Section 3: Peptide Mass Mapping for Protein Identification

Learning Objectives

At the end of this assignment, you 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.

**Discussion Question**

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

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?
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.

![Mass Spectrum of Proteins](image)

**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](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?

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

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.

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

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:

\[ KE = zV \]  

(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?

   b. If all of the ions have the same charge, what will be true about their kinetic energies?
5. You may remember another equation for kinetic energy from physics:

\[ KE = \frac{1}{2}mv^2 \]  

(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?

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

c. Is the TOF mass analyzer dispersive or scanning?

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

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

If \( +HV = 2000 \text{V} \), how much kinetic energy is imparted to the ion in Joules?

1 elementary charge = \( 1.6 \times 10^{-19} \text{C} \)  

1 \text{V} = 1 \text{J/C}

f. What is the velocity of the ion in \( \text{m/s} \)?

1 \text{J} = 1 \frac{\text{kg} \cdot \text{m}^2}{\text{s}^2}  

1 \text{amu} = 1.66 \times 10^{-27} \text{kg}

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

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

Image adapted from UC Davis Fiehn Metabolomics Lab. 

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

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?
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.

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?

   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.

   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.
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: $+HV_2 > +HV_1$.

![Diagram of reflectron](image)

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.

b. Which ion penetrates farther into the reflectron?

c. How does this arrangement correct for the ions’ different starting positions in the source?
10. Examine the diagram of a reflectron TOF mass analyzer.

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)

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

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

**Genome Database**
Predicts all proteins for organism

**In-silico digest generated**
A text file containing the amino acid single letter code sequences of all proteins is created.

**Compare** measured peptide masses with masses predicted by in-silico digest.

A **scoring algorithm** reviews all matches and determines the probability a particular protein was present in the sample. Score is converted into a statistical probability.
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.
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.

![Structures of lysine and arginine](image)

**Lysine (K)**

**Arginine (R)**

**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).

2. Why is a well annotated genome for the organism of interest needed in the peptide mass mapping technique for identifying a protein?
3. Classify the side chains of R and K as acidic or 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).

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

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.)

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

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

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

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?

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.
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 NCBInr 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:

\[
\text{R-S-H} + \text{H-CCCNH}_2 \rightarrow \text{R-S-C=CNH}_2
\]
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:

![Methionine Oxidation](image)

**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.

**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*?
   
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).

2. a. What is meant by the search parameter “missed cleavages?”
   
b. How will one missed cleavage affect the number of peptides created after digestion with trypsin?

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

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

<table>
<thead>
<tr>
<th>m/z</th>
<th>Intensity (×10^4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>832.312</td>
<td></td>
</tr>
<tr>
<td>842.53</td>
<td></td>
</tr>
<tr>
<td>1045.541</td>
<td></td>
</tr>
<tr>
<td>1179.58</td>
<td></td>
</tr>
<tr>
<td>1210.545</td>
<td></td>
</tr>
<tr>
<td>1247.598</td>
<td></td>
</tr>
<tr>
<td>1283.741</td>
<td></td>
</tr>
<tr>
<td>1307.679</td>
<td></td>
</tr>
<tr>
<td>1401.67</td>
<td></td>
</tr>
<tr>
<td>1403.711</td>
<td></td>
</tr>
<tr>
<td>1473.801</td>
<td></td>
</tr>
<tr>
<td>1521.775</td>
<td></td>
</tr>
<tr>
<td>1537.764</td>
<td></td>
</tr>
<tr>
<td>1618.801</td>
<td></td>
</tr>
<tr>
<td>1626.907</td>
<td></td>
</tr>
<tr>
<td>1648.88</td>
<td></td>
</tr>
<tr>
<td>1811.976</td>
<td></td>
</tr>
<tr>
<td>2141.078</td>
<td></td>
</tr>
<tr>
<td>2391.068</td>
<td></td>
</tr>
<tr>
<td>2753.496</td>
<td></td>
</tr>
</tbody>
</table>

m/z 23
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.
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 \(-10\times\log_{10}(P)\), 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.

**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?

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

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

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?

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?

**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?

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

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?

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?
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?

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

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?

Summarize how the selection of modifications affects score. Can you explain why the score changed?
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 choose carbamidomethyl as a fixed modification. (Do not select any fixed modifications if you cannot get a significant protein score.)

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?

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).

High temp. E. coli gel
Between MW band 5 (50 kDa)  
MW band 6 (75 kDa)
<table>
<thead>
<tr>
<th>Sample</th>
<th>m/z</th>
<th>Signal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>719.3664</td>
<td>262.6387</td>
</tr>
<tr>
<td></td>
<td>842.4825</td>
<td>4195.775</td>
</tr>
<tr>
<td></td>
<td>855.0103</td>
<td>204.8235</td>
</tr>
<tr>
<td></td>
<td>861.033</td>
<td>273.0225</td>
</tr>
<tr>
<td></td>
<td>864.4673</td>
<td>410.5345</td>
</tr>
<tr>
<td></td>
<td>877.0083</td>
<td>428.2205</td>
</tr>
<tr>
<td></td>
<td>892.9921</td>
<td>220.0347</td>
</tr>
<tr>
<td></td>
<td>1045.548</td>
<td>444.506</td>
</tr>
<tr>
<td></td>
<td>1567.886</td>
<td>1874.742</td>
</tr>
<tr>
<td></td>
<td>1694.763</td>
<td>476.5625</td>
</tr>
<tr>
<td></td>
<td>1707.792</td>
<td>276.5812</td>
</tr>
<tr>
<td></td>
<td>1845.939</td>
<td>1238.335</td>
</tr>
<tr>
<td></td>
<td>1939.982</td>
<td>243.3293</td>
</tr>
<tr>
<td></td>
<td>2402.261</td>
<td>209.4211</td>
</tr>
</tbody>
</table>