3.5: Protein Sequencing, Peptide Mapping, Synthetic Genes

Historically, some important disease states were identified as being caused by the lack of an important protein, or the presence of a dysfunctional mutated form of a protein.

- For example, diabetes, types of dwarfism and hemophilia were found to be due to deficiencies in insulin, growth hormone and clotting factor VIII, respectively.

These diseases could be treated by injecting supplemental doses of purified, or partially purified, preparations of these proteins.

- These proteins were isolated from natural materials, e.g. pig (insulin), human cadaver pituitaries (human growth hormone) or blood fractions pooled from normal donors (factor VIII).
- In most cases, even if the protein was found in relatively abundant supply, the cost of production was substantial.

More often than not, interesting bioactive properties were associated with proteins which could be isolated only in minute quantities (e.g. the blood clot dissolving protein tissue plasminogen activator).

Also, non-human proteins typically elicited an immune response when injected into humans, thus the human form of a protein was the only useful form.

- If the protein were not readily available from blood, or urine, it would prove impractical to obtain adequate starting material for production.
- Unfortunately, if the material were derived from human sources, the possibility existed for the spread of human disease (e.g. hepatitis and the AIDS virus).

If the genetic information for these proteins could be isolated, and then transcribed and translated in an easily scaleable biological system, potentially large amounts of protein could be obtained - and hopefully, relatively cheaply.
With the development of "molecular biology", i.e.

• the structure of DNA,
• the elucidation of the genetic code,
• the identification of transcriptional promoters and ribosome binding sites,
• the isolation of restriction endonucleases,
• the identification of the origin of DNA replication
• the development of plasmids with selectable markers, and
• the culturing of \textit{E. coli},

the possibility existed in the mid 1970's to put it all together and produce relatively large amounts of any human protein for therapeutic use.

\textbf{How would you go about the process of producing large amounts of some important human protein? (i.e. protein purification)}

The starting point is typically an \textbf{assay for a functionality of interest}. For example, we may have a hemophiliac whose blood does not clot. However, we find that if we take a sample of his blood and add to it a small amount of blood from a "normal" individual, the hemophiliac's blood will now clot. This will be the basis for our \textit{assay}.

Using this assay, we will fractionate normal blood using various means - chemical precipitation (with ethanol, or ammonium sulfate), and then various liquid chromatography steps, etc.

• Along the way we will follow where our clotting activity is going.
• Hopefully, at some point we will be unable to fractionate it further and will have a \textbf{pure protein}.

Once we have a pure protein we can begin to characterize it with regard to its \textbf{amino acid sequence}. From there we can ultimately get the gene for the protein and express it.

\begin{center}
\textbf{Figure 3.5.1: Protein production}
\end{center}
N-terminal peptide sequence analysis

Polypeptides can be sequenced from their amino-terminus by automated procedures based upon the Edman degradation reaction:

Figure 3.5.2: Edman degradation

- Note that with Edman chemistry only the N-terminal residue is attacked and removed, the rest of the polypeptide remains intact after the reaction.
- The new amino terminal group (previously the second amino acid in the polypeptide chain) is now available for another round of reactions. Thus, the method can be automated.
- The amino acid side chain of the phenylthiohydantoin derivative can be identified using liquid chromatography. Modern amino acid sequencers can probably sequence on the order of two to three dozen cycles (amino acids) of a polypeptide.
- Note that the reaction requires a free amino group on the N-terminal of the protein. If the amino-terminal residue is methylated or formylated then the reaction will not proceed (and the polypeptide is said to have a “blocked” N-terminal).

C-terminal peptide sequence analysis

C-terminal peptide sequence analysis is not as well developed as amino terminal analysis.

- The method usually makes use of non-specific carboxypeptidases.
- Carboxypeptidases will sequentially hydrolyze polypeptides from the carboxy-terminus end. The released amino acid can be identified using liquid chromatographic methods, and the remaining polypeptide is available for further reactions.
- Various carboxypeptidases are available, usually they are not entirely non-specific (i.e. they have certain preferences):

<table>
<thead>
<tr>
<th>Name</th>
<th>Source</th>
<th>Specificity</th>
</tr>
</thead>
</table>

Sometimes the choice of which carboxypeptidase to use is based upon the expected sequence information. In these types of experiments:

1. samples are taken at different time points during the digestion
2. free amino acids are separated from polypeptides
3. the released amino acids are identified via amino acid analysis (liquid chromatography).

*C-terminal analysis is usually only accurately for identification of the last half-dozen residues or so in a polypeptide.*

---

**Peptide Mapping**

One of the obvious problems with protein sequencing is that even if the N-terminal is not "blocked" only limited sequence information can be obtained from an intact polypeptide (i.e. only about two dozen from the N-terminal and half a dozen from the C-terminal).

*How can sequence information for the entire polypeptide be obtained?*

One method is that of **peptide mapping**. Peptide mapping makes use of **proteolytic cleavages** of the polypeptide to produce smaller polypeptides. These smaller polypeptides can then be isolated from one another and subject to sequence analysis.

*How do we order the different sequences which we obtain?*

One of the easiest ways is to repeat the experiment, but with a protease with a different specificity, and in this way obtain **overlapping sequence information**.
Overlapping sequence information can allow you to align the peptides in the correct order and determine the sequence of the original large polypeptide (i.e. protein).

One problem which can arise deals with Cysteine residues and the nature of any covalent disulfide bridges in the protein.

- Any "peptide" mobilities (on either liquid chromatographic or PAGE analyses) which split into two smaller peptides after treatment with a reducing agent (such as b-ME) indicate the presence of a cysteine mediated disulfide bond.
- Upon sequencing these peptides should each contain a cysteine residue. If each peptide has only one cysteine then the disulfide bond assignment is unambiguous.

\[
\begin{align*}
\text{Trypsin cleavage products:} & \quad R \quad E \quad R \quad C \\
\text{V8 Protease cleavage products:} & \quad E \quad C \\
\end{align*}
\]

\[
\begin{align*}
\text{Trypsin cleavage products (sorted by size):} & \quad N \quad C \quad C \\
\text{V8 Protease cleavage products (sorted by size):} & \quad N \quad C \\
\end{align*}
\]

\[
\begin{align*}
\text{Trypsin cleavage products:} & \quad N \quad C \quad R \quad E \quad C \quad R \\
\text{V8 Protease cleavage products:} & \quad C \quad C \\
\end{align*}
\]

\[
\begin{align*}
\text{Trypsin cleavage products (sorted by size):} & \quad N \quad N \\
\text{V8 Protease cleavage products (sorted by size):} & \quad N \\
\end{align*}
\]

\[
\begin{align*}
\text{Trypsin cleavage products:} & \quad N \quad C \\
\text{V8 Protease cleavage products:} & \quad N \\
\end{align*}
\]
Corresponding genetic information

Once we have partial, or complete, peptide sequence information we can begin to identify and isolate the corresponding genetic information. This is the main goal. Once we have the corresponding genetic information it may be possible to produce relatively large amounts of the desired polypeptide.

Back translation

Since we know the genetic code, we can back translate any polypeptide sequence into a corresponding genetic sequence.

- Thus, from the amino acid sequence we could synthesize an artificial gene which would code for the protein of interest.
- Since many amino acids are coded for by more than one codon, there is potential ambiguity with regard to the original exact genetic sequence.

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Number of Codons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Met, Trp</td>
<td>1</td>
</tr>
<tr>
<td>Phe, Tyr, His, Gln, Asn, Lys, Asp, Glu, Cys</td>
<td>2</td>
</tr>
<tr>
<td>Ile</td>
<td>3</td>
</tr>
<tr>
<td>Val, Pro, Thr, Ala, Gly</td>
<td>4</td>
</tr>
<tr>
<td>Leu, Arg, Ser</td>
<td>6</td>
</tr>
</tbody>
</table>

However, making sure we back translate in such a way as to faithfully duplicate the original genetic sequence may not be critical - a correct protein sequence is the overall goal.

In fact, if we are attempting to express the protein in another organism (say expressing a mammalian gene in a bacterial system) we may actually prefer to choose a codon bias appropriate for the expression host organism.

Synthetic genes for small proteins are a reasonable way to proceed; this is one way in which human insulin has been expressed in bacterial systems.

- However, automated synthesis of DNA oligonucleotides is practical for polymer lengths of approximately 60-90 bases or less (about 20-30 amino acids).
Furthermore, the method of construction of synthetic genes typically calls for overlapping complementary oligonucleotides (to be ligated into a single duplex DNA gene “cassette”).

Thus, many oligonucleotides are required for even a single small synthetic gene.

**Figure 3.5.5: Synthetic gene construction**

One way to improve upon the above method of synthetic gene construction is with a *direct PCR* approach. This method does not utilize ligase, or even oligonucleotides that butt together. Instead, with this method many (~100) different overlapping oligonucleotides are simultaneously used in a PCR reaction. Their sequence complementarity can be represented as follows:

The entire set of oligonucleotides may not line up to give the entire gene, but that is alright. We will do multiple rounds of PCR with the idea that some complementary oligo's will anneal and be extended and will lead, bit by bit, to construction of a contiguous synthetic gene:

On the next PCR cycle, some of these extended fragments will anneal with others:

These will be extended via the PCR and can go on to anneal with other larger PCR fragments. Eventually, the entire gene will be constructed. However, since the efficiency of construction of the full-length gene is probably not going to be very good, we need to conduct a subsequent PCR experiment to amplify the full-length gene (using outer primers). The principle features of this method are summarized as follows:

- Many (as many as 1-2 hundred) overlapping oligo's are combined in a single PCR reaction
- The oligos are designed to be as long as possible (~100mers) with limited overlap (~20 bases)
• The full-length gene is constructed in an initial (low yield) PCR experiment
• This full length gene is amplified with a subsequent typical PCR experiment using outer primers.