Lecture 8 (9/27/19)

- **Reading:** Ch3; 97-102
- **Problems:** Ch3 (text); 18, 19, 20, 23
  Ch3 (Study guide); 9
  Ch4 (Study guide); 3

**NEXT**

- **Reading:** Ch4; 119-122, 125-126, 131-133
  Ch4; 123-124, 130-131, 133, 137-138
- **Problems:** Ch4 (text); 2, 3, 4, 8, 13, 14

---

**OUTLINE**

I. Protein Structure
   A. Primary
      1. Determination
         a. Sequence determination; CHEMICAL
            i. Cleavage of peptides bonds
            ii. Amino acid composition and stoichiometry
            iii. Determine number of chains;
            iv. Divide & Conquer;
            v. Edman Degradation
         b. Sequence determination; PHYSICAL
            i. Mass Spectrometry for proteins
            ii. Use of tandem MS/MS for sequence determination
            iii. Isolation of proteins by 2D PAGE; Isoelectric focusing x SDS-PAGE
         c. Sequence determination; BIOLOGICAL
            i. Genome sequenced
            ii. Bioinformatics to predict protein sequences in predicted genes
            iii. Use of CHEMICAL and/or PHYSICAL methods to get partial sequence
Determination of primary structure

1) Purify protein
2) Determine the amino-acid composition, including stoichiometry
3) Disrupt structure (2°, 3°, 4°, and disulfides)
4) **Determine the number of peptide chains by counting number of amino terminal ends**
5) Divide into fragments and determine sequence
6) Divide into different set of fragments and determine sequence
7) Determine overlaps and piece original sequence back together

**Fig 3-26**

**Determine the number of peptide chains by counting number of amino terminal ends**

- Frederick Sanger (1918-2013)
- Sanger’s Reagent
- Example:
  - Small tripeptide
  - AA comp = A,K,L
  - C-term = K
  - DNP-A
  - Sequence is A-L-K

![Chemical structures](image)

- 2,4-Dinitrofluorobenzene (DNFB)
- Polypeptide
- DNP-Polypeptide
- Absorbs at 353 nm (yellow)
Determine the number of peptide chains by counting number of amino terminal ends

Excitation at 330 nm
Emission at 519 nm

Determination of primary structure

1) Purify protein
2) Determine the amino-acid composition, including stoichiometry
3) Disrupt structure (2°, 3°, 4°, and disulfides)
4) Determine the number of peptide chains by counting number of amino terminal ends
5) Divide into fragments and determine sequence
6) Divide into different set of fragments and determine sequence
7) Determine overlaps and piece original sequence back together
Determination of primary structure

1) Purify protein
2) Determine the amino-acid composition, including stoichiometry
3) Disrupt structure (2°, 3°, 4°, and disulfides)
4) Determine the number of peptide chains by counting number of amino terminal ends
5) Divide into fragments and determine sequence
6) Divide into different set of fragments and determine sequence
7) Determine overlaps and piece original sequence back together

**Divide into fragments**

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Source</th>
<th>Specificity</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trypsin</td>
<td>Bovine pancreas</td>
<td>R&lt;sub&gt;n-1&lt;/sub&gt; = positively charged residue; R&lt;sub&gt;n&lt;/sub&gt; ≠ Pro</td>
<td>Highly specific</td>
</tr>
<tr>
<td>Chymotrypsin</td>
<td>Bovine pancreas</td>
<td>R&lt;sub&gt;n&lt;/sub&gt; = hydrophobic residue; R&lt;sub&gt;n-1&lt;/sub&gt; ≠ Pro</td>
<td>Cleans more slowly for R&lt;sub&gt;n-1&lt;/sub&gt; = Arg, His, Met, Leu</td>
</tr>
<tr>
<td>Elastase</td>
<td>Bovine pancreas</td>
<td>R&lt;sub&gt;n-1&lt;/sub&gt; = small neutral residues: Ala, Gly, Ser, Val; R&lt;sub&gt;n&lt;/sub&gt; ≠ Pro</td>
<td></td>
</tr>
<tr>
<td>Thrombin</td>
<td>Bacillus thermoproteolyticus</td>
<td>R&lt;sub&gt;n&lt;/sub&gt; = Leu, Phe, Trp, Tyr, Val, R&lt;sub&gt;n-1&lt;/sub&gt; ≠ Pro</td>
<td>Occasionally cleaves at R&lt;sub&gt;n&lt;/sub&gt; = Ala, Asp, His, Thr; heat stable</td>
</tr>
<tr>
<td>Pepain</td>
<td>Bovine gastric mucosa</td>
<td>R&lt;sub&gt;n&lt;/sub&gt; = Leu, Phe, Trp, Tyr; R&lt;sub&gt;n-1&lt;/sub&gt; ≠ Pro</td>
<td>Also others; quite nonspecific; pH optimum = 2</td>
</tr>
<tr>
<td>Endopeptidase VII</td>
<td>Staphylococcus aureus</td>
<td>R&lt;sub&gt;n-1&lt;/sub&gt; = Glu</td>
<td></td>
</tr>
</tbody>
</table>

**Proteolytic Cleavage**

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Source</th>
<th>Specificity</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carboxypeptidase A</td>
<td>Bovine pancreas</td>
<td>R&lt;sub&gt;n&lt;/sub&gt; = C-terminal; R&lt;sub&gt;n-1&lt;/sub&gt; ≠ Pro</td>
<td></td>
</tr>
</tbody>
</table>
Divide into fragments

Cyanogen Bromide Cleavage

Similar result as an endopeptidase, but leaves a homoserine lactone as the C-terminal residue

Separation and isolation of peptide fragments

Example: paper electrophoresis
Also, TLC, silica gel, etc.
Determination of primary structure

- Determine amino-acid composition
- Dansyl chloride or FDNB to determine amino-termini and number
- Proteases: Cleaves peptide bonds only after specific residues.
  - **Cleave protein with 2 different proteases.**
- Sequence fragments with **Edman degradation.** Piece together sequence from overlapping fragments.

---

**Determination of primary structure: Divide & Conquer**

Tryptsin

Digestion and Edman degradation

(Ala₂, Gly, Lys₂, Phe, Thr, Trp, Val)
Determination of primary structure: Divide & Conquer

1) Purify protein
2) Determine the amino-acid composition, including stoichiometry
3) Disrupt structure (2°, 3°, 4°, and disulfides)
4) Determine the number of peptide chains by counting number of amino terminal ends
5) Divide into fragments and determine sequence
6) Divide into different set of fragments and determine sequence
7) Determine overlaps and piece original sequence back together
Determine the Sequence: Edman degradation

Determination of primary structure

1) Purify protein
2) Determine the amino-acid composition, including stoichiometry
3) Disrupt structure (2°, 3°, 4°, and disulfides)
4) Determine the number of peptide chains by counting number of amino terminal ends
5) Divide into fragments and determine sequence
6) Divide into different set of fragments and determine sequence
7) Determine overlaps and piece original sequence back together
Determine overlaps and piece original sequence back together

### Determination of primary structure

THREE basic ways to know the primary structure. Only the CHEMICAL method will give the entire covalent structure, including any disulfide bonds. But other methods are more sensitive. One can classify these methods by:

- **CHEMICAL**
- **PHYSICAL**
- **BIOINFOMATICAL**

We just went through the CHEMICAL.

The PHYSICAL method still requires the same strategy, including purification, fragmentation, chromatography, and alignment.

But, instead of an Edman degradation the use of tandem Mass Spectrometry (MS) is employed.

**Determining the use of MS in biochemistry**

- Ions “fly” in a vacuum toward a target with a velocity \( \propto \frac{z}{m} \) (charge-to-mass ratio)
- Molecules with higher charge and lower mass get detected first.
- Molecules with a lower charge and higher mass get detected last.
- Plotted as m/z to read peaks from left to right
- Instruments can distinguish molecules with same charge by < 1 Da
Determine the Sequence: Tandem MS

The major problem in using MS for macromolecules is getting them to "fly" in a vacuum with a charge. TWO major methods:

1) Electro-Spray Ionization (ESI)
2) Matrix-Assisted Laser-Desorption Ionization (MALDI)

What is MALDI?

Determine the Sequence: Tandem MS

The major problem in using MS for macromolecules is getting them to "fly" in a vacuum with a charge. TWO major methods:

1) Electro-Spray Ionization (ESI)
2) Matrix-Assisted Laser-Desorption Ionization (MALDI)

What is MALDI?
How do you use MS to get protein sequences?

- Mass spectrometry uses mass-to-charge ratio of different ions to determine mass.
- Tandem MS-MS: First selects a peptide, then fragmentation, and second determines mass.
- By comparing all of fragments, those that different by mass of one amino acid to determine sequence.

Determine the Sequence: Tandem MS

Matrix-Assisted Laser-Desorption Ionization (MALDI)
Determine the Sequence: Tandem MS

- **EXAMPLE**

  AVAW

<table>
<thead>
<tr>
<th>AA</th>
<th>MW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>90</td>
</tr>
<tr>
<td>Val</td>
<td>109</td>
</tr>
<tr>
<td>Trp</td>
<td>206</td>
</tr>
</tbody>
</table>

  There is an issue with K & Q, which have the same MW!
Determination of primary structure

THREE basic ways to know the primary structure:

- **CHEMICAL**
- **PHYSICAL**
- **BIOINFORMATICAL**

We just went through the CHEMICAL and PHYSICAL. The BIOINFORMATICAL method requires information from chemical or physical, but only a limited amount of sequence.

- Example: a sequence of 6 AA is only possible as one of 20\(^6\) possible hexa-peptide sequences (1 of 64 \(\times\) 10\(^6\)).
- There are no more than 50,000 protein-coding genes with \(\leq\) 400 AA on average. This is \(~20 \times 10^6\) possible unique sequences.
- So, a hexamer is not likely to appear more than once.
- Once you have at least 6 AA sequence, you can compare that to all possible proteins encoded in the entirety of the gene sequences (genome) for a species for which the genome is known using appropriate bioinformatic tools. This will then give you the entire protein sequence.

There is one remaining issue: Where are the Disulfides, if any?

…….This requires chemical and/or physical methods

---

Determination of primary structure

THREE basic ways to know the primary structure:

- **CHEMICAL**
- **PHYSICAL**
- **BIOINFORMATICAL**

We just went through the CHEMICAL and PHYSICAL. The BIOINFORMATICAL method requires information from chemical or physical, but only a limited amount of sequence.

- Example: a sequence of 6 AA is only possible as one of 20\(^6\) possible hexa-peptide sequences (1 of 64 \(\times\) 10\(^6\)).
- There are no more than 50,000 protein-coding genes with \(\leq\) 400 AA on average. This is \(~20 \times 10^6\) possible unique sequences.
- So, a hexamer is not likely to appear more than once.
- Once you have at least 6 AA sequence, you can compare that to all possible proteins encoded in the entirety of the gene sequences (genome) for a species for which the genome is known using appropriate bioinformatic tools. This will then give you the entire protein sequence.

There is one remaining issue: Where are the Disulfides, if any?

…….This requires chemical and/or physical methods
Determine which Cys are in Disulfide bonds

Recall that to separate polypeptides for sequence analysis, the disruption of \(-\text{S-S-}\) bonds is needed. And, that prevention of their re-oxidation can be done by alkylation.

\[
\text{Cys} \rightarrow \text{CH}_2\text{SH} + \text{ICH}_2\text{COO}^- \rightarrow \text{Cys} \rightarrow \text{CH}_2\text{S} - \text{CH}_2\text{COO}^- + \text{HI}
\]

Cysteine  | Iodoacetate  | S-Carboxymethylcysteine (CM-Cys)
---|---|---
\(\text{HCOOH}\)  | Paraformaldehyde  | Sulfonic acids

Prevention of their re-oxidation can be done by peroxidation as well.
Determine which Cys are in Disulfide bonds

We change the protection step for sulfhydryl groups.

- Cleave/Protect (reduction/alkylation or oxidation) AFTER fragmentation
- Separate fragments as before, but any linked by \(-S-S-\) will not separate and remain together (e.g., orange peptide).
- THEN break \(-S-S-\) bonds, and re-separate.

Determine the sequence of those peptides that fall off the diagonal by either Edman degradation or tandem MS/MS.

Technique is called "2D-diagonal electrophoresis."
Determine which Cys are in Disulfide bonds

We change the protection step for sulfhydryl groups.

- Cleave/Protect (reduction/alkylation or oxidation) AFTER fragmentation
- Separate fragments as before, but any linked by –S–S– will not separate and remain together (e.g., orange peptide).
- THEN break –S–S– bonds, and re-separate.

Determine the sequence of those peptides that fall off the diagonal by either Edman degradation or tandem MS/MS.

Technique is called "2D-diagonal electrophoresis."
4 levels of protein structure

In order to understand these levels of structure, you need to understand the nature of the polymer first.

In other words, the linkage or PEPTIDE BOND

Protein Structure-Secondary

The 4 S’s for secondary structure:

<table>
<thead>
<tr>
<th>S</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>dependent on number of amino acids</td>
</tr>
<tr>
<td>Solubility</td>
<td>dependent on AA composition and shape</td>
</tr>
<tr>
<td>Stability</td>
<td>complex and not well understood</td>
</tr>
<tr>
<td>Shape</td>
<td></td>
</tr>
</tbody>
</table>

Why is there Secondary Structure?
Protein Structure-Secondary

The Levinthal Paradox (1969):

Theoretical calculation:
Consider just the \(\alpha\)-carbon backbone.....
If there are 4 clearly different angles allowed of all the angles at the
\(\alpha\)-carbon (\(\phi\) and \(\psi\)), then each residue has 2x4=8 degrees of
freedom.
For a protein of 100 residues, there are \(8^{100}\) possible conformations
to “test” for optimal energetics
\[8^{100} = 2 \times 10^{90}\] different conformations
At 1000 billion “tests” per second (1/psec), this is \(2 \times 10^{78}\) seconds
to find the best.
\[\Leftrightarrow 7 \times 10^{70}\] years
Well ...... The age of the universe is \(14 \times 10^{9}\) years

The shortcut proteins use to fold is the use of 2\(^\circ\) structure where most of these degrees of freedom
are prescribed by a regular structure.

What are these “regular structures?”