigned out of class but could also be used during another class period.

Section 4: Sequencing peptides and identifying proteins from tandem MS (MS-MS) data

Class Period 1 and 2: If the reading and video questions are completed prior to class, the discussion questions can be completed in one to two class periods, depending on whether the groups work through sketching fragment ions and calculating masses together or divide the work among group members. To save time, the instructor may allow students to use the Institute for Systems Biology's [peptide calculator](#). The Challenge Question may be assigned as homework or worked in class during a class session. If readings, videos, and discussions will all occur in class, at least three class periods are recommended.

What is the difference between reading questions, video questions, and discussion questions?

Reading questions and video questions are short answer questions that are relatively straightforward. Individual students should be able to answer these questions on their own without group discussion. Reading questions are based on reading the text portion of the module, and video questions are based on linked videos. Students could be assigned to read and view videos before coming to class. This will ensure that students have the requisite background knowledge to discuss the more open-ended and challenging discussion questions in small groups during class. It is best to briefly re-cap the answers to the reading questions at the start of the class period.

Alternatively, students could complete the reading and video questions independently during class; this works especially well for shorter sections like the introduction to proteins (Section 1). Discussion questions should always be used in-class where students discuss the answers in groups of 3-4 students. A good strategy is to have each group discuss a question for several minutes while the instructor circulates to answer questions, correct misunderstandings, and monitor discussions. Then the instructor can optionally call the class together as a group to report out answers and for large group discussion before moving on to the next question. For the time-of-flight portion (Section 3B), there is an accompanying slide show that includes animations to be shown as a wrap-up to discussions.

Section 1: Proteins and Proteomics

Learning Objectives

At the end of this assignment, students will be able to:

1. Draw the chemical structure of an amino acid and small peptide.
2. Describe the difference between free and residue amino acid mass.
3. Calculate the monoisotopic and average mass of a peptide and explain why they are different.
4. Define proteomics and interpret a scientific paper on the application of proteomics in cancer research.

Section 1A. What is a protein?

Proteins are biological macromolecules consisting of long chains of amino acids. A shorter chain of amino acids is called a **peptide**. Proteins and peptides are biological polymers formed from amino acid monomers. Each amino acid is composed of an amine group, a side chain (R), and a carboxylic acid functionality (Figure 1).

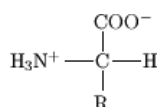


Figure 1. An amino acid.

There are 22 naturally occurring amino acids, 20 of which are encoded by the genome (Figure 2). These amino acids are usually grouped according to the character of their side chains, which may be acidic, basic, neutral, or hydrophobic. In a protein, portions of the linear sequence of amino acids may take on a secondary structure (such as a helix or a sheet) based on intermolecular forces between amino acid residues. The full-length protein will form a tertiary structure or overall shape, and the structure of the protein is intimately linked to the protein's function.

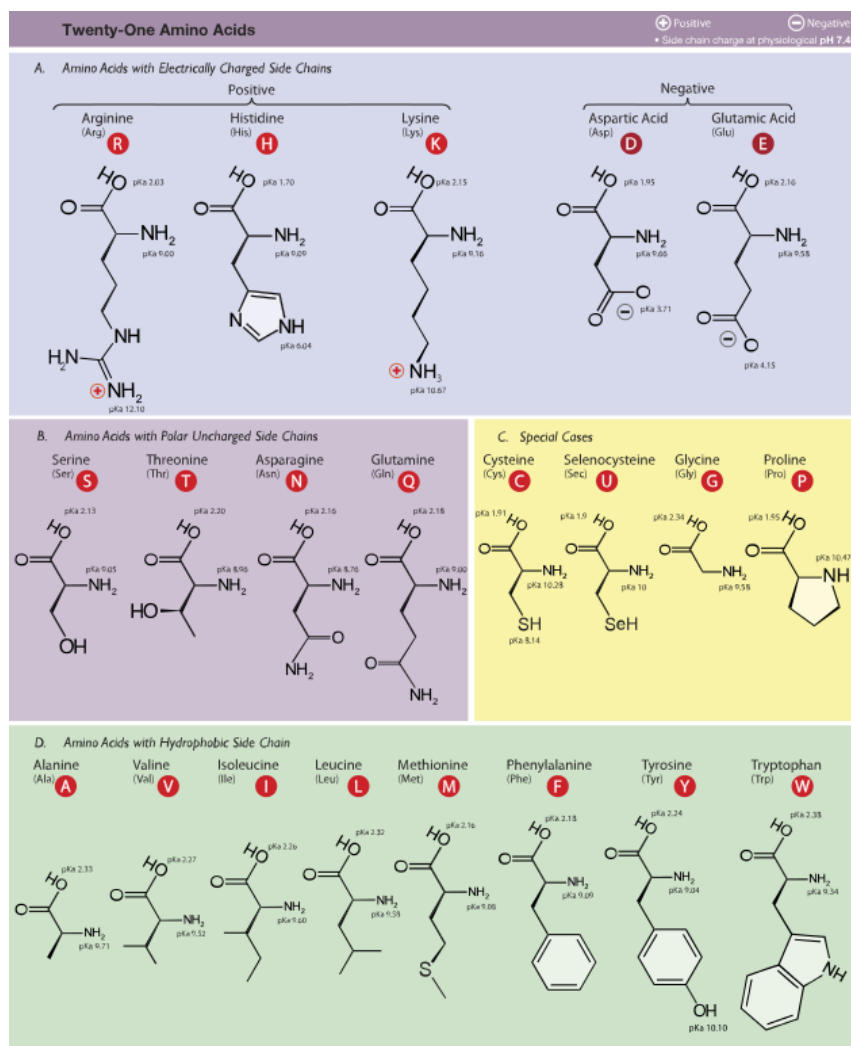
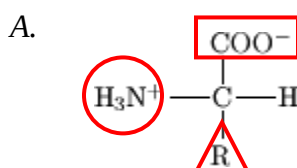


Figure 2. Table of naturally occurring amino acids sorted by side chain and with their three-letter and one-letter codes. (Reproduced under a Creative Commons license from Dancojocari.)

Reading Question

1. In Figure 1 above, circle the amine group, draw a triangle around the side chain, and draw a square around the carboxylic acid group of the amino acid.



The linear chains of amino acids that form proteins are connected by amide (or peptide) bonds that form by a dehydration reaction between two amino acids (Figure 3). By convention, we write and draw peptides from the N-terminus (the side with the free amine group) to the C-terminus (the side with the free carboxylic acid).

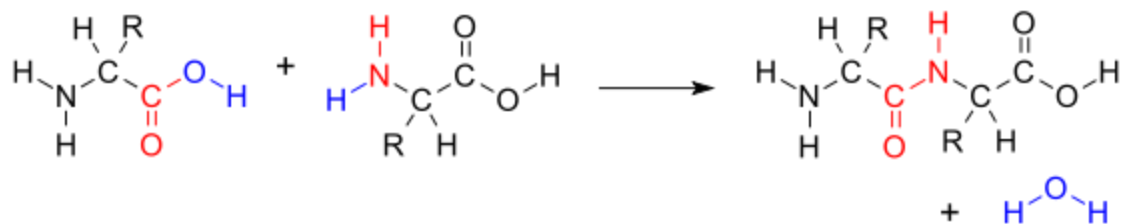
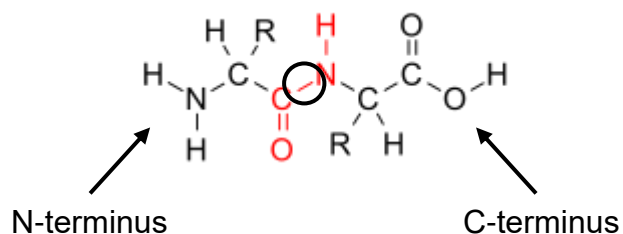


Figure 3. Formation of a peptide bond between two amino acids.

Reading Question

2. In Figure 3 above, label the N-terminus and the C-terminus of the dipeptide product. Circle the amide bond.

A.



While many properties of peptides and proteins can be measured to provide useful information, we will focus on measurements of peptide and protein masses using mass spectrometry. Mass spectrometry measures the mass and charge of molecules in the gas phase. As a result, the mass of individual molecules is important. The **monoisotopic mass** of a molecule is the mass of that compound when it is composed of the most abundant isotope of each atom in the molecule. For example, if we calculate the monoisotopic mass of CO_2 , we will use the masses of carbon-12 and oxygen-16 since these are the most abundant isotopes of these elements. This gives us a monoisotopic mass of $12.000 \text{ amu} + 2 (15.995 \text{ amu}) = 43.99 \text{ amu}$. This is different than the molar mass we would calculate using average atomic weights from the periodic table; on the periodic table, we find a molar mass for carbon of 12.011 amu . This represents the weighted average of all carbon atoms, which includes 98.9% carbon-12 and 1.1% carbon-13. Table 1 shows the monoisotopic mass of each amino acid and the mass of that amino acid as a residue in a protein or peptide chain.

Table 1. Monoisotopic molecular weight information for all 20 genetically encoded, naturally occurring amino acids.

Amino Acid	Single-Letter Code	Residue MW (amu)	Amino Acid MW (amu)
glycine	G	57.02	75.03
alanine	A	71.04	89.05
serine	S	87.03	105.04
proline	P	97.05	115.06
valine	V	99.07	117.08
threonine	T	101.05	119.06
cysteine	C	103.01	121.02
isoleucine	I	113.08	131.09
leucine	L	113.08	131.09
asparagine	N	114.04	132.05
aspartic acid	D	115.03	133.04
glutamine	Q	128.06	146.07
lysine	K	128.09	146.11
glutamic acid	E	129.04	147.05
methionine	M	131.04	149.05
histidine	H	137.06	155.07
phenylalanine	F	147.07	165.08
arginine	R	156.10	174.11
tyrosine	Y	163.06	181.07
tryptophan	W	186.08	204.09

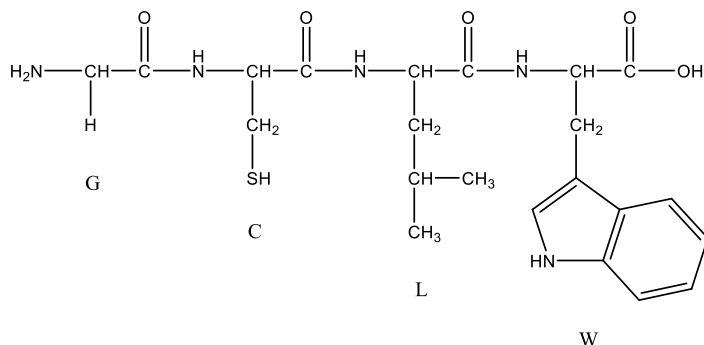
Discussion Questions

1. Consider the data in Table 1. By what value do the residue molecular weight (MW) and the amino acid MW differ? Why is the MW of an amino acid residue in a peptide chain different from the mass of the full amino acid?

A. The mass of the amino acid residue is 18 amu lighter than the mass of the amino acid because the formation of peptide bonds at the N- and C-termini results in the loss of water from the amino acid.

2. Draw the structure for the tetrapeptide G, C, L, W. Refer to Figure 2 for the structure of amino acid side chains.

A.



3. Calculate the monoisotopic molecular weight of the tetrapeptide using the data in Table 1.

A. When computing a peptide or protein molecular weight, sum up all residue masses and add 18.02 (for the H and OH at the termini):

$$57.02 + 103.01 + 113.08 + 186.08 + 18.01 = 477.20 \text{ amu}$$

Note: In protein mass spectrometry, the Da is a common unit. It is the same as amu.

4. Calculate the molecular weight of the tetrapeptide using molar mass information from the periodic table. Why is this molecular weight different from the monoisotopic mass you calculate in question 3? Which mass is the mass measured in mass spectrometry?

A. Glycine is $C_2H_5NO_2$. Cysteine is $C_3H_7NO_2S$. Leucine is $C_6H_{13}NO_2$. Tryptophan is $C_{11}H_{12}N_2O_2$.

This is a total of 22 carbons, 37 hydrogens, 5 nitrogens, 8 oxygens, and 1 sulfur atom. For the peptide, we must subtract the equivalent of 3 water molecules representing dehydration to form the peptide bonds (i.e., total atom count should be less 6 hydrogens and 3 oxygens). This gives us 22 C, 31 H, 5 N, 5 O, and 1 S.

$$22 (12.011) + 31 (1.008) + 5 (14.007) + 5 (15.999) + 1 (32.06) = 477.58 \text{ Da.}$$

The periodic table gives the weighted average molar masses of each element. This is the average mass of a population of GCLW peptides. No single GCLW peptide actually has this mass, and it is slightly heavier than the monoisotopic mass since the most abundant isotopes of H, C, N, and O are the lightest isotopes. The monoisotopic mass is the mass of a GCLW peptide composed of atoms that are the most abundant isotope of each element in the peptide. This will be the most intense peak on the mass spectrum.

Section 1B. What is proteomics?

The central dogma of molecular biology, DNA to RNA to protein, has given us an explanation of how information encoded by our DNA is translated and used to make an organism. It describes how a gene made of DNA is transcribed by messenger RNA and then translated into a protein by transfer RNA in a complex series of events utilizing ribosomal RNA and amino acids. Although in essence the central dogma remains true, studies of genes and proteins are revealing a complexity that we had never imagined. For example, distinct genes are expressed in different cell types and the physiological state of the cell alters which proteins are produced and at what level. Furthermore, chemical changes (i.e. phosphorylation) to proteins occur after translation and are critical to a protein's function. The importance and diversity of proteins started a whole new field termed proteomics.

Proteomics is the study of proteins, particularly their structures and functions. This term was coined to make an analogy with genomics. The Human Genome Project, started in 1990 and completed in 2003, sequenced three billion bases in genes (the human genome). The entire set of proteins in existence in an organism throughout its life cycle, or on a smaller scale the entire set of proteins found in a particular cell type under a specific set of conditions is referred to as the **proteome**.

Proteomics is much more complicated than genomics for several reasons. The genome is a rather constant entity while the proteome differs from cell to cell and is constantly changing through its biochemical interactions with the genome and the environment. Consequently, the proteome reflects the particular stage of development or the current environmental condition of the cell or organism. One organism will have radically different protein expression in different parts of the body, in different stages of its life cycle, and in different environmental conditions. For example, when *E. coli* cells are grown under conditions of elevated temperature a class of proteins known as heat shock proteins are upregulated. Many members of this group perform a chaperone function by stabilizing new proteins to ensure correct folding or by helping to refold proteins that were damaged by the cell stress. Ultimately, the comparison of proteomes of healthy and diseased tissues may identify the molecular nature of a disease and provide potential new targets for drug development. The field of proteomics also presents many analytical challenges when compared to genomics. In DNA there are only four nucleotide bases with similar molecular weights and properties. In a proteome there are thousands of different proteins with a wide range of concentrations, molecular weights, and properties.

Proteomics was initially defined as the effort to catalog all the proteins expressed in all cells at all stages of development. That definition has now been expanded to include the study of protein functions, protein-protein interactions, cellular locations, expression levels, and post-translational modifications of all proteins within all cells and tissues at all stages of development. It is hypothesized that a large amount of the non-coding DNA in the human genome functions to regulate protein production, expression levels, and post-translational modifications. It is regulation of our complex proteomes, rather than our genes, that makes us different from simpler organisms with a similar number of genes. An international collaboration of scientists in the human Proteome Project (HPP) is working to characterize all 20,300 genes of the known genome and generate a map of the protein based molecular architecture of the human body. Completion of this project will enhance understanding of human biology at the cellular level and lay a foundation for development of diagnostic, prognostic, therapeutic, and preventive medical applications.

Reading Questions

1. Define the term **proteome.**

A. The proteome is the entire set of proteins that an organism expresses.

2. Define the term **proteomics.**

A. Proteomics is the large scale and systematic study of a proteome and all the various information associated with it including cellular locations, protein abundances, modifications, and interacting partners and networks.

3. Why is the analysis of proteins in a cell more difficult than sequencing DNA?

A. The proteome is constantly changing based on the state of the cell.

Proteins have widely varying concentrations, molecular weights, and properties.

There are thousands of different proteins in a cell requiring separation methods.

4. What types of questions can be answered by studying the proteome?

A.

Example 1: How does protein expression change between different types of cells and stages of development?

Example 2: What protein(s) are related to a disease? What drug would be useful treatment? If a certain protein is implicated in a disease, its three dimensional structure provides the information needed to design drugs that interfere with the action of the protein. A molecule that fits the active site of an enzyme, but cannot be released by the enzyme, inactivates the enzyme. This is the basis of new drug-discovery tools, which aim to find new drugs to inactivate proteins involved in disease.

Example 3: What post-translational modifications occur and how are they involved in the regulatory mechanism of the cell?

Read the following research paper to learn how the field of proteomics can be useful in the treatment of cancer.

Comparative proteomics of oral cancer cell lines: identification of cancer associated proteins, Karsani *et al*, Proteome Science, 2014, **12**:3. doi:10.1186/1477-5956-12-3 (Open access journal)

Discussion Questions

1. What was goal of the scientific study reported in the paper?

A. To determine differences in protein expression in oral cancer cells compared to normal healthy cells. The goal is to identify cancer associated proteins.

2. Why would a study of the change in oral cancer proteins add to our understanding of the disease?

A. It is known that the mechanism of oral cancer involves an activation of oncogenes which results in a change in expression of various proteins. This research could help identify candidate proteins for biomarkers and early detection.

3. Refer to the results and discussion section and Figure 1 to answer the following questions.

a. How were the proteins in the healthy and cancerous cells separated and detected?

A. Two dimensional gel electrophoresis

First dimension is isoelectric point (pH where protein is neutral)

Second dimension is size (MW)

b. How many individual protein spots were resolved on the silver stained gels?

A. More than 1000

c. How many protein spots exhibited a significant difference in abundance from normal cells to cancerous cells?

A. 24

4. Table 1 is a list of proteins with different abundances in the cancer cell line.

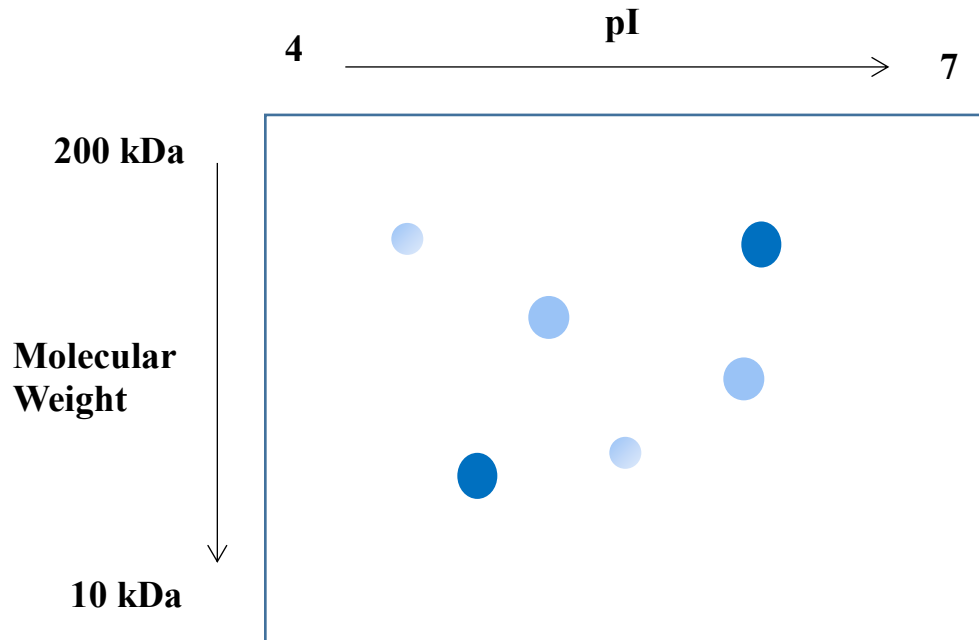
(The section in this module on peptide mass mapping describes how the identity of the protein in the gel spot was determined.)

Examine the data for two structural proteins: Stathmin (STMN1) and myosin regulatory light chain-2 (ML12A). What is the change in abundance for each protein? Can this change be visualized from the image of the spot? How is the change quantified?

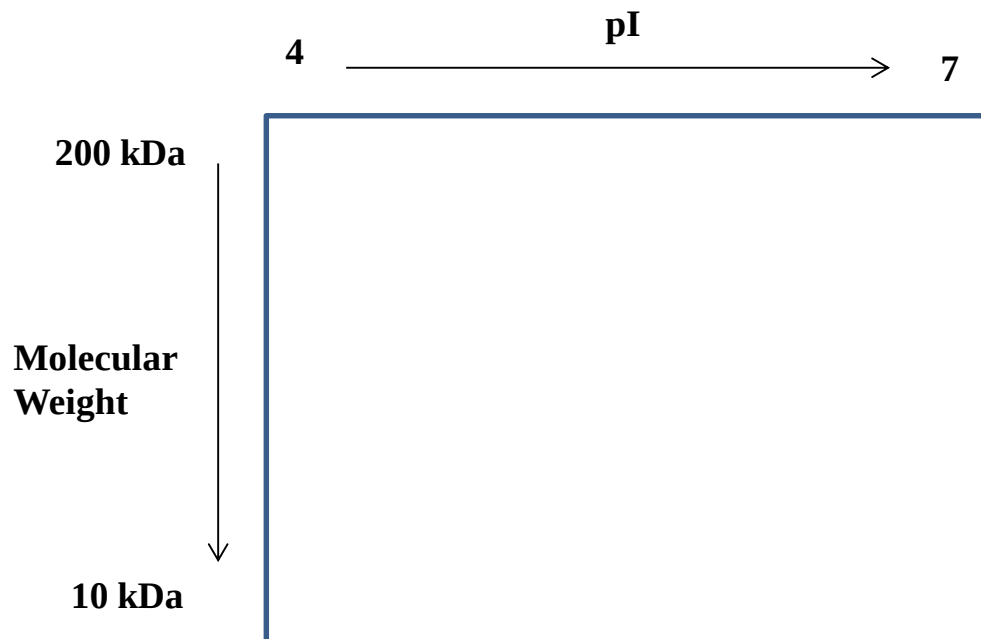
A. Stathmin (STMN1) increased (+1.6). The gel image is different. The spot in the cancerous cell line is darker and larger than the spot in the normal cell line.

Myosin regulatory light chain-2 (ML12A) decreased (-1.9) . The spot in the cancerous cell line is lighter and smaller than the spot in the normal cell line.

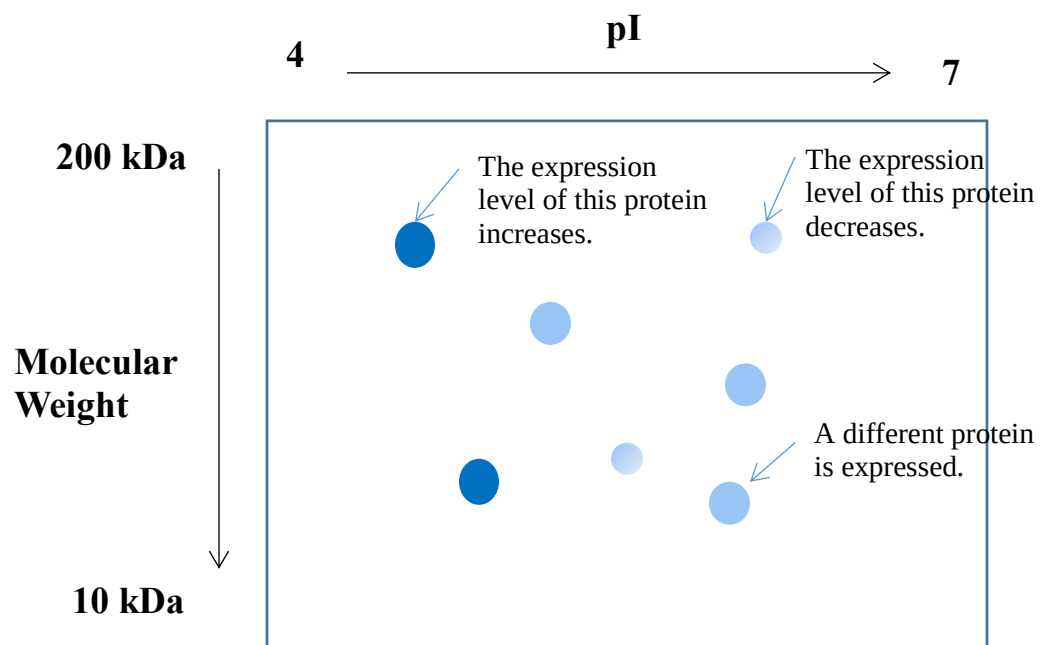
5. The simplified 2D gel shown represents the proteins from a healthy cell line.



Draw a new 2D gel which could represent the changes in protein expression that occur in a cancerous cell line.



A. A representative 2D gel is shown.



Section 2: Calculating the molecular weight of a protein from its electrospray ionization mass spectrum (ESI-MS)

Learning Objectives

At the end of this assignment, students will be able to:

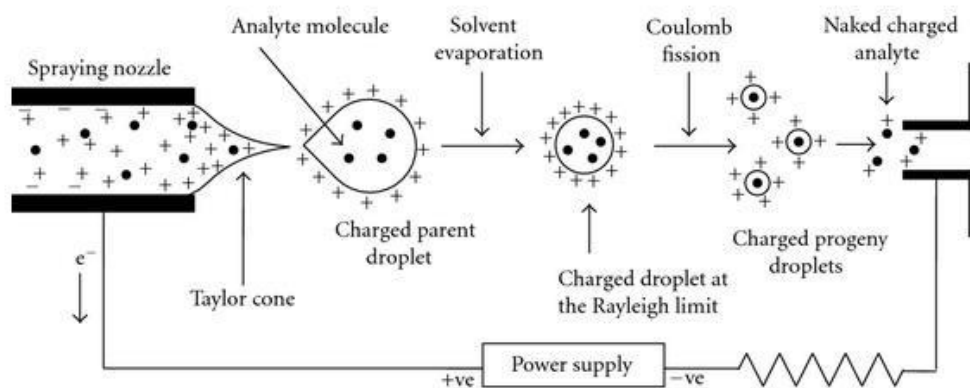
1. Calculate the charge state and mass of a protein from its ESI-MS spectrum.
2. Describe the principle of operation of a quadrupole ion trap mass analyzer.

Section 2A. Electrospray and Quadrupole Ion Trap Instrumentation

Gel electrophoresis gives some indication of the molecular weight of a protein based on its migration in a gel; however, the molecular weights determined from this method are not precise. In contrast, mass spectrometry can provide accurate and precise molecular weights that aid in identification of proteins. Initially, mass spectrometry experiments were limited to small molecules that could be readily volatilized into the gas phase and ionized before entering the mass analyzer. In the late 1980s, several ionization methods were developed and applied to biomolecules that permitted analytes from condensed phase samples, such as liquids and solids, to be volatilized and analyzed by mass spectrometry. Electrospray ionization (ESI) is one such technique.

ESI is a solution-phase “soft” ionization source which converts ions in solution to gaseous ions. The mechanism of ion formation in an ESI source is depicted schematically in Figure 1. Briefly, gaseous ions are formed when a solution containing the analytes of interest (from an LC column or from a syringe infusion pump) is sprayed through a stainless steel capillary to which a high voltage is applied, creating a fine mist of droplets which are charged on their surface. As solvent evaporates from the charged droplets, the charge density on their surface increases to a critical limit, at which point electrostatic repulsion causes the larger droplets to break up into smaller charged droplets. Finally, analyte ions are ejected into the gas phase by electrostatic repulsion, and these ions enter the mass analyzer for subsequent mass analysis. You can watch a [video](#) from the Johnson lab showing the formation of a stable electrospray as the voltage applied to the tip is increased.

Figure 1. The mechanism of electrospray ionization (ESI). High voltage is applied to a steel capillary to produce charged droplets containing the analyte molecules. As solvent evaporates, charge accumulates until the



Rayleigh limit is reached and the droplet undergoes Coulomb fission into smaller droplets. This process continues until gas phase analyte molecules enter the mass spectrometer. Figure is reproduced with permission from S. Banerjee and S. Mazumdar, *Int. J. Anal. Chem.*, **2012**, 2012, 282574 under a Creative Commons Attribution License.

For electrospray, the protein sample is prepared in a solution composed of water with acid and a low surface tension organic solvent, such as methanol. Relatively pure protein samples can be infused directly into the mass

spectrometer using a syringe pump, or more complex protein and peptide mixtures may be separated by chromatography first and electrosprayed directly from the column.

Electrospray ion sources are compatible with many types of mass analyzers. For this application, we will consider the quadrupole ion trap as a mass analyzer. The quadrupole ion trap is composed of two end cap electrodes at the entrance and the exit of the trap and a ring electrode (shaped like a donut) in the middle. The voltages and AC frequencies of the applied potentials on the electrodes are varied to control the motion of ions in and through the trap. The ion trap mass analyzer is a small, relatively inexpensive mass analyzer that typically generates mass spectra with unit resolution ($\Delta m \approx 1$) or slightly better. This [video](#) shows the operating steps involved in generating a mass spectrum using an ion trap.

Video Question

1. Summarize what you saw in the video on the ion trap. What are the operating steps used to generate a mass spectrum using an ion trap mass analyzer?

A. The ions are formed by electrospray and then enter the trap. In the trap, the ions collide with buffer gas particles (gray) and their motion is damped. The ions are then scanned out of the trap sequentially by m/z value and detected.

Section 2B. ESI-MS Data

One of the advantages of ESI is that it is a “soft” ionization technique in which little fragmentation of large, thermally fragile biomolecules occurs. Consequently, molecular weight information is readily obtained with this technique. Additionally, the ions formed are often multiply charged, which enables the analysis of molecular masses exceeding 100,000 Da because the multiple charges bring the m/z values into the mass range of conventional mass analyzers, such as ion traps. With proteins and peptides, the mass spectrum consists of a series of peaks, call the “peak envelope” which represents a distribution of multiply charged analyte ions.

Ubiquitin is a small protein with a monoisotopic molecular weight of 8560 Da. Electrospray ionization of this small protein typically results in major charge states of +8, +9, +10, +11, +12, and +13.

Reading Questions

1. Using this information, complete the table below, assuming that the charges on each ion come from protonation rather than sodium or potassium adducts. Round masses and m/z values to the ones place.

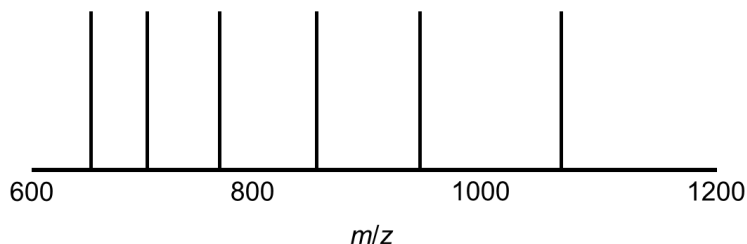
z	mass of $[M+zH]^{+z}$	m/z
8		
9		
10		
11		
12		
13		

A.

z	mass of $[M+zH]^{+z}$	m/z
8	8568	1071
9	8569	952
10	8570	857
11	8571	779
12	8572	714
13	8573	659

2. Using the data you entered in the table above, sketch an expected ESI-MS spectrum for ubiquitin. Label each peak with its charge state. What do you notice about the spacing of the peaks along the x-axis?

A.



The spacing of peaks widens along the length of the x-axis as the charge state decreases, as seen in the predicted spectrum.

Based on the m/z values and peak spacing observed in the charge envelope, we can determine the charge state, z , for each peak and the molecular weight of the analyte using Equations 1 and 2,

$$z = \frac{M_2 - A}{M_1 - M_2} \quad (1)$$

$$MW = \frac{(M_1 - A)(M_2 - A)}{M_1 - M_2} \quad (2)$$

where MW is the molecular weight of the analyte, M_1 is the m/z value for the first ion, z is the charge state of the first ion, M_2 is the m/z value for a second ion of lower m/z , and A is the mass of the adduct ion, which is usually a proton (H^+) but can be sodium (Na^+) or potassium (K^+) ions from glassware or buffers used in the experiment.

Discussion Questions

1. Why do aqueous samples for electrospray typically include an acid and a low surface tension solvent such as methanol?

A. *The acidity of the solution helps to ionize analytes by protonation, and low surface tension solvent promotes droplet fission.*

2. Figure 2 shows an experimentally obtained mass spectrum for ubiquitin. Compare this spectrum to the spectrum you predicted in Reading Question 2. Are there any differences? If so, what might cause these differences?

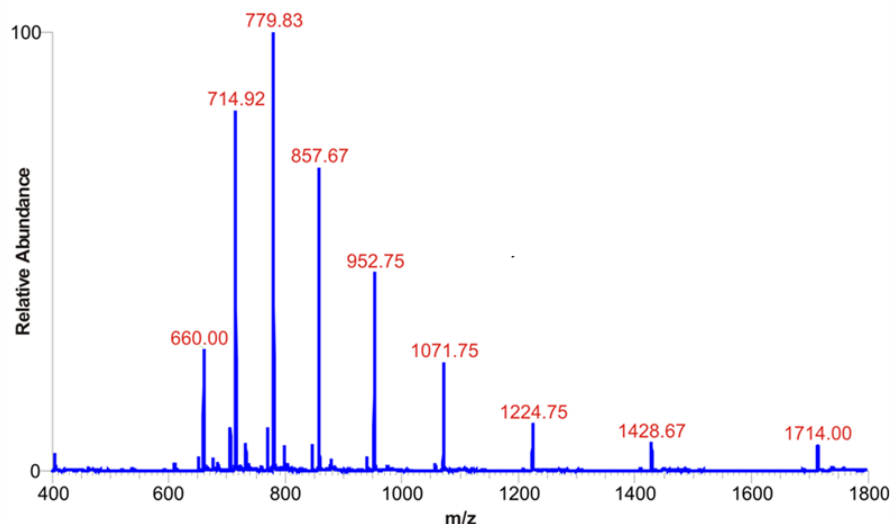


Figure 2. ESI-MS spectrum of bovine ubiquitin from Protea Biosciences (<https://proteabio.com/products/PS-143>)

A. The experimental spectrum agrees with the predicted spectrum with respect to the m/z values of the major peaks, which differ only slightly from the predicted values. The spectra differ in two ways: first, the intensity of the peaks varies because some charge states are more favorable than others; second, the experimental spectrum has additional smaller peaks at higher m/z values arising from low abundances of ions with charges of less than +8.

3. Using Equation (2) and any pair of peaks from Figure 2, calculate the molecular weight of ubiquitin and its percent error compared to the theoretical monoisotopic mass of 8560 amu.

A. Any pair of peaks can be used. An example solution is shown below using the peaks at 714.92 and 779.83.

$$MW = \frac{(M_1 - A)(M_2 - A)}{M_1 - M_2} = \frac{(779.83 - 1.01)(714.92 - 1.01)}{779.83 - 714.92} = 8566 \text{ amu}$$

$$\% \text{ error} = \frac{8566 - 8560}{8560} \times 100\% = 0.07\%$$

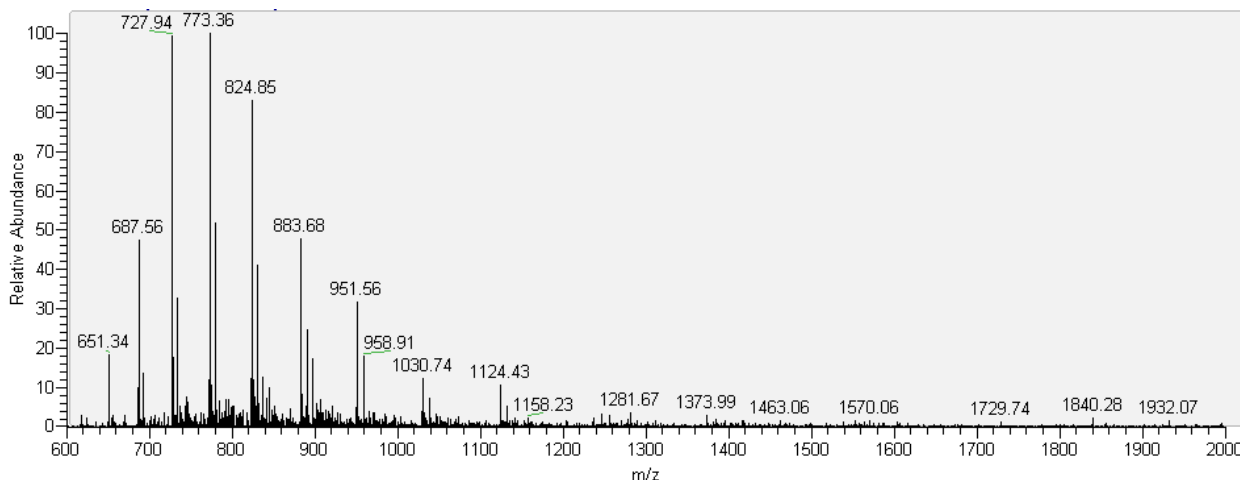


Figure 3. ESI-MS spectrum of cytochrome C. Data obtained at Trinity College.

4. Figure 3 shows the ESI-MS spectrum for cytochrome C electrosprayed from a mixture of water, methanol, and acetic acid with pH of 2.5. What is the charge state of the peak at $m/z = 773.36$?

A. Using equation (1), $z = \frac{727.94 - 1.008}{773.36 - 727.94} = \frac{726.932}{45.42} = +16$

If students use the peak at $m/z = 824.85$ the charge comes out negative. It is helpful to remind them that the sample is usually prepared in acid and that they should expect positive charges from protonation.

5. Determine the MW of the analyte, cytochrome C, using the data in Figure 3.

A. Any pair of peaks can be used. An example solution is shown below using the peaks at 773.36 and 727.94 and equation (2).

$$MW = \frac{(M_1 - A)(M_2 - A)}{M_1 - M_2} = \frac{(773.36 - 1.01)(727.94 - 1.01)}{773.36 - 727.94} = 12360 \text{ amu}$$

6. How would you expect the mass spectrum to change if the cytochrome C sample was electrosprayed from a solution of higher pH? Explain your answer.

*A. Using a higher pH sample results in less protonation of the cytochrome C analyte and therefore lower charge states. This shifts the charge envelope to higher m/z values. For example, at pH 4.2 the z=+9 charge state is the base peak, and at pH 6.4 the +8 charge state is the base peak. See L. Konermann and D. J. Douglas, Biochem., **1997**, 36, 12296-12302 for example data.*

7. Note that in the ESI-MS spectra for ubiquitin and cytochrome C, each major peak is accompanied by a series of less intense peaks of slightly different m/z. What is the source of these peaks?

A. These peaks represent the isotopic distribution of atoms in the protein. Because the m/z value of a mass spectrum represents the mass-to-charge ratio of an individual molecule, they reflect the exact isotopic composition of each molecule. The most intense peak arises from the most likely composition of isotopes. Less intense peaks arise from molecules that happen to contain a different isotopic distribution, with the relative intensities of these peaks representing the isotopic abundance of each element in the ion and the number of atoms.

Note to instructors: For small molecules, the most intense peak arises from ions composed solely of the most abundant isotopes of each atom, such as carbon-12, oxygen-16, hydrogen-1, and nitrogen-13. For organic molecules composed of these elements, the most abundant isotope is also the lightest isotope, and the most intense peak gives the monoisotopic mass of the ion. For larger molecules, the most intense peak may include heavier isotopes, and for very large molecules the most intense peak may actually present the average mass as calculated from atomic weights for each atom. For example, 1.10% of all carbon atoms are carbon-13, so molecules that contain more than 100 carbon atoms are likely to include at least one carbon-13 atom, making their most intense peak heavier than the monoisotopic mass. More information on isotopic distributions in large molecules, such as proteins, can be found in the 1983 Yergey et al. reference.

8. Compare your description of the ion trap mass analyzer video with your group mates' descriptions. As a group, write a consensus explanation for how an ion trap works.

A. The video shows that the trap starts out with some gas molecules already in the trap. (This buffer gas is usually helium. Collisions with the buffer gas "cool" the ions as they enter the trap. This reduction in kinetic energy makes them easier to trap.) Ions from the ESI source enter the trap and follow a complex path between the end cap electrodes and the ring electrode. Eventually, the ions are damped to the middle of the trap. Then ions are scanned out of the trap by m/z value and detected to generate a mass spectrum.

References and Resources

G.L. Glish and R.W. Vachet, "The basics of mass spectrometry in the twenty-first century," *Nat. Rev. Drug Discov.*, **2003**, 2, 140-150.

C. Hendrickson, "Electrospray Ionization (ESI)", <https://nationalmaglab.org/user-facilities/icr/techniques/esi>

C.S. Ho, *et al.* "Electrospray ionization mass spectrometry: principles and clinical applications," *Clin. Biochem. Rev.* **2003**, 24, 3-12.

J.R. Yates, "A century of mass spectrometry: from atoms to proteomes," *Nat. Methods*, **2011**, 8, 633-637.

J. Yergey, *et al.* "Isotopic distributions in mass spectra of large molecules," *Anal. Chem.* **1983**, 55, 353-358.

Section 3: Peptide Mass Mapping for Protein Identification

Learning Objectives

At the end of this assignment, students will be able to:

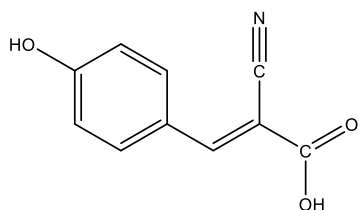
1. Describe the technique of Matrix Assisted Laser Desorption Ionization (MALDI) for creating gas phase ions of peptides and proteins.
2. Describe the principle of operation of Time of Flight (TOF) mass spectrometer and how the reflectron design affects resolution.
3. Describe the process of peptide mass mapping for identifying a protein from mass spectrometry data.
4. Search mass spectrometry data in the MASCOT database and interpret the results.

Overview

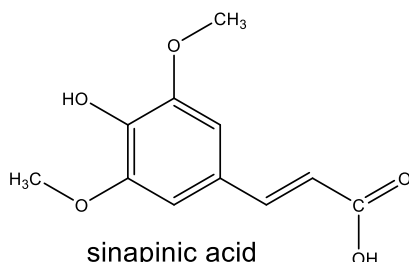
Often the first step in analyzing a complex mixture of proteins is to separate them using two-dimensional (2D) gel electrophoresis. (The Introduction section of this module includes a paper on protein analysis using 2D gel electrophoresis.) If 2D gel analysis indicates that new protein is being expressed or the expression level of a protein has changed, that protein must be identified. **Peptide mass mapping** is a method used to determine the identity of a protein spot in a gel. The protein is cut from the gel, destained, and extracted. Then the protein is digested (cut into pieces) with a proteolytic enzyme such as trypsin. The digest contains a mixture of peptides whose mass is analyzed using Matrix-Assisted Laser Desorption Ionization Time of Flight (MALDI-TOF) mass spectrometry. In this module you will learn the principles of MALDI-TOF mass spectrometry and the method of peptide mass mapping for identifying a protein.

Section 3A. MALDI-TOF Mass Spectrometry

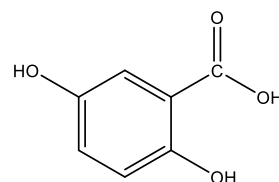
Matrix-assisted laser desorption ionization (**MALDI**) is a technique that generates gas-phase peptide and protein ions. Proteins are nonvolatile and cannot simply be heated to make the gaseous form. (Think about a steak on the grill. The steak chars but does not vaporize.) Instead proteins are **desorbed** (released from a surface) by mixing the protein with a matrix molecule and irradiating the mixture with a pulse from an ultraviolet laser. The following figure shows the structure of various molecules used as the MALDI matrix.



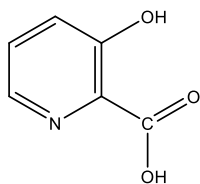
a-cyano-4-hydroxycinnamic acid
CCA $C_{10}H_7NO_3$
peptides and small molecules



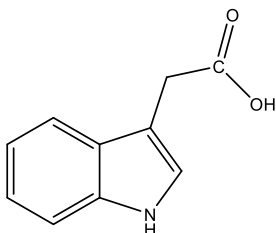
sinapinic acid
SA $C_{11}H_{12}O_5$
proteins



2,5 dihydroxybenzoic acid
DHB $C_7H_6O_4$
oligosaccharides



3-hydroxypicolinic acid
HPA $C_6H_5NO_3$
oligonucleotides



3-indoleacrylic acid
IAA $C_{10}H_9NO_2$
organic polymers

Discussion Question

1. What structural features do these molecules have in common? Why are these common features important for a MALDI matrix?

A. They have aromatic rings so they have strong absorption in the ultraviolet region of the spectrum. They have a carboxylic acid group so they can transfer a proton to the biomolecule giving it a positive charge.

The α -cyano-4-hydroxycinnamic acid (CCA) molecule is a MALDI matrix used for peptide and protein analysis. The matrix is combined with a mixture of peptides and placed on a metal plate. A pulsed nitrogen laser with a wavelength of 337 nm is used to irradiate the peptides and matrix. The absorption spectrum for a peptide and the CCA matrix is shown.

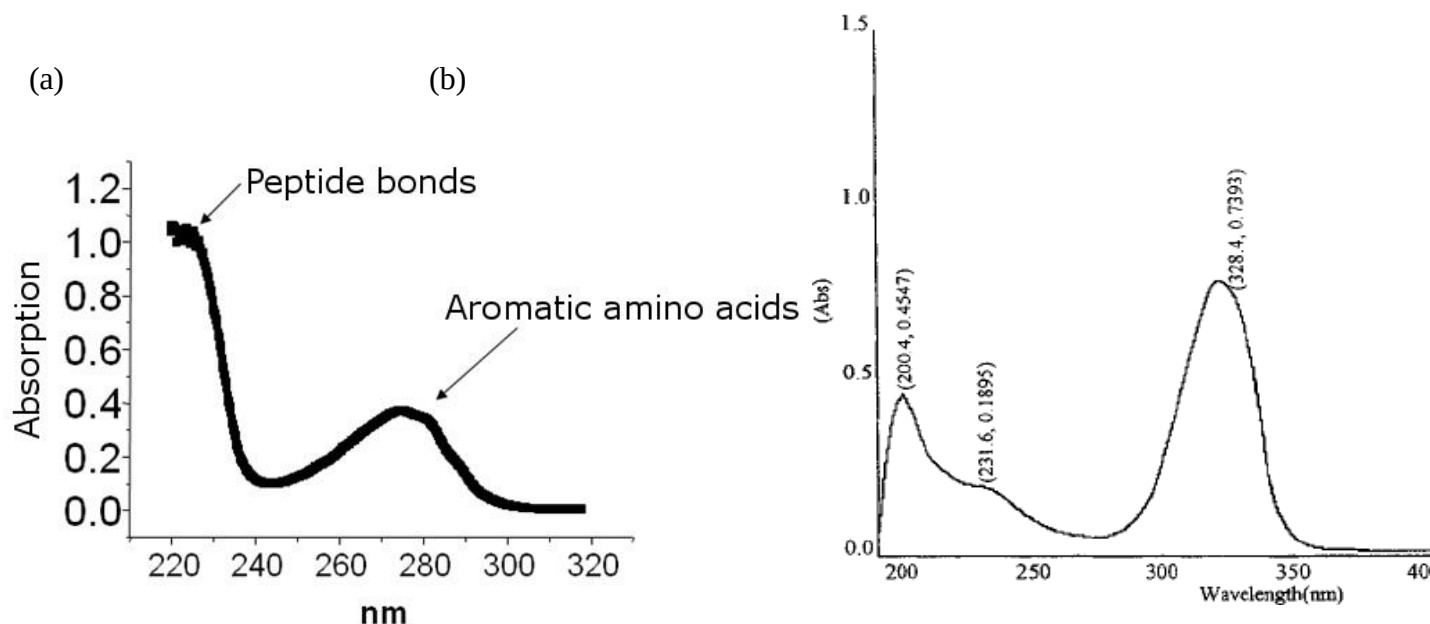


Figure. (a) Absorbance spectrum of a peptide. (b) Absorbance spectrum of CCA matrix.

Discussion Questions

1. Does the peptide absorb the laser light? Does the matrix absorb the laser light?

A. No, the peptide does not absorb the laser light. The peptide absorbs strongly at 220 nm and 280 nm. The matrix does absorb the laser light.

2. Propose a reason that the laser is matched to the absorption of the matrix (CCA) and not the peptide. What would happen to the peptide if it absorbed a high energy pulse of laser light?

A. It would break bonds in the peptide causing fragmentation.

In the MALDI process, the laser is directed at the mixture of CCA matrix and peptides. The matrix absorbs the laser light, but the peptides do not. The matrix heats up and “explodes” from the surface carrying the peptide with it. (It is similar to skydiving. The airplane (matrix) carries the people (peptides) into the air and when they jump, they are temporarily flying.) After absorbing light the matrix is in an excited state and transfers a proton (H^+) to the peptide giving it a positive charge. The invention of MALDI enabled the mass spectrometric analysis of large nonvolatile biomolecules and was so important that its inventor Koichi Tanaka of Shimadzu Corporation was awarded one-third of the 2002 Nobel Prize in Chemistry. The MALDI process is depicted in the figure below.

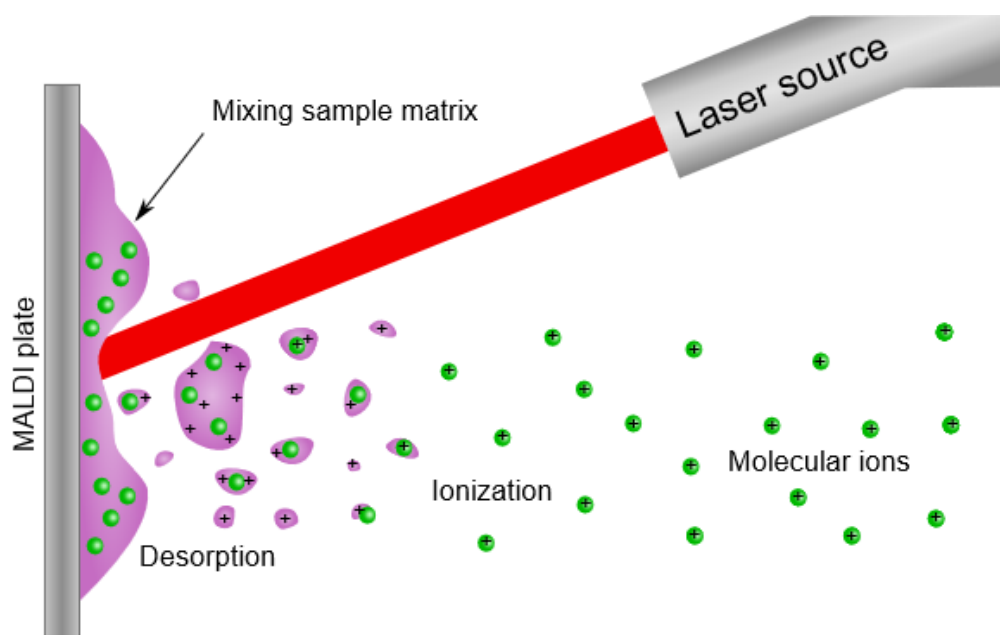


Figure. The MALDI process for creating gas-phase ions of peptides. (Figure from Wikimedia Commons)

The purple color represents the MALDI matrix and the green color represents peptides. Energy from the laser is absorbed by the matrix, causing the matrix to desorb from the surface. The matrix carries the peptides into the gas phase and gives them a charge by donating a proton. Molecular ions of the peptides are created because peptides do not absorb laser light which would break bonds within the molecule.

An example of a typical mass spectrum of a mixture of proteins ionized by the MALDI method is shown below. The matrix is 2,5-dihydroxybenzoic acid (DHB) and the mixture contains three proteins; ubiquitin, cytochrome C and equine myoglobin.

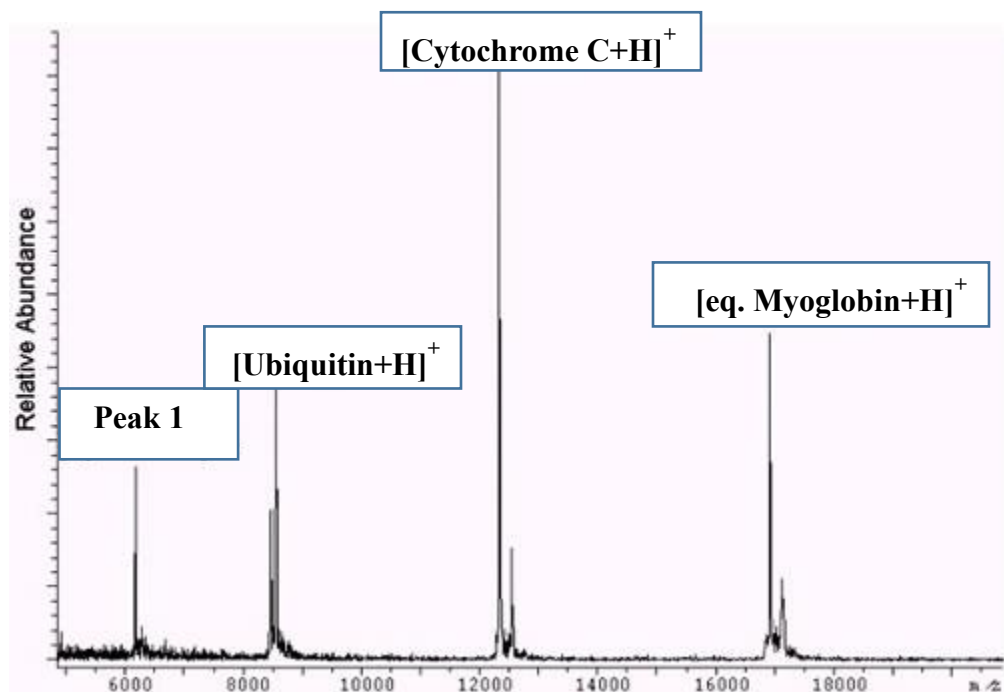


Figure. MALDI-TOF mass spectrum of three proteins; ubiquitin, cytochrome C, and equine myoglobin. Figure adapted from King's College London

<http://www.kcl.ac.uk/innovation/research/corefacilities/smallrf/mspec/cemsw/instr/maldi-tof-ms.aspx>

Discussion Questions

1. Examine the mass spectrum of the mixture of three proteins obtained using MALDI as the ionization method. Would you classify MALDI as a “soft” (little to no fragmentation) ionization technique or a “hard” (fragmentation) ionization technique?

A. Soft ionization method. Fragmentation is not evident in this mass spectrum.

2. What is the identity of Peak 1 in the mass spectrum?

A. Cytochrome C + 2 H with a +2 charge. The m/z ratio of cytochrome C + H with a +1 charge is 12400. If cytochrome C accepts 2 protons from the matrix, then the m/z ratio is 6200.

Section 3B. Time of Flight (TOF) Mass Analyzer

This is set of questions guides students through the principle of a TOF mass analyzer. There is very little reading and all questions are for small group discussion.

(Source for figures and animations: Dr. Jon Karty of Indiana University Mass Spectrometry Facility)



Consider the diagram above, which shows 4 ions (the circles), between two electrical plates. The ions have different masses and higher mass is represented by a larger circle. Imagine the entire schematic is contained in a long tube under vacuum. After the ions enter, the left plate is brought to a positive high voltage (+HV) such as +15,000 V. The right plate is a grid, like a screen door, so that ions can pass through, and is held at electrical ground (0 V).

1. In this scenario, the ions will accelerate once the +HV is applied. If the ions are positively charged, which direction will they move?

A. They will move away from the plate because like charges repel.

2. If all of the ions have a charge of +1, which ion will reach the detector first? The larger circles represent ions with higher mass.

A. The smallest ion (red) with the lowest mass will move fastest and reach the detector first.

The powerpoint file, TOF Animations Slide 2 has a nice animation showing the answer to this question.

3. What factors other than mass will influence the velocity of the ions?

A. The magnitude of the positive voltage and the magnitude of the charge on the ion. Also, the distance the ion travels through the electric field.

Note: Most students will not know about the effect of distance through the electric field. This phenomena limits resolution and will be developed later on this worksheet.

4. This time of flight mass analyzer uses an applied voltage (+HV) to give kinetic energy (KE) to the ions. The amount of energy imparted to the ions is given by Equation 1:

$$KE = zV \quad (\text{Eq. 1})$$

where z is the charge on the ion and V is the magnitude of the high voltage (HV).

- a. Does this equation support your answer to number 3?

A. Yes (*The charge on the ions and the applied voltage influence kinetic energy.*)

- b. If all of the ions have the same charge, what will be true about their kinetic energies?

A. *They will all have the same kinetic energy (KE) because they are all accelerated with the same voltage.*

5. You may remember another equation for kinetic energy from physics:

$$KE = \frac{1}{2}mv^2 \quad (\text{Eq. 2})$$

where m is the mass of the object and v is its velocity.

- a. Consider the four ions at the start of this worksheet. Which ion will have the highest velocity?

A. *If ions with the same charge have the same kinetic energy, then lighter ions will have a higher velocity. The red ion has the smallest mass and will have the highest velocity.*

- b. What is the order of ions reaching the detector?

A. *The order of ions reaching the detector is lowest to highest mass. Red, blue, black, yellow.*

- c. Is the TOF mass analyzer dispersive or scanning?

A. *Dispersive because changes do not need to be made to electric or magnetic fields to pass an ion with a given m/z to the detector. The flight times are very short (on the order of μs and therefore ions are essentially detected all at once allowing for collection of many signals in a short amount of time. The laser can pulse 200 times per second allowing 200 signals to be averaged in 1 second greatly improving signal to noise ratio. Note to instructors: This questions could be omitted if you have not previously discussed dispersive vs. scanning mass analyzers in the course.*

- d. Why do you think it is called a “time-of-flight” mass analyzer?

A. *Ions of different m/z are separated because they take different amounts of time to reach the detector.*

- e. Consider a +1 ion with $m/z = 115$ which enters the time-of-flight source region (between the two plates).

If +HV = 2000 V, how much kinetic energy is imparted to the ion in Joules?

$$1 \text{ elementary charge} = 1.6 \times 10^{-19} \text{ C} \quad 1 \text{ V} = 1 \text{ J/C}$$

A. $KE = zV = (+1)(1.6 \times 10^{-19} \text{ C})(2000 \text{ V}) = 3.2 \times 10^{-16} \text{ C} \cdot \text{V} = 3.2 \times 10^{-16} \text{ J}$

f. What is the velocity of the ion in m/s?

$$1 \text{ J} = 1 \frac{\text{kg} \cdot \text{m}^2}{\text{s}^2} \quad 1 \text{ amu} = 1.66 \times 10^{-27} \text{ kg}$$

Answer.

$$KE = \left(\frac{1}{2} \right) mv^2$$

$$3.2 \times 10^{-16} \frac{\text{kg} \cdot \text{m}^2}{\text{s}^2} = \left(\frac{1}{2} \right) (115 \text{ amu}) \left(\frac{1.66 \times 10^{-27} \text{ kg}}{\text{amu}} \right) v^2$$

$$v = 5.8 \times 10^4 \text{ m/s}$$

g. If the flight tube is 2 m long, how long will the ion take to reach the detector? Remember that $v = d/t$, where v is velocity, d is distance traveled, and t is time.

A. $3.5 \times 10^{-5} \text{ s}$ or $35 \mu\text{s}$

6. The ability of a mass spectrometer to distinguish between two different m/z ions is called resolving power.

A common definition for resolving power is the m/z of the peak (m) divided by the width of the peak at half maximum height (Δm).

$$\text{Resolving Power} = \frac{m}{\Delta m}$$

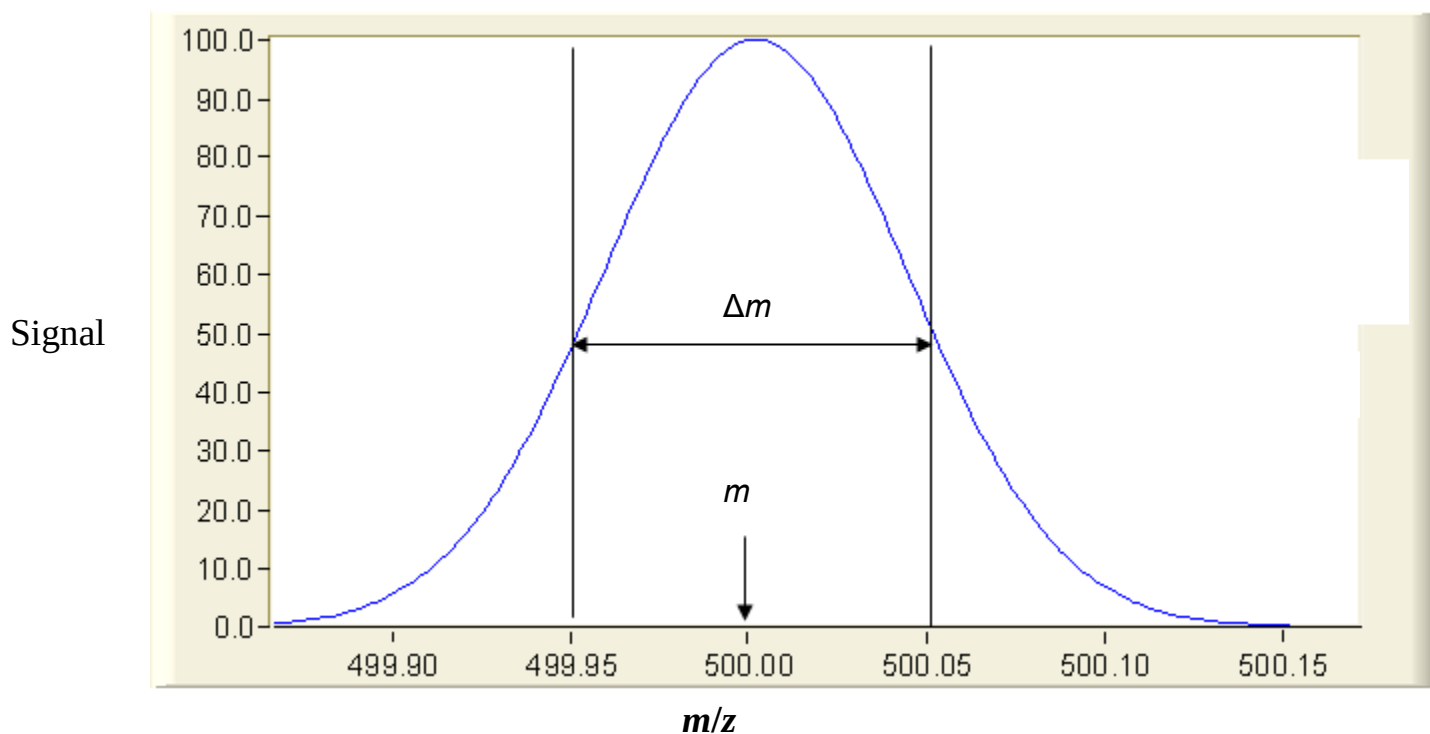


Image adapted from UC Davis Fiehn Metabolomics Lab.

http://fiehnlab.ucdavis.edu/projects/Seven_Golden_Rules/Mass_Resolution

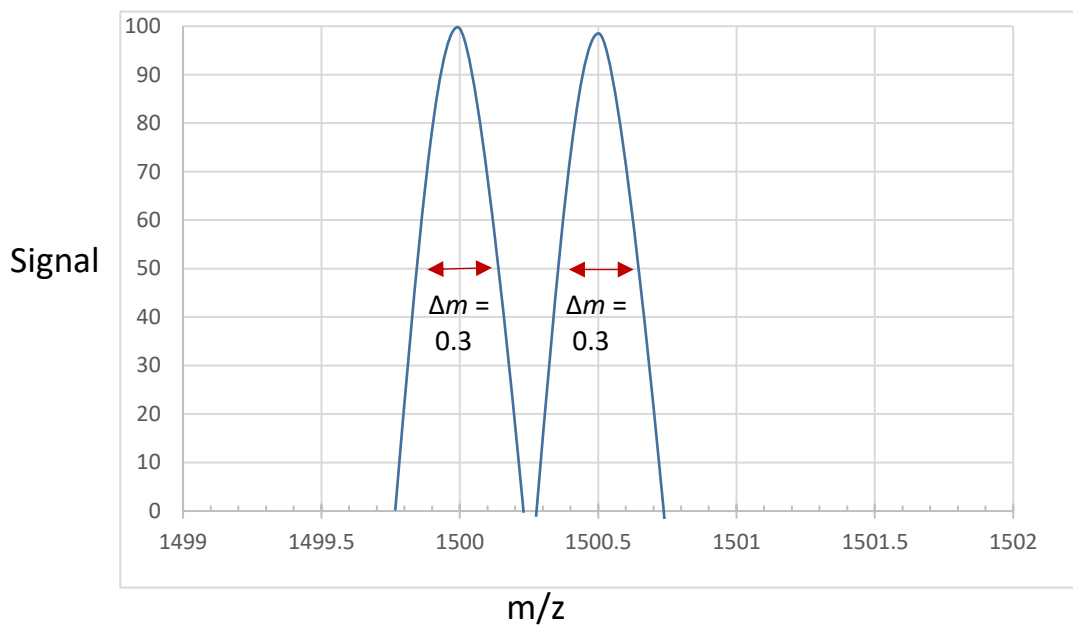
- a. Calculate the resolving power of the mass analyzer from the peak in the figure.

A.

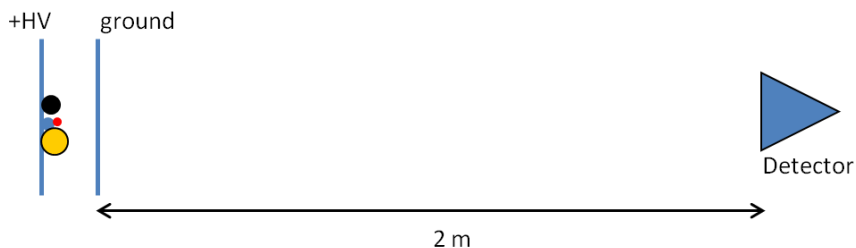
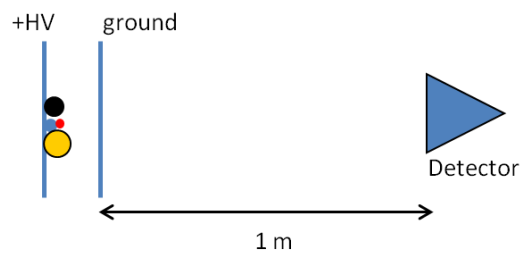
$$\text{Resolving Power} = \frac{m}{\Delta m} = \frac{500.00}{0.10} = 5000 = 5.0 \times 10^3$$

- b. Using the resolving power calculated in part (a), sketch the appearance of the two peaks with m/z ratios of 1500 and 1500.5. Would you say these peaks are resolved?

A. The figure shows the appearance of the two peaks. They are resolved because they do not overlap at the baseline.

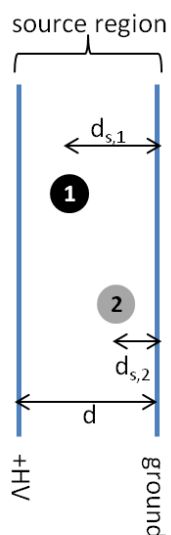


7. Consider the two instrument schematics for a time of flight mass analyzer below. Which one do you think will be better at resolving small differences in the masses of the ions? Explain.



A. The longer flight tube will result in better resolution because the ions with different speeds will have more time to separate. Imagine two runners. One runner has a pace of 7:00 per mile, the other runner has a pace of 8:00 per mile. After 100 yards there will not be much separation between the two, but after a mile the separation will be much larger.

8. Consider the possibility diagrammed below, in which two ions of the same mass and charge start at slightly different positions in the source region.



a. Which ion will leave the source region with the higher velocity, Ion 1 or Ion 2?

A. Ion 1

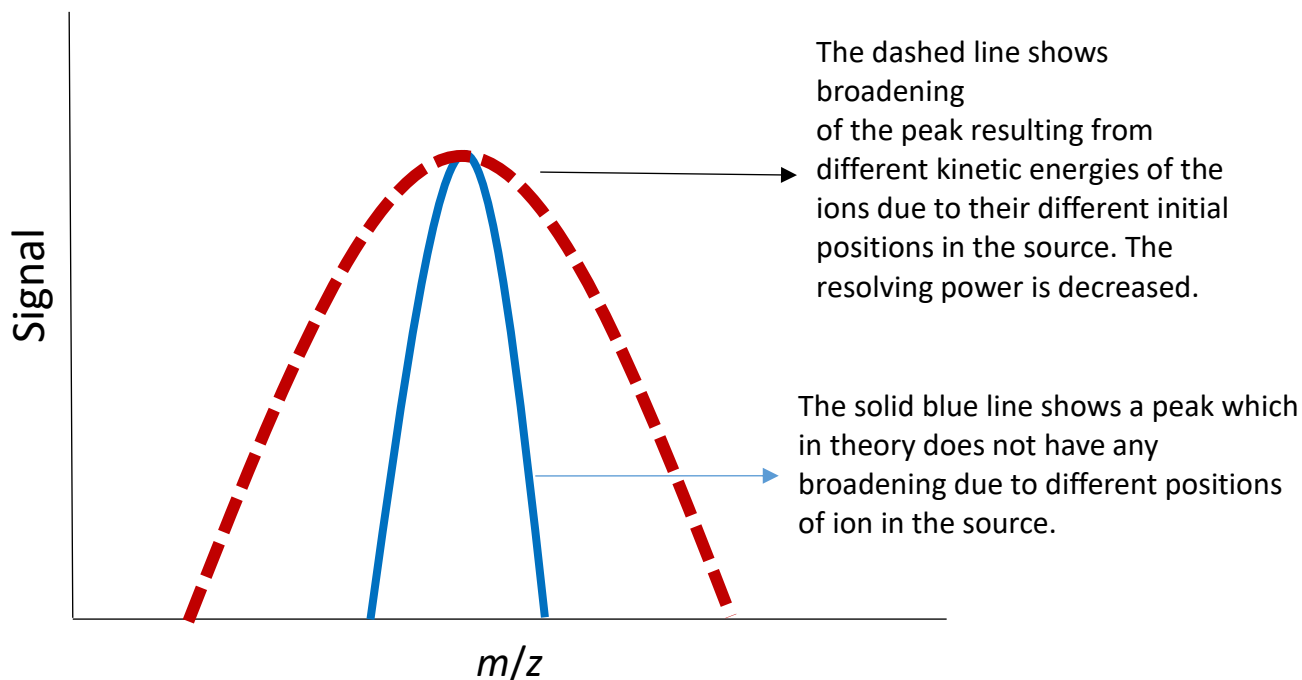
Equation 1 can also be written as : $KE = zEd_s$ (Eq. 1b)

where E is the electric field (V/d) and d_s is the distance traveled in the source region. Use this equation to explain which ion (1 or 2) will leave the source with the highest velocity.

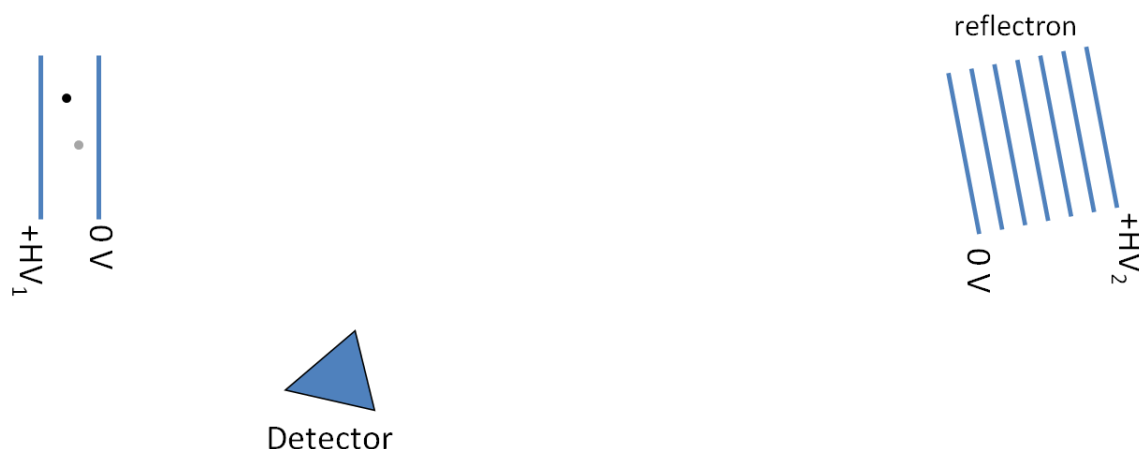
A. Ion 1 will be accelerated a larger distance through the electric field thus acquiring greater kinetic energy and velocity. The powerpoint file, TOF Animations, Slide 9 has a nice animation showing how the position of the ion in source affects the flight time.

b. How will the position of different ions in the source affect the resolving power of the mass analyzer? Draw a peak for a given m/z similar to the one in question 7 assuming that the position of the ion in the source did not affect kinetic energy. Draw a new peak taking into account how different positions of ions in the source affect the kinetic energy of the ions.

A. The resolving power will decrease because ions of the same m/z formed at different locations in the source will acquire different kinetic energies and reach the detector at different times. In theory, ions with the same m/z should have the same kinetic energy and reach the detector at the same time. The effect of different positions of ions in the source will broaden the peak.



9. Consider the instrument diagram below, which adds a “reflectron” to the mass analyzer. The reflectron consists of a series of grids held at increasingly high positive potentials from ground (0 V) up to $+HV_2$. Note : $+HV_2 > +HV_1$.



- a. On the diagram above, sketch the flight paths for Ion 1 and Ion 2 (these ions have the same m/z ratio and are the same ions in question 8). Note that the reflectron is angled toward the detector.

A. Ion 1 and 2 turn around at the reflection and head toward the detector.

- b. Which ion penetrates farther into the reflectron?

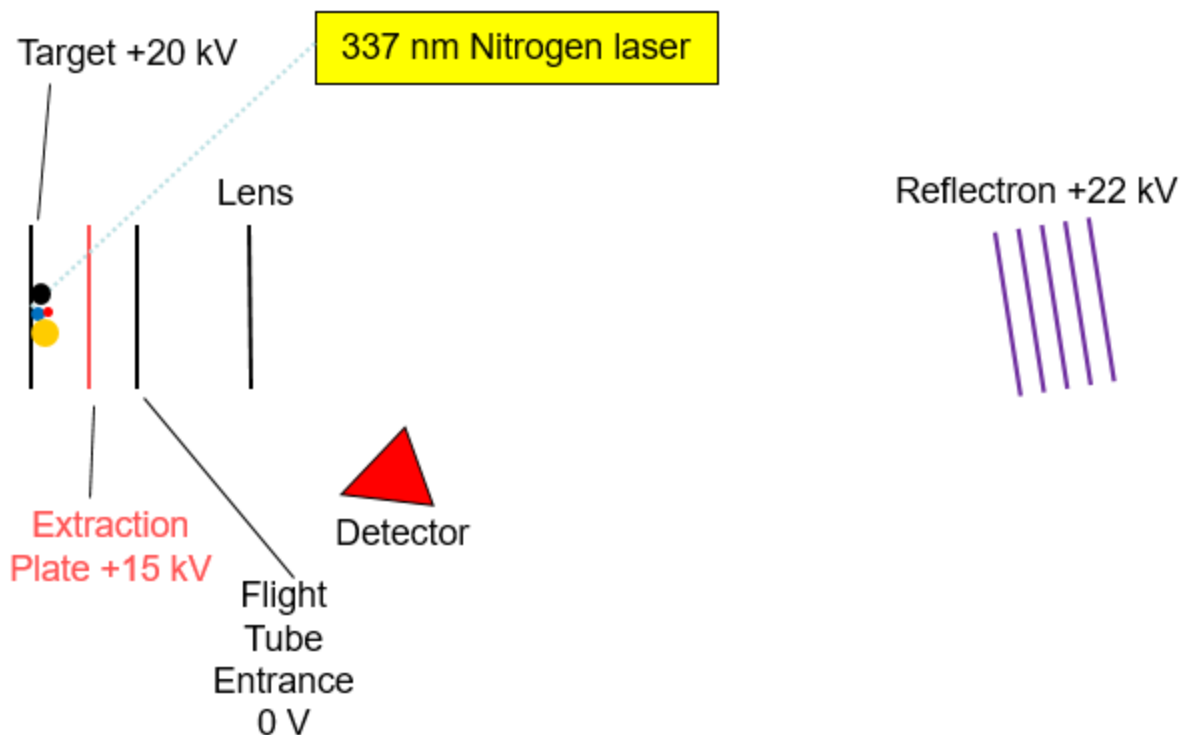
A. Ion 1 will leave the source with greater kinetic energy and will be more difficult to stop and turn around. Therefore Ion 1 will penetrate further into the reflectron.

- c. How does this arrangement correct for the ions' different starting positions in the source?

A. Ion 1 is moving faster but travels a greater distance because it penetrates further into the reflectron. Ion 1 and Ion 2 should reach the detector at nearly the same time.

The Powerpoint file, TOF Animations Slide 10 shows the effect of the reflectron on the flight time of two ions with the same m/z formed at different positions in the source.

10. Examine the diagram of a reflectron TOF mass analyzer.



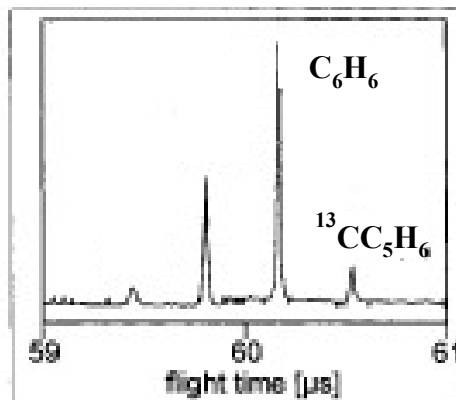
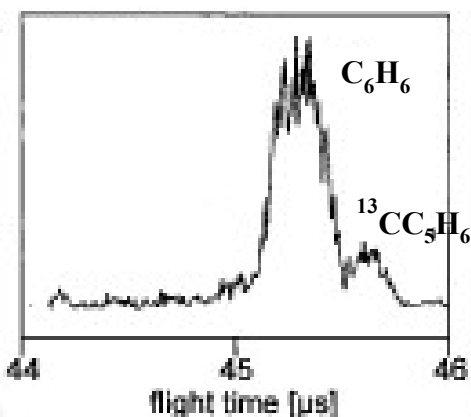
- a. What is the order of ions reaching the detector in the reflectron time of flight mass spectrometer? (all ions have a +1 charge)

A. Lowest mass ion arrives first and highest mass ion arrives last. (red, blue, black, then yellow)

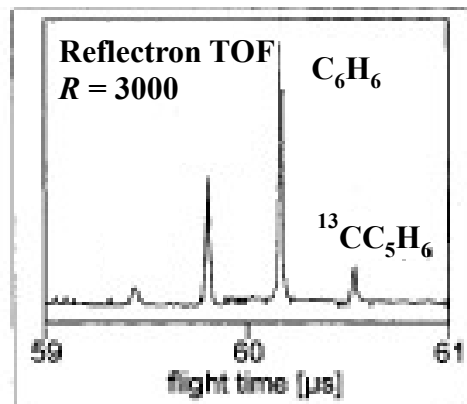
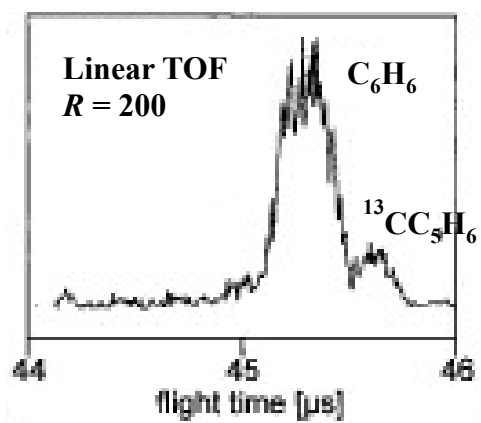
- b. Why is the ionization method MALDI frequently coupled with a time of flight mass analyzer?

A. To measure flight times a set of ions must enter the mass analyzer and be accelerated all at the same time. The pulsed nitrogen laser, produces a “pulse” of ions for time of flight analysis. The nitrogen laser is often pulsed 100-200 times per second leading to improved signal to noise ratio due to signal averaging.

- c. Examine the two mass spectra below. Identify which spectrum is from a reflectron TOF instrument and which is from a linear TOF instrument?



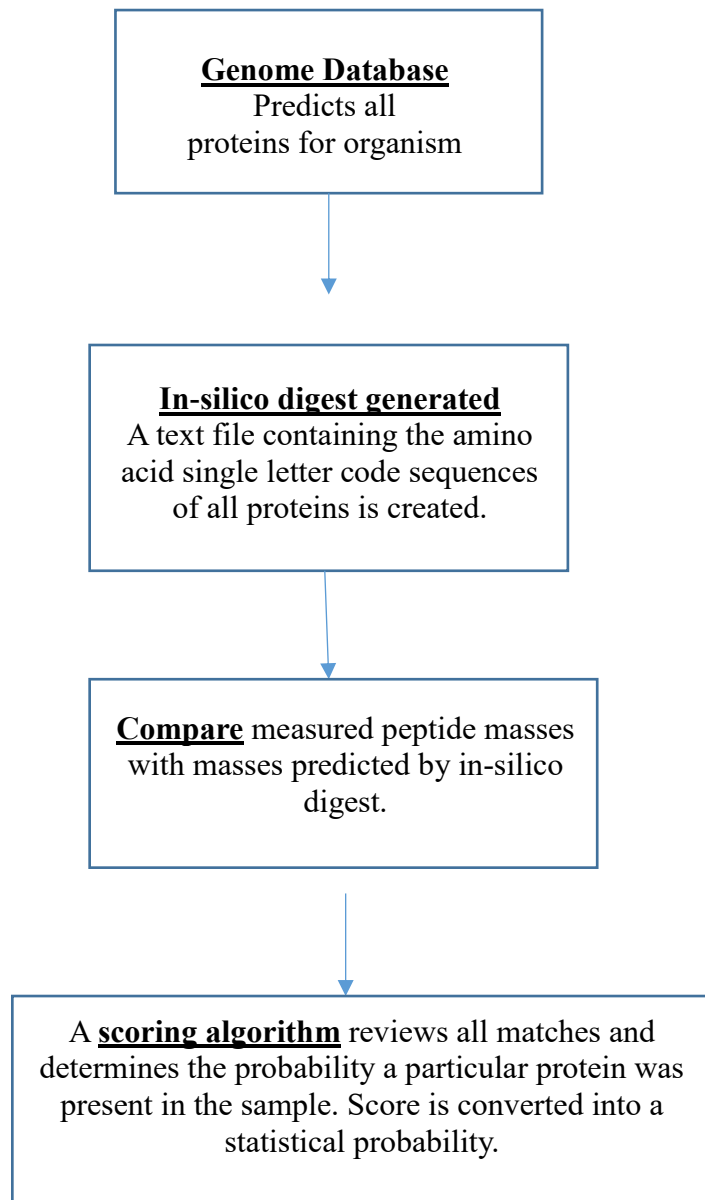
A.



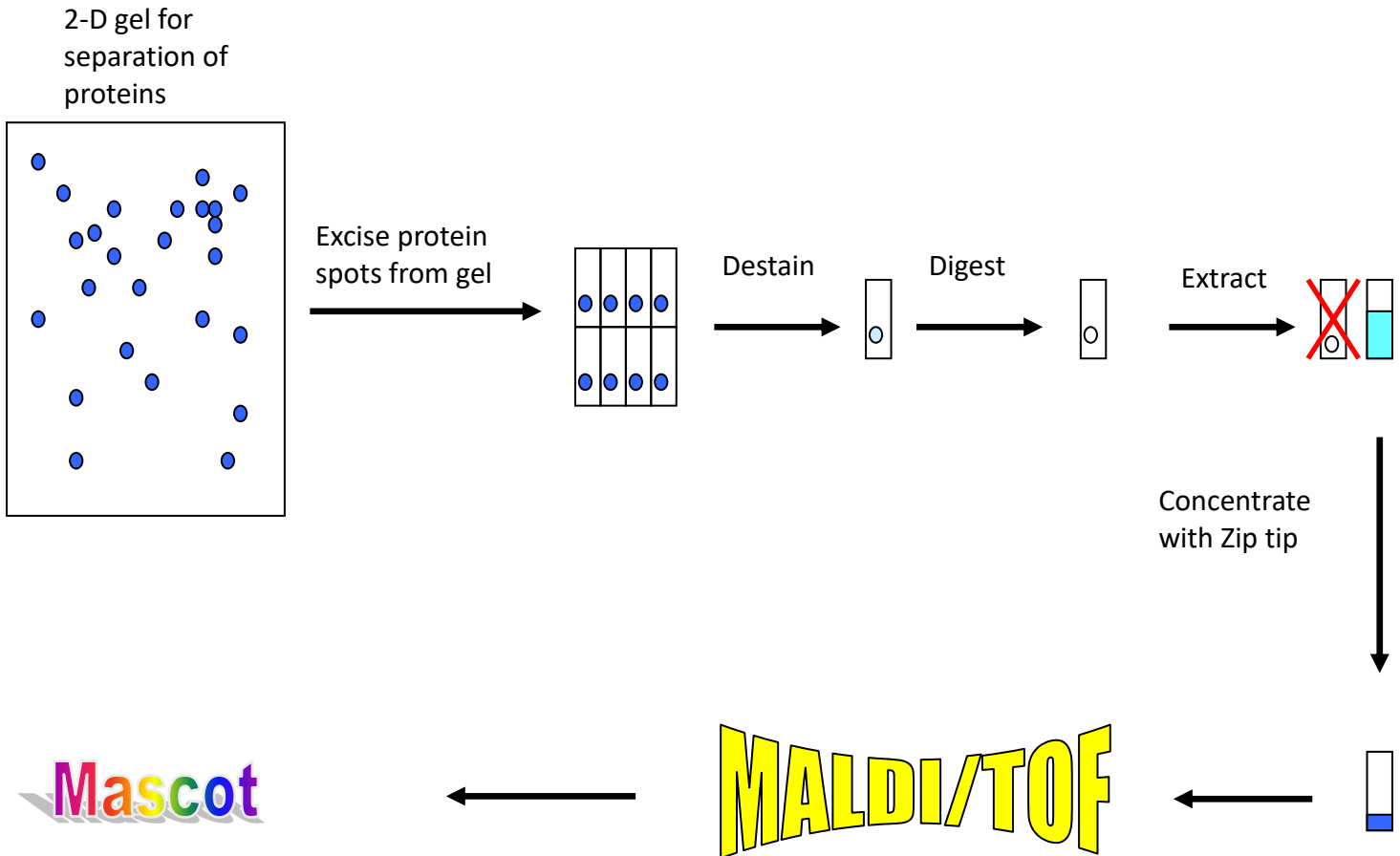
Section 3C. Peptide Mass Mapping for Protein Identification

Peptide mass mapping is a technique that uses powerful search engines (e.g. Mascot) to identify a protein from mass spectrometry data and primary sequence databases. The general approach is to take a small sample of the protein and digest it with a proteolytic enzyme, such as trypsin. Trypsin cleaves the protein after lysine and arginine residues. The resulting mixture of peptides is analyzed by MALDI-TOF mass spectrometry.

The experimental mass values of the peptides are then compared with theoretical peptide mass values. Theoretical mass values of peptides are obtained by using the genome sequence for an organism and predicting all the proteins that can be expressed. Once all the proteins are predicted then the cleavage rules for the digest enzyme are applied and the masses of the resulting peptides calculated by the computer (in-silico digest). By using an appropriate scoring algorithm, the closest match or matches can be identified. If the "unknown" protein is present in the sequence database, then the aim is to pull out that precise entry. If the sequence database does not contain the "unknown" protein, then the aim is to pull out those entries which exhibit the closest homology, often equivalent proteins from related species. The steps in peptide mass mapping are outlined in the flow chart.



The analysis of a complex mixture of proteins from an organism always involves some type of separation step to isolate a certain protein. The separation methods frequently used are two dimensional (2D) gel electrophoresis or liquid chromatography. The figure shows the experimental workflow used to identify a protein spot from a 2D gel.



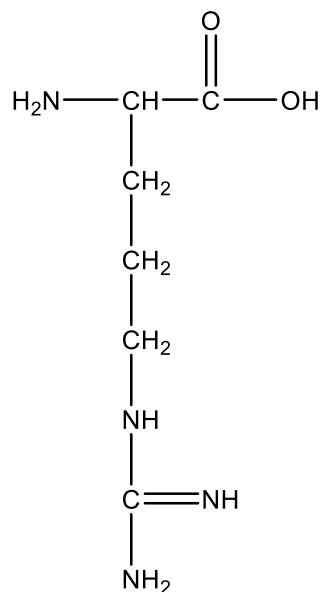
Reading Questions

1. In your own words, describe the general principle of peptide mass mapping for protein identification.

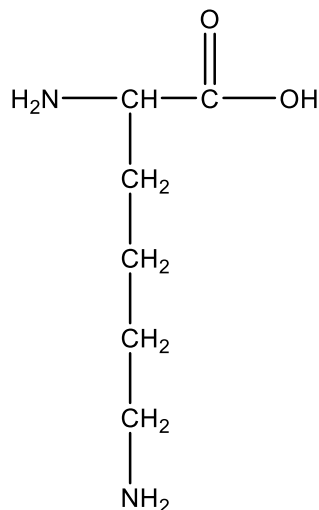
A. The sequence of DNA bases in an organism's genome is used to predict all the proteins the organism will synthesize. Once the proteins are predicted, then the cleavage rules for a specific enzyme such as trypsin are applied and the masses of the different peptides are predicted for each protein. Experimentally, a protein is removed from a gel and digested with an enzyme. The masses of the peptides from the digest are measured using MALDI-TOF mass spectrometry and compared with the theoretical peptide matches. A statistically significant number of peptide mass matches will identify a protein.

The enzyme trypsin is frequently used to digest proteins in the peptide mass mapping technique. Trypsin cleaves the amide bond after lysine (K) and arginine (R) residues. K and R make up about 10% of the amino acids in a protein and digesting with trypsin typically results in peptides in a useful range mass range for mass spectrometry (500-3,000 Da). (Remember: The mass unit amu is the same as Da.)

The structures of lysine and arginine are shown.



Arginine (R)



Lysine (K)

Discussion Questions

1. What is the purpose of digesting the protein with trypsin? (Hint: Think about how mass spectrometry is used to identify and/or elucidate the structure of small organic molecules such as caffeine).

A. The molecular weight of a protein is not a unique characteristic. There are many proteins that could have similar or identical molecular weights and the molecular weight alone cannot be used for identification. Digesting the protein with trypsin cleaves the protein into smaller fragments in a predictable way. The matching of many masses (from different peptide fragments) will be a unique characteristic that can be used to identify a protein. Peptide mass mapping is often referred to as a "peptide fingerprint" because just like a fingerprint analysis there must be several unique similarities for a match to be made.

2. Why is a well annotated genome for the organism of interest needed in the peptide mass mapping technique for identifying a protein?

A. A protein is identified by taking experimentally measured peptide masses and matching them with predicted masses. The computer predicts the peptide masses in the following way. First, the sequence of DNA bases is used to predict all the proteins produced by a given organism. Then, peptide masses for each protein are predicted by cleaving the protein after lysine and arginine. There would not be theoretical masses to match to experimental data without the DNA sequence for the organism.

3. Classify the side chains of R and K as acidic or basic.

A. Basic

4. Digesting the protein with trypsin ensures that there is an R or K residue in each peptide. Why is this helpful for MALDI-TOF analysis? (Hint: Think about the function of the matrix in MALDI).

A. The basic residues readily accept a proton from the matrix giving the peptide a positive charge. The positive charge is needed so the ion can be accelerated with a voltage in the TOF mass analyzer.

5. a. Can the following two peptides with the same amino acid composition be distinguished using MALDI-TOF mass spectrometry?

Peptide 1: GASPVRTCILKMHFY

Peptide 2: GMFHRATIKYPVCSL

A. No they cannot be distinguished because they have the same mass.

b. Calculate the expected (monoisotopic) masses if the enzyme trypsin was used to digest peptides 1 and 2. A table of amino acid masses is provided. Can the peptides be distinguished after digestion with trypsin?

(Refer to the Introduction section of this module if you need assistance in calculating the mass of a peptide or defining the difference between a monoisotopic mass and average mass.)

A.

Peptide 1 has the following masses after digestion with trypsin.

GASPV (585.33 amu)

TCILK (146.11 amu)

MHFY (596.25 amu)

Peptide 2 has the following masses after digestion with trypsin.

GMFHR (646.31 amu)

ATIK (431.28 amu)

YPVCSL (593.29 amu)

Yes, peptide 1 and 2 can be distinguished once they are digested with trypsin. The masses of the three smaller peptides are different.

6. There are other enzymes that could be used to digest the proteins.

Pepsin is most efficient in cleaving peptide bonds between hydrophobic amino acids (leucine) and aromatic amino acids such as phenylalanine, tryptophan, and tyrosine. Pepsin is less specific and results in many small peptides. Would pepsin be a good choice for digesting proteins for peptide mass mapping? Explain your reasoning.

There are also enzymes that are highly specific and result in only a few cleavage sites in a protein. Is a highly specific enzyme that creates a few large peptides be a good choice for peptide mass mapping? Explain your reasoning.

A. An enzyme that results in many small peptides would not be ideal for peptide mass mapping because there would be a significant chance that other proteins would have small peptides with the exact same amino acid composition. The goal is to produce peptides that are unique to one protein.

An enzyme that only has a few cleavage sites and results in a few large peptides is not ideal for peptide mass mapping either. Multiple peptide matches are required for identification. The greater number of matches the higher the probability that the protein has been identified correctly. A few large peptides would only give a few masses to be matched

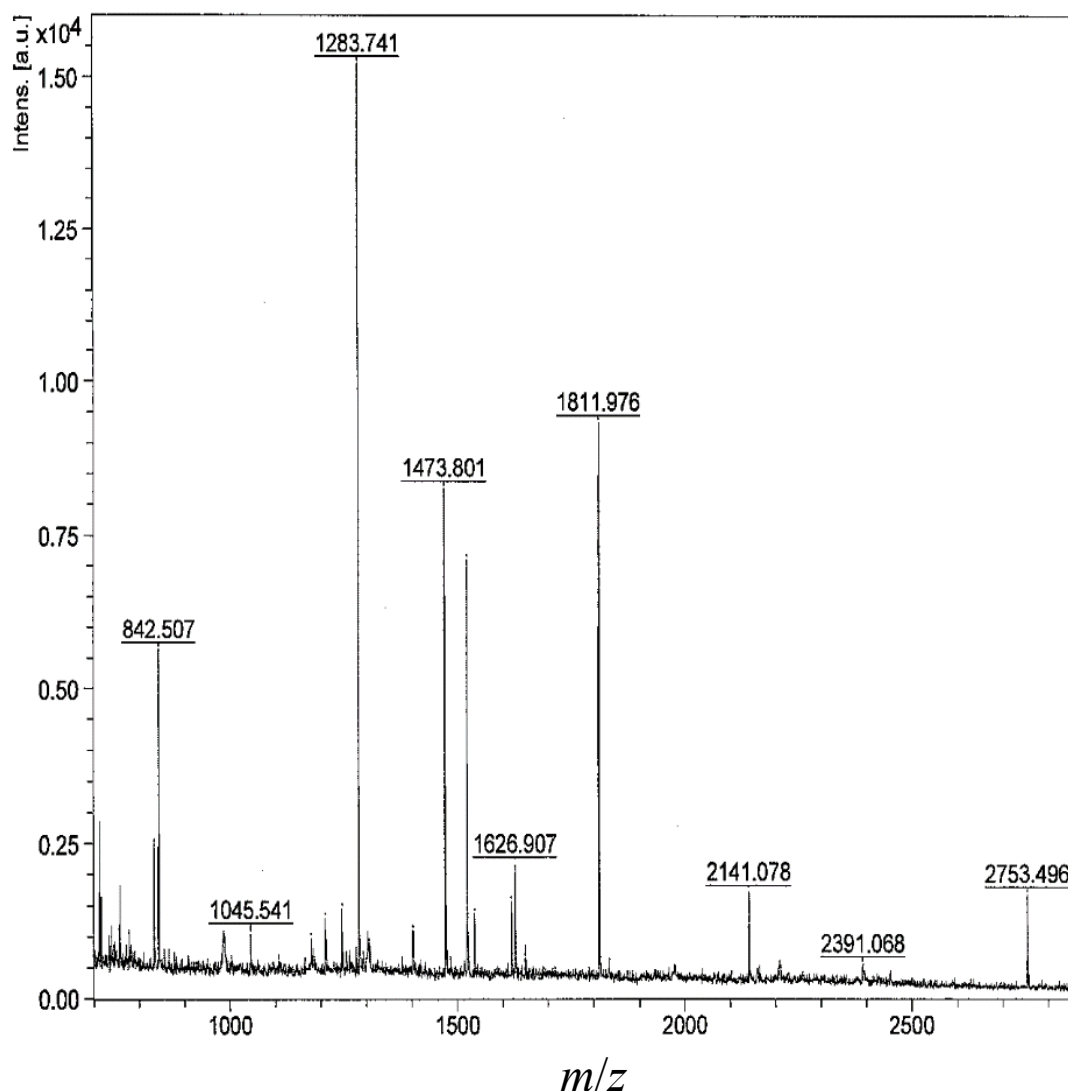
Table 2. Molecular weight information for all twenty naturally occurring amino acids.

Amino Acid	Single-Letter Code	Residue MW (amu)	Amino Acid MW (amu)
glycine	G	57.02	75.03
alanine	A	71.04	89.05
Serine	S	87.03	105.04
proline	P	97.05	115.06
Valine	V	99.07	117.08
threonine	T	101.05	119.06
cysteine	C	103.01	121.02
isoleucine	I	113.08	131.09
leucine	L	113.08	131.09
asparagine	N	114.04	132.05
aspartic acid	D	115.03	133.04
glutamine	Q	128.06	146.07
Lysine	K	128.09	146.11
glutamic acid	E	129.04	147.05
methionine	M	131.04	149.05
histidine	H	137.06	155.07
phenylalanine	F	147.07	165.08
arginine	R	156.10	174.11
tyrosine	Y	163.06	181.07
tryptophan	W	186.08	204.09

Section 3D. MASCOT Database Search

The peptide mass fingerprint search in the MASCOT database is used to identify a protein from mass spectrometry data. The following MALDI-TOF mass spectrum shows the masses detected after digesting a protein with trypsin. This protein is from the organism *Escherichia coli* (*E. coli*) and was cut from a 2D gel and analyzed by students in Analytical Chemistry at Indiana University.

m/z	S/N
832.312	10.3
842.53	23
1045.541	3.7
1179.58	2.3
1210.545	5.2
1247.598	5.7
1283.741	74.7
1307.679	2.1
1401.67	3.1
1403.711	2.3
1473.801	43.2
1521.775	37.2
1537.764	6
1618.801	7.3
1626.907	10.4
1648.88	2.6
1811.976	53
2141.078	8.5
2391.068	2.3
2753.496	12.3



Some peaks in the mass spectrum do not come from peptides from the protein we are trying to identify. Peaks that should be eliminated before performing the Mascot search are:

Trypsin autolysis peaks: Trypsin cleaves the protein of interest after lysine and arginine residues. However, trypsin will also cleave other trypsin molecules. The peaks due to trypsin autolysis are: 514.63, 842.5, 906.05, 1006.15, 1045.12, 1736.97, 1768.99, 2158.48, 2211.4, 2239.1

Matrix Clusters: The MALDI matrix may also combine with Na⁺ and K⁺ during ionization. These peaks are 855.1, 861.1, 871.1, 877.1, 1060.1. If the scientist had good lab technique, these ions would be eliminated or greatly reduced in a sample preparation step.

Reading Question

1. Examine the m/z data in the table and figure on the previous page. Which peaks should be removed before entering m/z data into the MASCOT database?

A. 842.53, 1045.541

Peptide Mass Mapping Search Parameters

Before m/z information is entered into the MASCOT database, there are also a number of **search parameters** that must be set appropriately. The image shows a data entry page for a peptide mass fingerprint search. Let's examine the meaning of each of the following search parameters; Database, Enzyme, Taxonomy, Fixed Modifications, Variable Modifications, Protein Mass, Peptide Tolerance, Mass Values, Monoisotopic or Average Mass, and Data Input.

MASCOT Peptide Mass Fingerprint

Your name	Jill Robinson	Email	jirobins@indiana.edu
Search title	protein 1		
Database(s)	SwissProt NCBI nr contaminants cRAP	Enzyme	Trypsin
		Allow up to	1 missed cleavages
Taxonomy	All entries		
Fixed modifications	Carbamidomethyl (C)	> <	Acetyl (K) Acetyl (N-term) Acetyl (Protein N-term) Amidated (C-term) Amidated (Protein C-term) Ammonia-loss (N-term C) Biotin (K) Biotin (N-term) Carbamyl (K) Carbamyl (N-term) Carboxymethyl (C)
	Display all modifications ☐		
Variable modifications	Oxidation (M)	> <	
Protein mass	kDa	Peptide tol. $\pm$	1 Da
Mass values	☒ MH^+ ☐ M_r ☐ $M-H^-$		
	☐ Data file Choose File No file chosen		
	☒ Query		
Data input	1179.58 1210.545 1247.598 1283.741 1307.679 1401.67		
Decoy	☐		
	Report top	5 hits	
Start Search ...		Reset Form	

The following section on search parameters has been adapted from the Peptide Mass Fingerprint Tutorial in the MASCOT Database.

Database and Taxonomy: The first choice you have to make is which **database** to search. Some databases contain sequences from a single organism. Others contain entries from multiple organisms, but usually include the taxonomy for each entry, so that entries for a specific organism can be selected during a search using a **taxonomy** filter.

If your target organism is well characterized, such as human or mouse or yeast, Swiss-Prot is the recommended choice. The entries are all high quality and well annotated. Because Swiss-Prot is non-redundant, it is relatively small, which makes it easier to get a statistically significant match. If you know what is in the sample, you can restrict the search to an organism or family by means of the taxonomy filter, but remember that you can never rule out contaminants.

If you are interested in a bacterium or a plant, you may find that it is poorly represented in Swiss-Prot, and it would be better to try one of the comprehensive protein databases, which aim to include all known protein sequences. The two best known are NCBIInr and UniRef100. These are very large databases, and you will almost certainly want to select a limited taxonomy. But, never choose a narrow taxonomy without looking at the counts of entries and understanding the classification. In the current Swiss-Prot, for example, there are 26,139 entries for rodentia, of which all but 1,602 are for mouse and rat. So, even if your target organism is hamster, it isn't a good idea to choose 'other rodentia'. Better to search rodentia and hope to get a match to a homologous protein from mouse and rat.

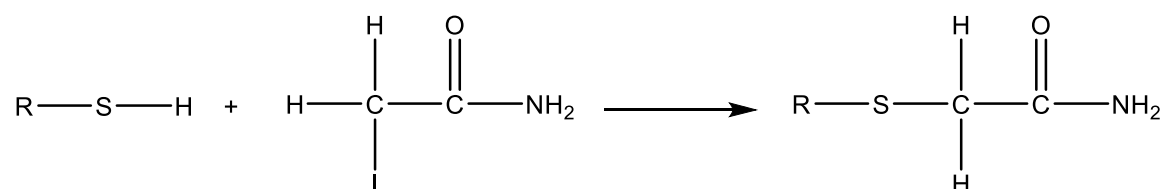
Enzyme: Choose the enzyme used to digest the protein. Trypsin is commonly used and will be the enzyme utilized in all the data in this module.

Missed Cleavages: The number of missed cleavages refers to the completeness of the enzyme digest. Did the enzyme trypsin cleave after every lysine and arginine residue in the protein? Or were some cleavages missed? The number of allowed **missed cleavages** should be set empirically, by running a standard and/or trying different values to see which gives the best score.

Modifications in database searching are handled in two ways.

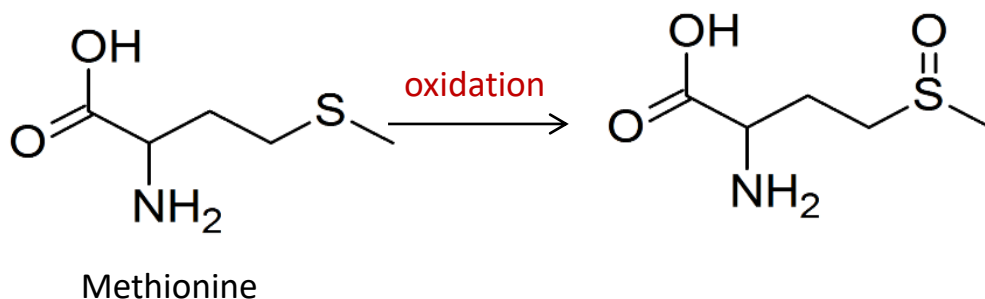
First, there are the **fixed** modifications. The most common example is the reduction and alkylation of cysteine. This reaction is performed to break disulfide bonds and prevent them from reforming. In the absence of disulfide bonds, the protein will be unfolded and the enzyme will be more effective in digesting the protein. Since all cysteines are modified, this is effectively just a change in the mass of cysteine. It carries no penalty in terms of search speed or specificity.

The alkylation agent used is iodoacetamide (select modification carbamidomethyl). In proteins, the reduced thiol group in cysteine is alkylated with iodoacetamide in the reaction shown:



In contrast, most post-translational modifications do not apply to all instances of a residue. For example, phosphorylation might affect just one serine in a protein containing many serines and threonines. These **variable** or non-quantitative modifications are expensive in the sense that they increase the search space. This is because the software has to permute out all the possible arrangements of modified and unmodified residues that fit to the peptide molecular mass. As more and more modifications are considered, the number of combinations and permutations increases geometrically, and we get a so-called combinatorial explosion.

One common **variable modification** is the oxidation of methionine shown:



Protein Mass: If the protein mass is known from its position in a 2D gel, this value can be entered. Usually, this adds little to the score, and the general advice is to leave this field blank.

Peptide Tolerance: Making an estimate of the **mass accuracy** doesn't have to be a guessing game. The Mascot Protein View report includes graphs of mass errors.

One way to evaluate the mass accuracy of the mass spectrometer is to run a standard and look at the error graphs for the correct match. Another method of evaluating mass accuracy is to compare the experimental value of a trypsin autolysis peak with the theoretical value.

In the data set provided, one trypsin autolysis peak had a measured mass of 1045.54 and the theoretical mass is 1045.12. The measurement indicates the mass spectrometer has mass error of approximately 0.42 Da. (Note: Da is the same as amu).

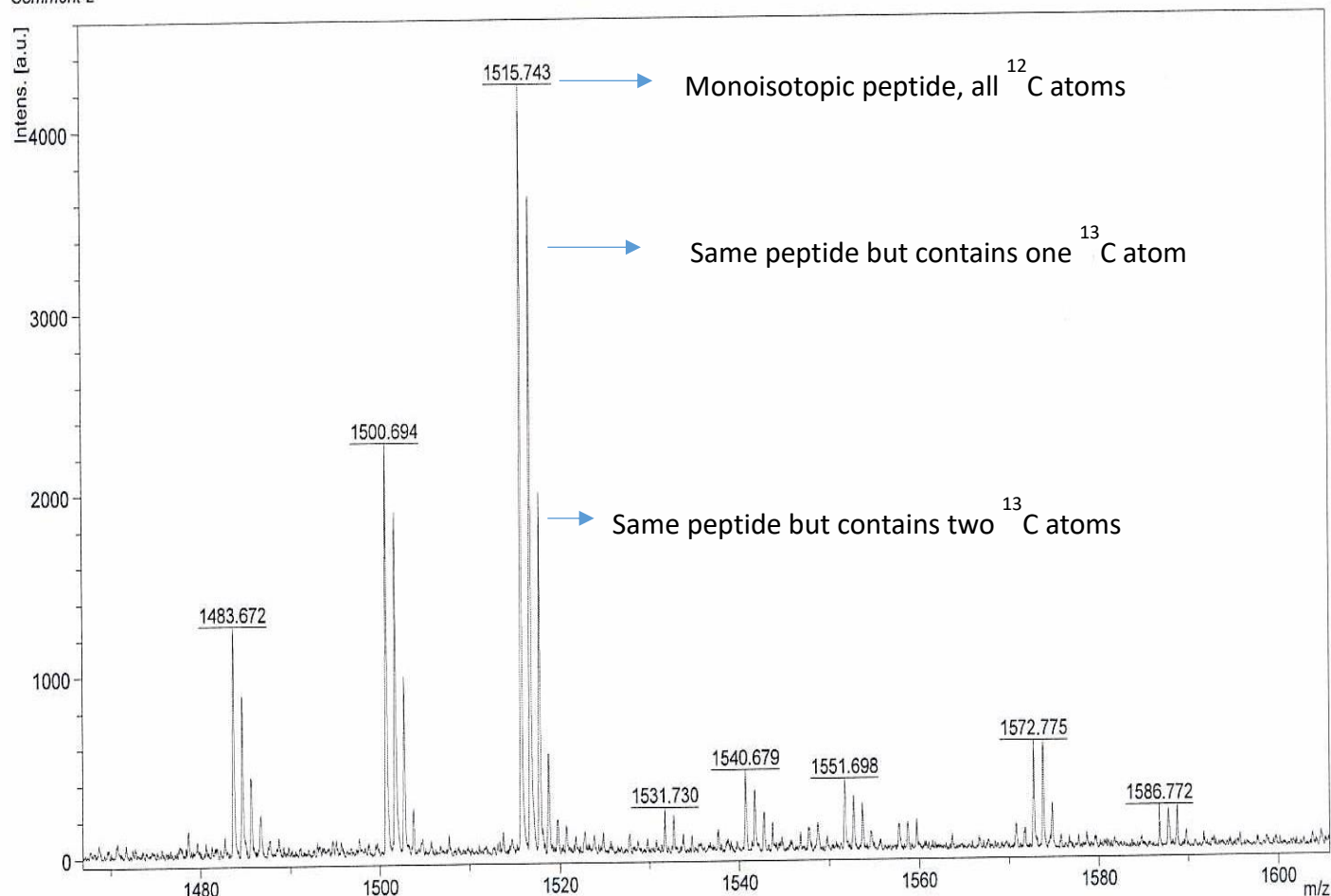
Mass values: Most frequently MALDI produces the singly charged molecular ion (MH^+). Your peak list will only contain M_r values (relative molecular mass) if the peak picking software has 'de-charged' the measured m/z values. Peak picking software may be programmed to do this because the data contained a mixture of charge states.

Most modern instruments produce **monoisotopic** mass values. You will only have **average** masses if the entire isotope distribution has been centroided into a single peak, which usually implies very low resolution.

The following MALDI-TOF mass spectrum of a protein digest zooms in on the mass region of different peptides near m/z 1500. The isotope distribution in a peptide with m/z 1515.7 is shown. The natural abundance of carbon-12 is 98.90% and carbon-13 is 1.10%. Therefore, peptides with a large number of carbon atoms will contain significant contributions to the $M+1$ peak and $M+2$ peak from carbon-13 atoms. The monoisotopic peptide contains all carbon-12 atoms. The $M+1$ peak has one carbon-13 atom and the $M+2$ peak has two carbon-13 atoms.

Comment 1 Gel band A or L (as labeled) after C18 Ziptip

Comment 2



Data Input: The first requirement for a Peptide Mass Fingerprint (PMF) search is a **peak list** (a list of m/z values). Peak lists are text files and come in various different formats. You can also copy and paste a list of values into the **query** area of the search form, or even type them in. Each m/z value goes on a separate line. If you also have an intensity value for the peak, this follows the m/z value, separated by a space or a tab.

Reading Questions

1. You are analyzing a protein from *E. coli*.

a. What is the advantage of setting the taxonomy to *E. coli*?

A. There are a fewer number of theoretical masses for peptides that could be matched. Therefore a mass match is more significant.

b. What is a disadvantage of setting the taxonomy to *E. coli*, instead of a more general class of bacteria to which *E. coli* belongs (Proteobacteria).

A. If the database for E. coli. is not complete then the protein will not be identified. If the more general class of bacteria were chosen a homologous protein from a different bacteria could be matched.

2. a. What is meant by the search parameter “missed cleavages?”

A. The enzyme trypsin cleaves after lysine and arginine. If trypsin did not cleave after a lysine residue, this would be a missed cleavage.

b. How will one missed cleavage affect the number of peptides created after digestion with trypsin?

A. There will be one less peptide than if there were zero missed cleavages. The peptide with the missed cleavage will have a larger mass than the predicted peptide because it contains two peptides.

3. A common fixed modification is carbamidomethyl. Why is a protein chemically modified in this way?

A. The cysteine residues are reduced and alkylated to prevent the formation of disulfide bonds. This keeps the protein in an unfolded state and increases digestion efficiency (fewer missed cleavages).

4. Briefly describe one method for determining the peptide tolerance (or the mass accuracy) of the mass spectrometer.

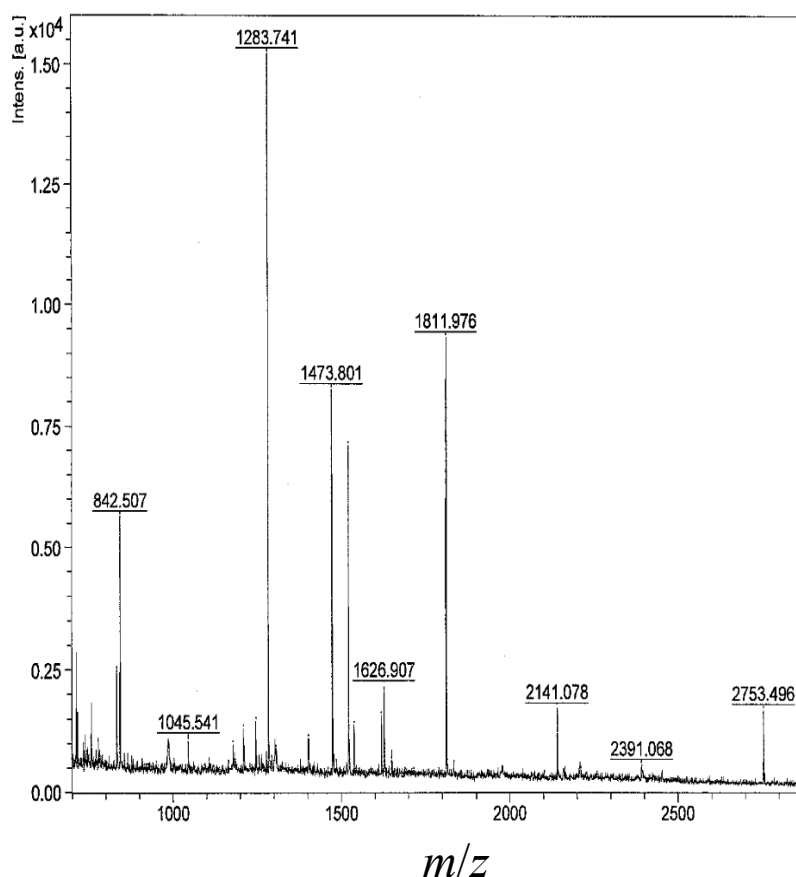
A. An experimental value of a trypsin autolysis peak can be compared to a theoretical value. For example, the theoretical value of a trypsin autolysis peak is 1045.12. If the experimental value is 1045.62, then the peptide tolerance would be 0.50 Da.

Instructor Note: The next part of the module involves working with protein digest data and varying search parameters in order to understand the effect these parameters on the protein score. The protein score is related to the probability that the protein has been correctly identified.

Performing a Peptide Mass Mapping Search: Now that you understand the various search parameters, you are now ready to perform a peptide mass fingerprint search in MASCOT.

1. Go to www.matrixscience.com and choose “Mascot search database” “Peptide mass fingerprint”, and “Perform search”
2. A good set of search parameters to start with are:
 Database: SwissProt
 Taxonomy: *Escherichia Coli*
 Enzyme: Trypsin
 Missed Cleavages: 1
 Fixed Modification: Carbamidomethyl
 Variable Modification: Oxidation of M
 Protein Mass: leave blank
 Peptide Tolerance: ± 1 Da
 Mass Values: MH^+
 Monoisotopic
 Report Top 5 Hits
3. We will start with the MALDI-TOF data for the protein from *E. coli* cut from a 2D gel. Copy and paste the m/z values in the table. Don’t forget to remove the trypsin autolysis peaks or matrix clusters from the data set.

m/z
832.312
842.53
1045.541
1179.58
1210.545
1247.598
1283.741
1307.679
1401.67
1403.711
1473.801
1521.775
1537.764
1618.801
1626.907
1648.88
1811.976
2141.078
2391.068
2753.496



4. Record the search results.

What protein has the highest score?

What is its protein score? What score is needed for significance?

Click on the identity of the protein for information about sequence and which peptide masses were found experimentally.

How many mass values were searched?

How many mass values were matched?

What was the percent sequence coverage?

Instructor Note: Students can run the search on computers and then compare with the results shown here.

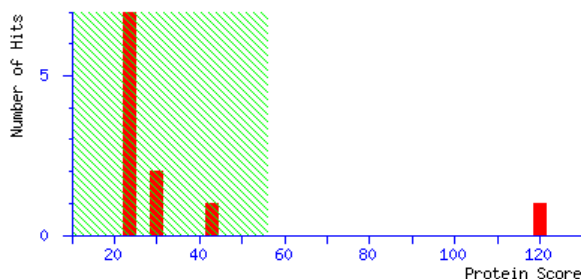
Results Summary: The first results screen identifies the protein as cysteine synthase A with a protein score of 120. Scores outside the green region (>56) are significant. A score of 120 indicates that there is a high probability that the protein has been correctly identified.

MATRIX SCIENCE Mascot Search Results

User : Jill Robinson
Email : jirobins@indiana.edu
Search title : E coli protein
Database : SwissProt 2016_06 (551385 sequences; 196948568 residues)
Taxonomy : Escherichia coli (23011 sequences)
Timestamp : 8 Jul 2016 at 13:15:56 GMT
Top Score : 120 for **CYSK_EC057**, Cysteine synthase A OS=Escherichia coli O157:H7 GN=cysK PE=:

Mascot Score Histogram

Protein score is $-10 \cdot \log(P)$, where P is the probability that the observed match is a random event.
Protein scores greater than 56 are significant ($p < 0.05$).



Click on the protein identity for more information.

Concise Protein Summary Report

Format As: Concise Protein Summary [Help](#)

Significance threshold $p < 0.05$ Max. number of hits 5

Preferred taxonomy: All entries

Re-Search All Search Unmatched

Rank	Protein Identity	Mass	Score	Expect	Matches
1.	CYSK_EC057	34525	120	2.3e-08	13
	Cysteine synthase A OS=Escherichia coli O157:H7 GN=cysK PE=3 SV=2				
	CYSK_EC01	34525	120	2.3e-08	13
	Cysteine synthase A OS=Escherichia coli (strain K12) GN=cysK PE=1 SV=2				
	MDTP_EC06	53557	31	17	6

Protein Score

The optimum data set for a peptide mass fingerprint is, of course, all of the correct peptides and none of the wrong ones. By correct, we mean that the textbook enzyme cleavage rules were followed, and only specified modifications are present. Sadly, real life data are generally far from ideal, and it is almost unknown to get every single experimental mass value matching and 100% sequence coverage. However, it is not always recognized that having too many peptide mass values can create similar difficulties to having too few.

Imagine a tryptic digest of a 20 kDa protein. We would expect something around 20 perfect cleavage peptides. If the digest was incomplete, or there was a non-quantitative modification, we might expect to double the number of peptides observed.

If 100 peaks are taken from the mass spectrum of this digest and submitted to Mascot then either 60 to 80 peaks are noise or there are extensive non-quantitative modifications. Either possibility is bad news for search specificity. The peaks which cannot be matched correctly will still contribute to the population of random matches.

The Mowse Scoring Algorithm is described in [Pappin, 1993]. (Reference available on MASCOT database)

The first stage of a Mowse search is to compare the calculated peptide masses for each entry in the sequence database with the set of experimental data. Each calculated value which falls within a given mass tolerance of an experimental value counts as a match.

Rather than just counting the number of matching peptides, Mowse uses empirically determined factors to assign a statistical weight to each individual peptide match.

Probability Based Scoring

Mascot incorporates a probability based implementation of the Mowse algorithm. The Mowse algorithm is an excellent starting point because it accurately models the behavior of a proteolytic enzyme. By casting the Mowse score into a probabilistic framework a simple rule can be used to judge whether a result is significant or not.

Matches using mass values are always handled on a probabilistic basis. The total score is the absolute probability that the observed match is a random event. Reporting probabilities directly can be confusing because they encompass a very wide range of magnitudes, and also because a "high" score is a "low" probability. For this reason, we report scores as $-10 \cdot \text{LOG}_{10}(P)$, where P is the absolute probability. A probability of 10^{-20} thus becomes a score of 200.

Significance Level

Given an absolute probability that a match is random, and knowing the size of the sequence database being searched, it becomes possible to provide an objective measure of the significance of a result. A commonly accepted threshold is that an event is significant if it would be expected to occur at random with a frequency of less than 5%. This is the value which is reported on the master results page.

The master results page for typical peptide mass fingerprint search reports that "Scores greater than 56 are significant ($p < 0.05$).” The protein with the score of 120 is a nice result because the highest score is highly significant, leaving little room for doubt.

After clicking on the protein with the top score, additional information from the search is displayed (as shown in the screen capture below). The molecular weight (34,525 Da) and pI value (5.83) are provided. If the protein was cut from a 2D gel, the position in the gel should correlate with the molecular weight of pI value of the protein identified. The protein sequence coverage was 49% and 13 of the 18 mass values that were searched matched the protein of interest.

MATRIX SCIENCE MASCOT Search Results

Protein View: CYSK_ECO57

Cysteine synthase A OS=Escherichia coli O157:H7 GN=cysK PE=3 SV=2

Database: SwissProt
 Score: 120
 Expect: 2.3e-08
 Nominal mass (M_r): 34525
 Calculated pI: 5.83
 Taxonomy: [Escherichia coli O157:H7](#)

Sequence similarity is available as [an NCBI BLAST search of CYSK_ECO57 against nr.](#)

Search parameters

Enzyme: Trypsin: cuts C-term side of KR unless next residue is P.
 Fixed modifications: [Carbamidomethyl \(C\)](#)
 Variable modifications: [Oxidation \(M\)](#)
 Mass values searched: 18
 Mass values matched: 13

Protein sequence coverage: 49%

Matched peptides shown in **bold red**.

```

1 MSKIFEDNSL TIGHTPLVRL NRIGNGRILA KVESRNPSFS VKCRIGANMI
51 WDAEKRGVLK PGVELVEPTS GNTGIALAYV AAARGYKLTLPMPETMSIER
101 RKLKALGAN LVLTEGAKGM KGAIQKAEEL VASNPEKYL LQQFSNPANP
151 EIHEKTTGPE IWEDTDGQVD VFIAGVGTGG TLTGVSRYIK GIKGKTDLIS
201 VAVEPTDSPV IAQALAGEEI KPGPHKIQGI GAGFIPANLD LKLVDKVIGI
251 TNEEAISTAR RLMEEEGILA GISSGAARVA ALKLQEDSEF TNKNIVVILP
301 SSGERYLSTA LFADLFTEKE LQQ
  
```

Discussion Questions

1. The protein identified has a very high score; however, less than half of the sequence was matched.

Why can a protein have a high score even with low sequence coverage?

A. There is a very small chance that any other protein can have the exact same peptides as found in the identified protein.

2. What are some experimental reasons for low sequence coverage? In other words, why are some peptides not found in the MALDI-TOF data?

A. Some peptides such as the first one (MSK) are in a low mass range and not displayed in the MALDI-TOF data from 500-3000 Da. Some very large peptides also may have a mass above 3000 Da.

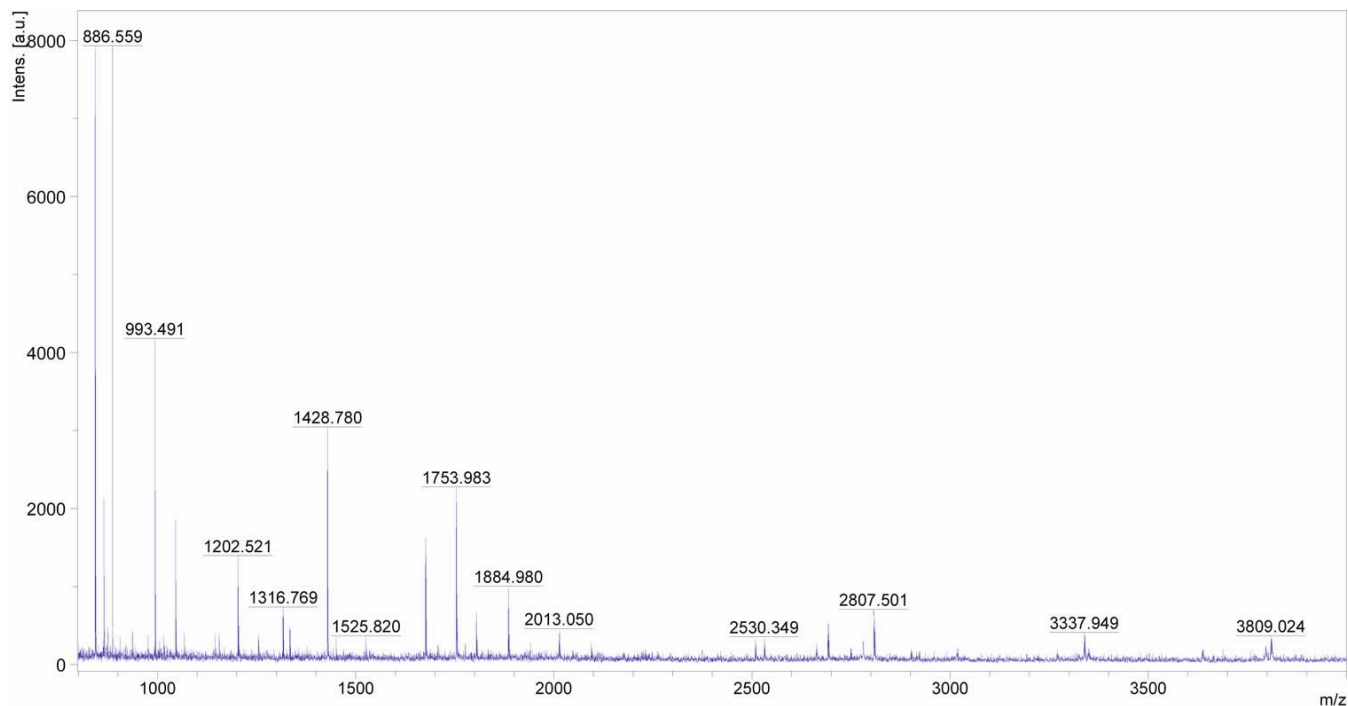
Some peptides can be very hydrophobic and may not dissolve in aqueous solution in sample preparation steps. Some peptides may not ionize easily and therefore have low intensity (or be missing) in the mass spectrum. Modification such as phosphorylation may have changed the mass of the peptide. The enzyme may have been very inefficient and many cleavages were missed.

Peptide Mass Mapping for Protein Identification

In this computer exercise, Mascot search parameters will be varied to explore their effect on protein score. The following MALDI-TOF data for an *E. coli* protein cut from a 2D gel will initially give a low protein score using default search parameters. The parameters will then be changed in a systematic way to see if a significant protein score can be achieved.

Open the Excel File: Proteomics Data AST4

The data is shown here as well.



m/z	Intensity
842.589	1962
864.585	483
874.496	99
886.559	93
936.473	91
976.436	87
993.491	1285
1015.477	93
1045.643	571
1067.664	99
1145.477	125
1155.598	142
1202.521	520
1254.658	153
1316.769	246
1333.777	186
1428.78	1282
1450.796	69

1525.82	119
1675.941	862
1753.983	1171
1804.047	307
1884.98	508
2013.05	271
2094.001	147
2508.378	155
2530.349	141
2662.519	162
2691.481	333
2807.501	476
3337.949	307
3794.978	221
3809.024	338

1. Use the following default search parameters to run a Peptide Mass Fingerprint search in Mascot search to identify the protein with the file name AST4.

Database: SwissProt
Taxonomy: All
Enzyme: Trypsin
Missed Cleavages: 1
Fixed Modification: Carbamidomethyl
Variable Modification: Oxidation of M
Protein Mass: leave blank
Peptide Tolerance: ± 1 Da
Mass Values: MH^+
Monoisotopic
Report Top 5 Hits

What protein has the highest score?

What is its protein score? What score is needed for significance?

How many mass values were searched?

How many mass values were matched?

What was the percent sequence coverage?

A. Students should remember to remove the m/z data for trypsin autolysis peaks (842 and 1045) to obtain these results.

What protein has the highest score? Lysozyme C (Chicken)

Protein score? 62 What score is needed for significance? 70

How many mass values were searched? 8

How many mass values were matched? 31

What was the percent sequence coverage? 62%

2. Use the same masses and initial search parameters and **change the taxonomy to Metazoa (Animals)** because the sample was from chicken.

What protein has the highest score?

Protein score? What score is needed for significance?

How many mass values were searched?

How many mass values were matched?

What was the percent sequence coverage?

The score for significance changed between all taxonomies and Metazoa. What do you think is the reason for the change?

A.

What protein has the highest score? Lysozyme C (Chicken)

Protein score? 62 What score is needed for significance? 63

How many mass values were searched? 8

How many mass values were matched? 31

What was the percent sequence coverage? 62%

The scientist typically knows the species from which their sample was obtained so it is not unreasonable to specify Metazoa if the protein came from a chicken. The score for significance dropped from 70 to 63 because there are fewer total proteins in the search once other organisms such as bacteria and plants are eliminated.

3. Retain the taxonomy as Metazoa and **vary the number missed cleavages.**

Zero missed cleavages:

What protein has the highest score?

Protein score? What score is needed for significance?

How many mass values were searched?

How many mass values were matched?

What was the percent sequence coverage?

A. *What protein has the highest score? Lysozyme C (Chicken)*

Protein score? 62 What score is needed for significance? 63

How many mass values were searched? 6

How many mass values were matched? 31

What was the percent sequence coverage? 54%

Two missed cleavages:

What protein has the highest score?

Protein score? What score is needed for significance?

How many mass values were searched?

How many mass values were matched?

What was the percent sequence coverage?

*A. What protein has the highest score? Lysozyme C (Chicken)
Protein score? 52 What score is needed for significance? 63
How many mass values were searched? 8
How many mass values were matched? 31
What was the percent sequence coverage? 62%*

Summarize how the number of missed cleavages parameter affects protein score. Can you explain why the score changes?

A. Zero missed cleavages lowered the number of peptide matches and sequence coverage from 62% to 54%, but the protein score did not change. Zero missed cleavages lowers the number of possible masses to match and even though sequence coverage dropped, the score did not.

Two missed cleavages retains the same number of peptide matches as the original search and the same sequence coverage (62%) but the score is lowered from 62 to 52 because the number of possible masses increases making it more likely to get a random match.

4. Vary the mass tolerance.

Repeat the search choosing 1 missed cleavage and a peptide tolerance of 0.5 Da.
What protein has the highest score?
Protein score? What score is needed for significance?
How many mass values were searched?
How many mass values were matched?
What was the percent sequence coverage?

*A. What protein has the highest score? Lysozyme C (Chicken)
Protein score? 82 What score is needed for significance? 63
How many mass values were searched? 8
How many mass values were matched? 31
What was the percent sequence coverage? 62%*

Repeat the search choosing 1 missed cleavage and a peptide tolerance of 0.3 Da.
What protein has the highest score?
Protein score? What score is needed for significance?
How many mass values were searched?
How many mass values were matched?
What was the percent sequence coverage?

*A. What protein has the highest score? Lysozyme C (Chicken)
Protein score? 94 What score is needed for significance? 63
How many mass values were searched? 8
How many mass values were matched? 31
What was the percent sequence coverage? 62%*

Repeat the search choosing 1 missed cleavage and a peptide tolerance of 0.1 Da.

What protein has the highest score?

Protein score? What score is needed for significance?

How many mass values were searched?

How many mass values were matched?

What was the percent sequence coverage?

A. What protein has the highest score? Ribosomal Protein S7 (Human)

Protein score? 45 What score is needed for significance? 63

How many mass values were searched? 6

How many mass values were matched? 31

What was the percent sequence coverage? 30%

Summarize how the mass tolerance affects score. Can you explain why the score changes?

A. As the mass tolerance parameter was decreased the protein score was increased if all the number of peptide matches did not change. The score increased because the chance the experimental mass of a peptide could be randomly matched was decreased. Once the mass tolerance was lower than the accuracy of the mass spectrometer, experimental peptide masses were no longer matched. The score was then lowered significantly and a different protein was identified.

5. Vary modifications.

Set the number of missed cleavages to 1 and mass tolerance to 0.30 Da. **Choose only carbamidomethyl (C) as a fixed modification and no variable modification.**

What protein has the highest score?

Protein score? What score is needed for significance?

How many mass values were searched?

How many mass values were matched?

What was the percent sequence coverage?

A. What protein has the highest score? Lysozyme C (Chicken)

Protein score? 103 What score is needed for significance? 63

How many mass values were searched? 8

How many mass values were matched? 31

What was the percent sequence coverage? 62%

Summarize how the selection of modifications affects score. Can you explain why the score changed?

A. The number of experimental peptides matched was still 8 of 31 and the sequence coverage was 62%. The score greatly increased when the variable modification, oxidation of M, was removed. This indicates that this modification did not occur and the score increased because the possible number of masses decreased. The less masses in the search, the chance of a random match decreases.

Homework (Peptide Mass Mapping)

1. Analyze data from three of the proteins in the data file (Proteomics homework data).

To turn in for each protein:

- a. Record the search parameters that were used to generate the highest probability score.
- b. What masses were used and which were discarded? Explain your reasoning.
- c. Record the protein identity, probability score, molecular weight, pI, number of mass values searched and matched, and percent sequence coverage.
- d. Interpret the results.

***Note: The reduction and alkylation procedure was not efficient in this set of data. It may help scores to not use carbamidomethyl as a fixed modification. (Do not select any fixed modifications if you cannot get a significant protein score.)

A.

N2- lysozyme
N3- carbonic anhydrase
N4 - Albumin
N5- Fetuin
N6- myoglobin
N7 - cytochrome C

2. Challenge Problem

Imagine that you have performed a 2D gel separation of proteins from healthy and cancerous cells and have identified a protein implicated in the cancerous state by cutting out the spot, digesting with trypsin, and performing peptide mass mapping. What else could you do experimentally to increase your level of certainty that the protein from the gel spot was identified correctly?

A. The protein could be digested with a different enzyme and a peptide mass fingerprint search performed in MASCOT. If the same protein were identified using a different enzyme, it is very likely that the identification is correct.

3. Challenge Problem

The following data were obtained by analytical chemistry students at Indiana University. The students grew *E. coli* samples at two different temperatures (37°C and 46°C). The cells were lysed, proteins isolated, and 2D gel electrophoresis performed. Based upon differences in the 2D gel pattern between the high and low temperature *E. coli* samples, one protein spot (from high temp. sample) was analyzed as it was suspected to be a **heat shock protein**.

Heat shock proteins (HSP) are a class of functionally related proteins involved in the folding and unfolding of other proteins. Their expression is increased when cells are exposed to elevated temperatures or other stress. This increase in expression is transcriptionally regulated. The dramatic upregulation of the heat shock proteins is a key part of the heat shock response and is induced primarily by heat shock factor (HSF). HSPs are found in virtually all living organisms, from bacteria to humans.

Change search parameters for the following data set to see if you can achieve a significant score for a heat shock (chaperone protein).

Sample	m/z	Signal
	719.3664	262.6387
	842.4825	4195.775
	855.0103	204.8235
	861.033	273.0225
	864.4673	410.5345
	877.0083	428.2205
	892.9921	220.0347
	1045.548	444.506
	1567.886	1874.742
	1694.763	476.5625
	1707.792	276.5812
	1845.939	1238.335
	1939.982	243.3293
	2402.261	209.4211

High temp. *E. coli* gel
Between MW band 5 (50 kDa)
MW band 6 (75 kDa)

A.

Search Parameters:

Enzyme: Trypsin

Database: Swiss Prot

Taxonomy: E coli

Fixed Modifications: Carbamidomethyl

Variable Modifications: none

Peptide Tolerance: 0.1 Da

Mass removed: Trypsin autolysis peaks 842 and 1045 Da

60 kDa Chaperonin 1 from *E. coli*.

Learning Objectives

At the end of this assignment, students will be able to:

1. Describe the operations of a triple quad mass spectrometer for tandem MS experiments
2. Identify *b* and *y* ions and use them to determine the sequence of a short peptide by *de novo* sequencing

Section 4A. Tandem MS

While enzymes, such as trypsin, can be used to cleave proteins and peptides at specific amino acid linkages, we can also fragment peptides inside of a mass spectrometer to obtain additional information. These types of experiments are called tandem MS or MS-MS experiments. These experiments are particularly helpful for “shotgun” proteomics or bottom-up proteomics. In these experiments, protein mixtures are first digested with enzymes (such as trypsin), then separated by one or more chromatography steps, and then electrosprayed into a mass spectrometer. A mass analyzer is then used to select a precursor ion with a specific m/z value for fragmentation. Fragmentation requires that some energy be added to the system. The most common method of fragmentation in MS-MS experiments is collision-induced dissociation (CID). In CID, the precursor ion is accelerated into an interaction cell that contains a collision gas, such as helium or nitrogen. When the precursor ion collides with the collision gas, the ion can fragment into two fragments, an ion and a neutral. The fragment ions are then analyzed to produce the MS/MS spectrum.

To better understand the operation of a commonly used triple quadrupole or “triple quad” MS/MS, watch [an animation](#), then answer the questions below.

Video Questions

1. What is the purpose of the skimmer in the instrument?

A. The skimmer helps to remove neutrals so that only ions enter the mass analyzer.

2. The animation uses color to indicate difference m/z value ions. Which “color” of ion is selected as the precursor ion?

A. The green ions are the precursor. They reach the end of the first quadrupole and enter the collision cell, while other m/z value ions are removed by the mass filter effect of the quadrupole.

3. Although this style of instrument is commonly called a “triple quad,” the collision cell is not actually a quadrupole. What is it?

A. It is a hexapole.

4. The last quadrupole selected fragment ions to be sent to the detector. Neutrals also pass through this quadrupole. Why don’t they produce a signal at the detector?

A. The conversion dynode (detector) is at 90° relative to the quadrupole ion path. The electrical potential on the dynode draws ions down, where their impact produces electrons that become the amplified current at the detector. Neutrals are not affected by this electrical potential so they continue along the quadrupole axis and do not result in current.

MS/MS experiments are useful because the fragment m/z values give information about the analyte's molecular structure. Low energy CID often produces small neutral losses, such as H_2O , CH_3OH , CO , CO_2 , NH_3 , and CN . Higher energy collisions can lead to retrosynthetic reactions, in which characteristic bonds in the precursor ion are broken. For example, CID of peptides often results in cleavage of peptide bonds along the backbone.

Reading Question

1. Complete the table below to summarize the expected mass differences from common neutral losses. Round expected masses to the nearest amu.

Table 1. Common neutral losses produced by CID of peptide ions.

Neutral Loss	Mass (amu)
NH_3	
H_2O	
CN	
CO	
CH_3OH	
CO_2	

A.

Table 1. Common neutral losses produced by CID of peptide ions.

Neutral Loss	Mass (amu)
NH_3	17
H_2O	18
CN	26
CO	28
CH_3OH	32
CO_2	44

The identity of the precursor ion can be scanned through the mass range so that a mass spectrum of each precursor ion's corresponding fragments is obtained. These experiments produced large quantities of data extremely rapidly. Interpreting these large data sets usually involves specialized software programs that identify peptides from their fragments and then identify proteins from their peptides; however, for simple mixtures, the data may be interpreted manually since peptide fragmentation in tandem MS experiments is well-characterized.

Section 4B. CID of Peptides and *De Novo* Sequencing

Peptide fragments produced in tandem MS experiments are named using a letter-number scheme that identifies which bond was broken and which side of the peptide (the N-terminus or the C-terminus) became the charged fragment (Figure 1). As noted above, CID typically results in cleavage of the peptide bond and therefore produces b and y ions. When the N-terminal side of the peptide is the charged fragment and the C-terminus is a neutral loss, the result is a b ion. When the C-terminal side is the charged fragment, the result is a y ion. In either case, the ions are numbered sequentially from the charged terminus, meaning that the first N-terminal residue gives the b_1 fragment, and the first C-terminal residue gives the y_1 ion. Other fragmentation methods,

such as photodissociation, electron capture, and electron transfer dissociation can cleave other bonds within the peptide, giving rise to *a*, *x*, *c*, and *z* ions. In some cases, CID may also produce *a* ions, which differ from *b* ions by the absence of a carbonyl, making them 28 amu lighter than the corresponding *b* ion.

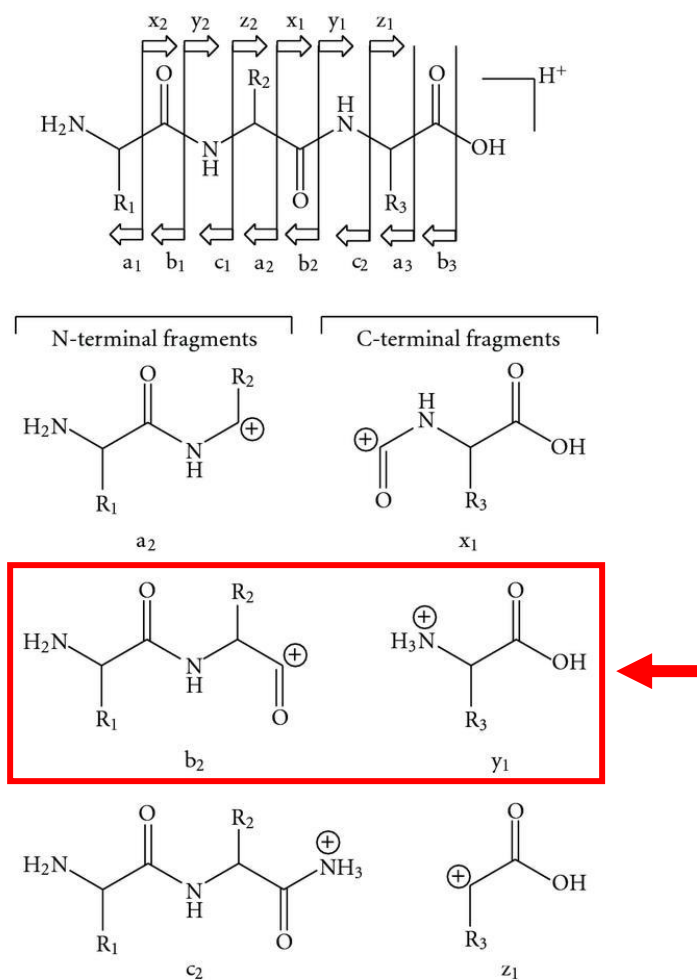


Figure 1. Fragment ions are named based on the bond cleaved and the location of the charge on the resulting fragments. CID most commonly produces *b* and *y* ions, which result from cleavage of the peptide bond with the charged fragment occurring on the N- or C-terminal fragment, respectively. Figure is adapted with permission from S. Banerjee and S. Mazumdar, *Int. J. Anal. Chem.*, **2012**, 2012, 282574 under a Creative Commons Attribution License.

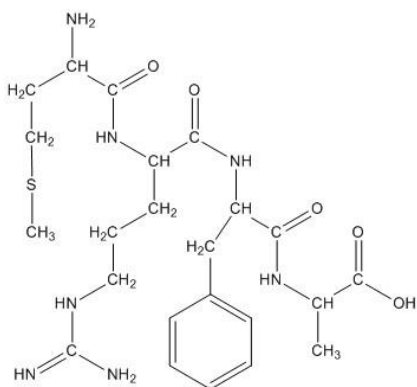


Figure 2. Structure of the tetrapeptide MRFA.

For appropriate CID conditions, it is often possible to obtain a series of *b* and *y* ions for a peptide (Figure 1). The mass differences between *b* ions (or between *y* ions) are characteristic of the amino acid residues which have been lost, and careful examination of the mass spectrum can yield the peptide sequence from the MS-MS data. For example, Figure 3 shows model data for fragmentation of the MRFA peptide. The mass difference between peaks from right to left are 71, 147, 156 amu, corresponding in mass to sequential losses of A, F, and R amino acid residues (Table 2), indicating that these amino acids form a series in the peptide backbone of the analyte.

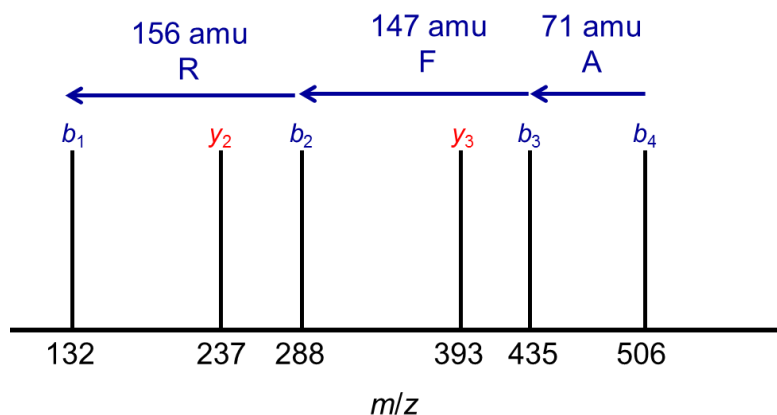


Figure 3. A model mass spectrum for MS-MS of the peptide, MRFA, which has a precursor m/z of 524.27.

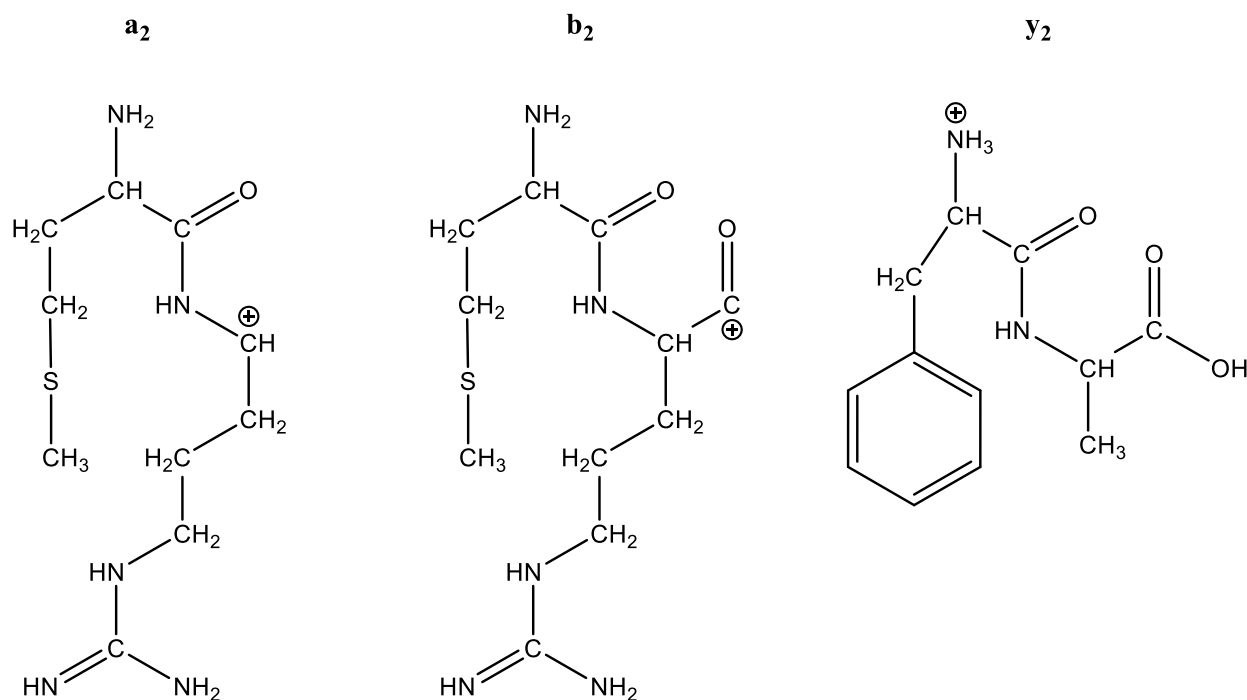
Table 2. Molecular weight information for all 20 naturally occurring amino acids.

Amino Acid	Single-Letter Code	Residue MW (amu)	Amino Acid MW (amu)
glycine	G	57.02	75.03
alanine	A	71.04	89.05
serine	S	87.03	105.04
proline	P	97.05	115.06
valine	V	99.07	117.08
threonine	T	101.05	119.06
cysteine	C	103.01	121.02
isoleucine	I	113.08	131.09
leucine	L	113.08	131.09
asparagine	N	114.04	132.05
aspartic acid	D	115.03	133.04
glutamine	Q	128.06	146.07
lysine	K	128.09	146.11
glutamic acid	E	129.04	147.05
methionine	M	131.04	149.05
histidine	H	137.06	155.07
phenylalanine	F	147.07	165.08
arginine	R	156.10	174.11
tyrosine	Y	163.06	181.07
tryptophan	W	186.08	204.09

Reading Questions

2. Sketch the a_2 , b_2 , and y_2 ions for the tetrapeptide MRFA shown in Figure 2. What are the expected masses of these fragments?

A.



The a_2 ion has an expected mass of 260. The b_2 ion (MR) has a mass of 288; note that this is 28 amu heavier than the a_2 fragment since the b_2 fragment has an additional carbonyl group. The y_2 ion (FA) has a mass of 237. These values are monoisotopic masses. Students may need to be reminded to use monoisotopic masses (from Table 2 below) rather than isotopic weighted averages for calculations of mass for mass spectrometry experiments.

3. What is the mass difference between the y_3 and the y_2 ion in Figure 3?

A. The mass difference is $393 - 237 = 156$ amu.

4. What amino acid residue corresponds to this mass difference? Does this make sense given the sequences of these two ions?

A. An arginine residue has a mass of 156 amu. This makes sense because the y_2 fragment ion is FA, and the y_3 fragment ion is RFA.

In practice, a complete b and y ion series may not be obtained, but it is often possible to deduce the peptide sequence from MS-MS data without referring to external databases or genomic data. This method of determining peptide structure is called *de novo* sequencing since the sequence is determined without reference to outside data. Because *de novo* sequencing does not rely on an external database, peptides can be identified even if they have unexpected post-translational modifications or arise from organisms with unsequenced genomes. Several research groups have developed algorithms to automate *de novo* sequencing from MS-MS data. These algorithms are designed to account for missing b and y ions, identify post-translational modifications, and address other challenges, including identification of oxidation and other chemical changes to proteins and peptides. Database searching is an alternative to *de novo* sequencing for longer peptides and proteins. Similarly to database searching for peptide mass fingerprinting, MS-MS database searching relies on

genomic data to predict the expected spectra for MS-MS analysis. The experimental spectra are then matched to the predicted spectra to make peptide and protein assignments.

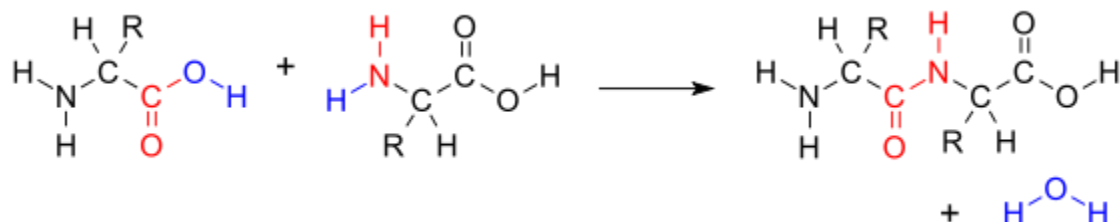
Discussion Questions

1. Consider the data in Table 2. By what value do the residue MW and the amino acid MW differ?

A. The mass of the amino acid is 18 amu heavier than the mass of the amino acid residue.

2. If needed, review the module introduction to peptides and proteins. Why is the MW of an amino acid residue in a peptide chain different from the mass of the full amino acid? Sketch a reaction to support your answer.

A. The mass of the amino acid residue is 18 amu lighter than the mass of the amino acid because the formation of peptide bonds at the N- and C-termini results in the loss of water from the amino acid.



3. Are there any amino acids that could not be distinguished from one another using a mass analyzer with unit resolution (e.g., a quadrupole ion trap)?

A. Leucine and isoleucine are isomers, so they cannot be distinguished regardless of resolving power since they have the same mass. Glutamine and lysine residues differ by 0.03 amu, meaning that they cannot be resolved by an instrument with unit resolution; however, they could be resolved by a high-resolution instrument.

4. Leu-enkephalin is a pentapeptide with the sequence YGGFL that is involved in neurotransmission. Tabulate the amino acid sequence and the expected monoisotopic MWs, and fragments of leu-enkephalin. Work in groups, dividing the required calculations among members.

Table 3. Expected precursor and fragment ions for CID MS-MS of leu-enkephalin.

Ion	Peptide Sequence	Expected m/z
$[M+H]^+$		
b_1		
b_2		
b_3		
b_4		
b_5		
y_1		
y_2		
y_3		
y_4		

A. Table 3. Expected precursor and fragment ions for CID MS-MS of leu-enkephalin.

Ion	Peptide Sequence	Expected m/z
$[M+H]^+$	YGGFL	556.28
b_1	Y	164.07
b_2	YG	221.09
b_3	YGG	278.11
b_4	YGGF	425.18
b_5	YGGFL- H_2O	538.27
y_1	L	132.10
y_2	FL	279.17
y_3	GFL	336.19
y_4	GGFL	393.21

Instructor note: If time is limited, you may also allow your students to calculate these masses using the Institute for Systems Biology's [peptide calculator](#).

5. Using your completed Table 3 above, identify as many of the peaks in the leu-enkephalin MS-MS spectrum shown in Figure 4 below as possible, remembering that CID can result in the small neutral losses that you noted in Table 1 in addition to retrosynthetic fragmentation.

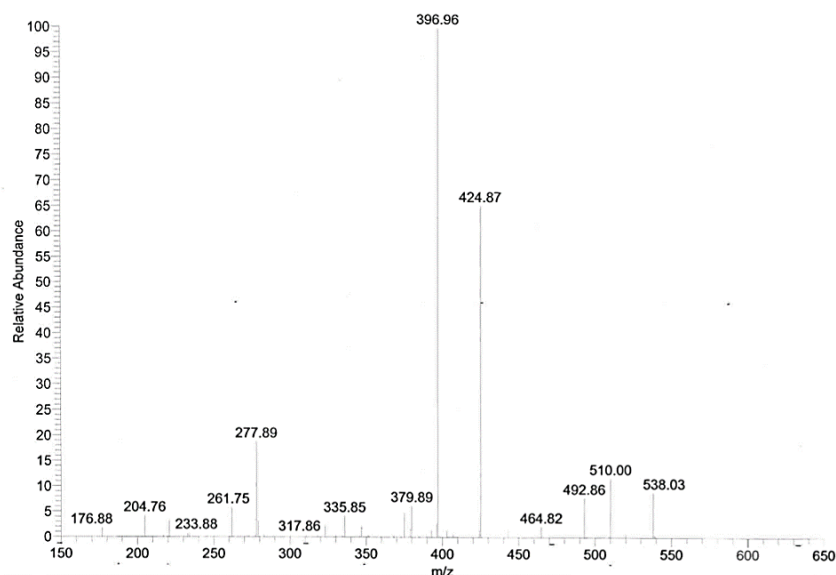
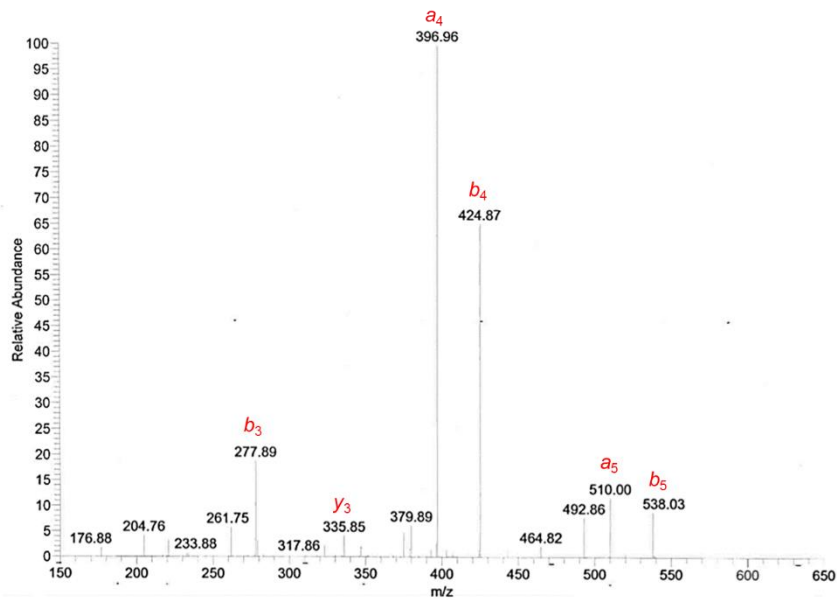


Figure 4. ESI-MS-MS spectrum of leu-enkephalin. Data obtained at Trinity College.

A. Note that the a_4 fragment may also be correctly identified as the b_4 fragment with the neutral loss of CO. Similarly, the a_5 fragment is b_5 -CO.



6. Imagine researchers take a sample of blood serum from a patient and perform a shotgun proteomics experiment with a trypsin digestion and tandem MS detection. Would you recommend a separation step in between the tryptic digest of the serum proteins and the mass spectrometric detection? Why or why not?

A. A separation step will be crucial in this application. Serum contains many different proteins, and without a separation step before the MS analysis each mass spectrum will contain peptides and fragments from multiple proteins, making data interpretation difficult and potentially impossible.

Challenge Question

To the best of your ability, use the MS-MS spectrum in Figure 5 below to *de novo* sequence an unknown tetrapeptide. To support your proposed sequence, make a table showing the expected mass of each fragment and the actual mass you observe for the peak.

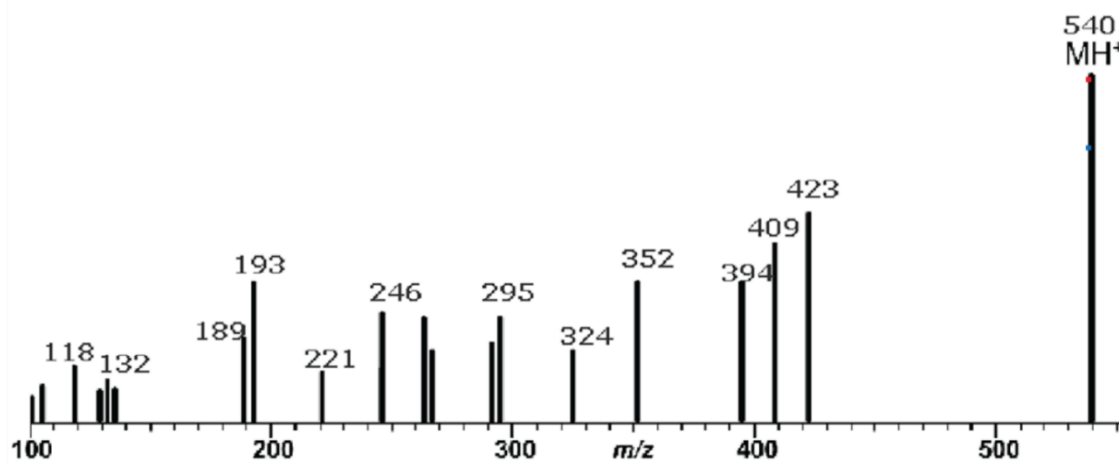
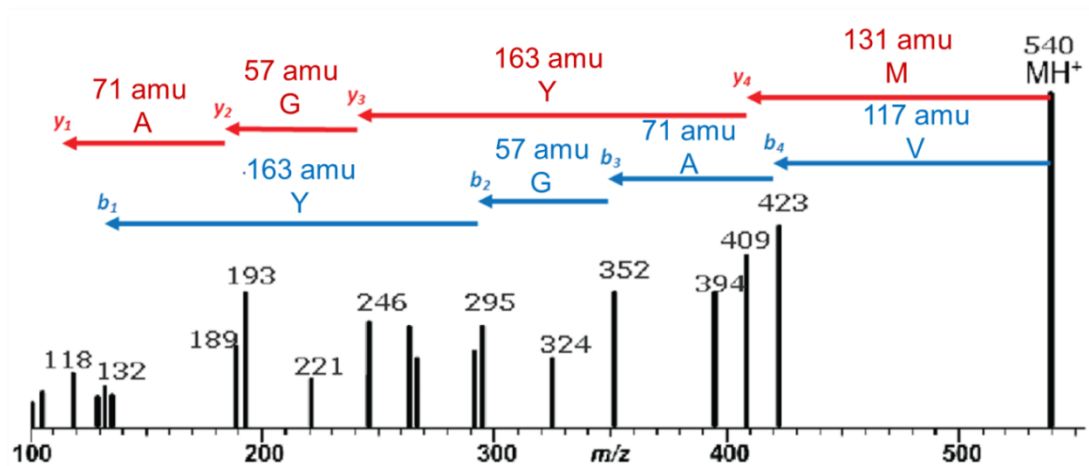
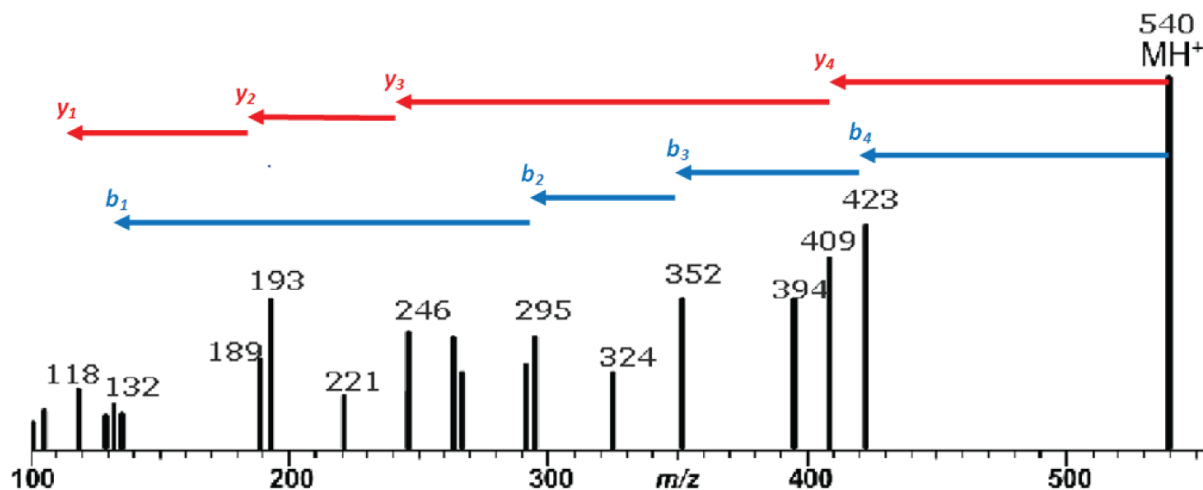


Figure 5. MS-MS mass spectrum of an unknown peptide. Data from F. Klink, *Separation Science*.

A. The plot below labels the b and y ion series along with the mass difference between fragments and corresponding amino acid residue. The peptide sequence is MYGAV, and the data is from F. Klink, "MS Solutions #10: Peptide Sequencing with Electrospray LC/MS Part 2: Interpretation of a Simple Spectrum," *Separation Science*, <http://www.sepscience.com/Information/Archive/All-Articles/246-/MS-Solutions-10-Peptide-Sequencing-with-Electrospray-LCMS-Part-2-Interpretation-of-a-Simple-Spectrum>



Note to instructors: To determine this sequence from the raw data, one key is to identify which peaks are b ions and which are y ions. This is best done using a print-out to calculate mass differences between various combinations of peaks, looking for mass differences that correspond to an amino acid residue. In particular, identifying the C-terminal valine is difficult because there is no b_5 ion. Instead, the difference between the $[M+H]^+$ peak and the b_4 ion corresponds to the mass of the entire valine residue. To make this assignment easier, you could identify which ions belong to which series (as in the figure shown below), provide a few of the peak assignments to start, or provide the sequence and ask students to identify the ions.



For those who are interested, [a more detailed tutorial](#) on de novo sequencing is available from Ion Sense.

References and Resources

R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, **2003**, 422, 198-207.

The Broad Institute, "MRM (Multiple Reaction Monitoring)," <https://www.broadinstitute.org/scientific-community/science/platforms/proteomics/mrm-multiple-reaction-monitoring>

F. Klink, "MS Solutions #9: Peptide Sequencing with Electrospray LC/MS Part 1: Ion Types and Nomenclature," *Separation Science*, <http://www.sepscience.com/Information/Archive/MS-Solutions/244-/MS-Solutions-9-Peptide-Sequencing-with-Electrospray-LCMS-Part-1-Ion-Types-and-Nomenclature>

F. Klink, "MS Solutions #10: Peptide Sequencing with Electrospray LC/MS Part 2: Interpretation of a Simple Spectrum," *Separation Science*, <http://www.sepscience.com/Information/Archive/All-Articles/246-/MS-Solutions-10-Peptide-Sequencing-with-Electrospray-LCMS-Part-2-Interpretation-of-a-Simple-Spectrum>

Ion Source, "De Novo Peptide Sequencing Tutorial," **2012**,
<http://ionsource.com/tutorial/DeNovo/DeNovoTOC.htm>