

1.22: Proteins

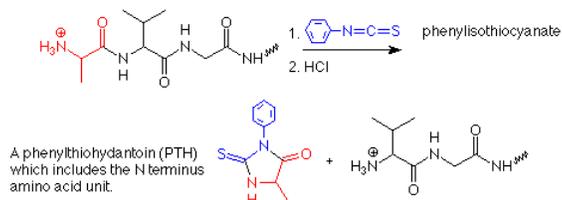
Sequencing a Peptide

Last time we looked at the structural characteristics of amino acids and the peptide bond which joins individual amino acids together to make proteins and peptides. We also learned about the sequence (order) in which amino acid units are joined in peptides. Today we'll study the ways in which the specific sequence of a peptide may be discovered and the methods which are used to synthesize such a peptide.

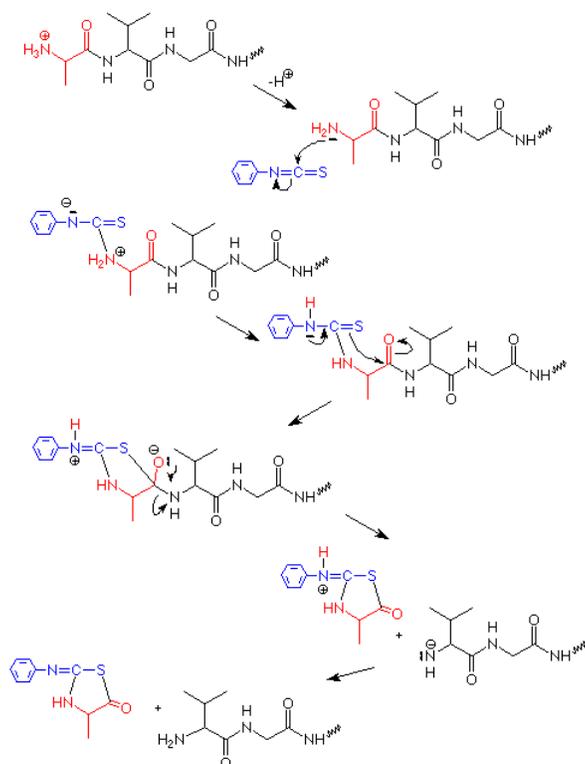
The first thing that is done in determining the sequence of a peptide is to find out which amino acids are present and in what ratios. This is much like beginning the process of determining the structure of an organic compound by determining the ratios of atoms such as carbon, hydrogen and oxygen. This is done by hydrolyzing the peptide bonds which hold the peptide together using HCl as an acid catalyst. (The mechanism is very much like acid catalyzed ester hydrolysis.)

This "amino acid analysis" tells us what the building blocks are in the peptide, but it tells us nothing about their sequence, the order in which they are joined. This information is lost when the peptide bonds which preserve that sequence are hydrolyzed. Even with as few as two amino acids, there are two possible sequences. Consider a dipeptide which amino acid analysis gives us **gly** and **ala**. Either of these could be the N terminus, so the dipeptide could be either **gly-ala** or **ala-gly**. Problem 18.6 in Brown gives you some experience with a pentapeptide, and things rapidly get more complex as the number of amino acid units in the peptide increases.

The next step is to determine which amino acids occupy the N terminus and C terminus positions in the peptide. N terminus determination is commonly done by a process called the Edman degradation. The chemistry is outlined as follows:



This reaction can be understood if we look for some analogies that will help us apply the patterns we used in the past. The $-N=C=S$ group resembles a CO_2 ($O=C=O$) molecule in that the carbon atom is connected to two electronegative atoms by a double (sigma and pi) bond. We know from reacting Grignard reagents with CO_2 that the nucleophile attacks the carbon in CO_2 , so we can expect the same type of pattern in the Edman degradation. The nucleophile is the free NH_2 group at the N terminus of the peptide, formed by loss of a proton from the NH_3^+ to some unspecified base. As we have seen with other reactions of NH_2 groups, this step is followed by a proton shift.



The product of the addition of N and H to the C=N double bond has a nucleophilic sulfur atom located just in reach of the carbonyl carbon at the other end of the N terminal amino acid. Attack of this sulfur at that carbonyl group is followed by departure of the NH group of the next amino acid. This cleaves the peptide bond between the N terminal amino acid and the next amino acid. Further reshuffling of protons yields an isomer of the phenylthiohydantoin. This isomer is converted to the phenylthiohydantoin during the treatment with HCl and the phenylthiohydantoin is identified. Since the phenylthiohydantoin includes the R group of the N terminal amino acid, identification of the phenylthiohydantoin also identifies the N terminal amino acid.

The other product of the Edman degradation is also a peptide -- it is the original peptide minus the original N terminal amino acid. It now has a new N terminal amino acid, which was adjacent to the N terminal amino acid in the original peptide. The new peptide can also be subjected to Edman degradation. When this is done, we learn the identity of the second amino acid (from the N terminal end) of the original peptide and obtain again a peptide which is now two amino acids shorter than the original. In principle, repetition of this sequence would allow us to run successive Edman degradations, clipping off an N terminal amino acid with each degradation, and thus learn the entire sequence of of a peptide or protein. In practice, such a process is practical only for about 20 to 40 amino acids.

Since the laboratory steps in an Edman degradation are very repetitive -- the same thing is done at each cycle, it has been possible to automate this process. Computer controlled "protein sequencers" are common in biochemistry laboratories.

Overlapping Sequences

When peptides and proteins larger than 20 to 40 amino acid units are to be sequenced, they are first broken into smaller fragments either by chemical or enzymatic partial hydrolysis. Here's an example of how partial chemical hydrolysis might be used.

Angiotensin II

Amino acid analysis gives Arg, Asp, His, Ile, Phe, Pro, Tyr, Val.

Partial hydrolysis gives four fragments, which are sequenced by several cycles of Edman degradation as:

Asp-Arg-Val
Ile-His-Pro
Pro-Phe
Arg-Val-Tyr
Val-Tyr-Ile

The structure of Angiotensin II is then pieced together by looking for overlaps in these sequences. For example, we notice that the Arg-Val sequence occurs in two fragments, so we line those two fragments up with the Arg-Val pair aligned:

Asp-Arg-Val
Arg-Val-Tyr

Similarly, the Val-Tyr pair occurs twice, so we add it to the alignment.

Asp-Arg-Val
Arg-Val-Tyr
Val-Tyr-Ile

If we continue to look for these "overlaps," we can assemble the whole sequence.

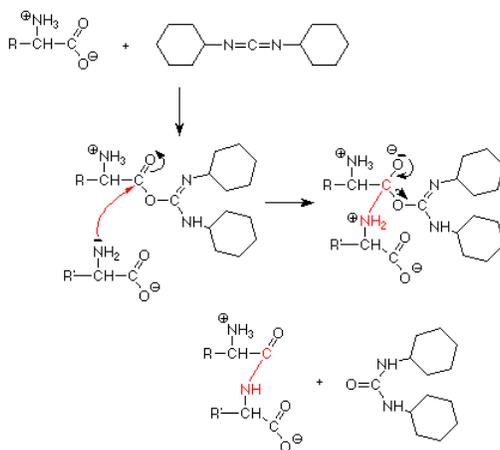
Asp-Arg-Val
Arg-Val-Tyr
Val-Tyr-Ile
Ile-His-Pro
Pro-Phe
Asp-Arg-Val-Tyr-Ile-His-Pro-Phe

Of course, angiotensin II (a peptide involved in blood pressure regulation) is small enough that a protein sequencer could give its sequence directly, but the example illustrates the way fragments may be overlapped to give a complete sequence for a larger protein. The complete sequence of a protein is called its *primary structure*

Peptide Synthesis

When a sequence has been obtained for a peptide, attention can be turned to its synthesis. There are two issues to resolve in synthesizing a peptide. One is to develop a method for making the peptide bond which does not damage anything else in the peptide. This is called "coupling" the two amino acids. The other is to be sure that the amino acids are added to the peptide in the proper sequence.

The key to the first issue is to convert the O^- an amino acid's carboxylate group into a better leaving group. We've seen something similar when we've converted an OH into a Cl as we made the reactive acyl chlorides. This is done by the use of a reagent called dicyclohexyl carbodiimide, or DCC for short. DCC works by bonding to the O. and converting it to a good leaving group -- good because it has many electronegative atoms which can help stabilize the negative charge as it leaves. An amino acid which has a good leaving group is said to be "activated."

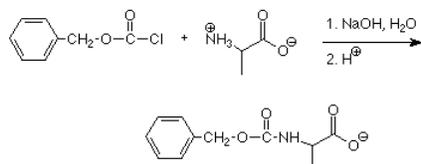


The actual coupling reaction occurs when the amino group of the amino acid "to the right" in the sequence attacks the carbonyl carbon of the "activated" amino acid and the DCC leaves with the oxygen it is bonded to. As usual, there are some proton shifts needed to tidy things up.

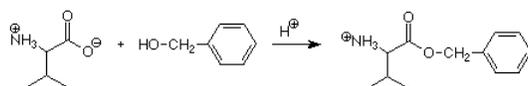
The second issue, adding amino acids in desired sequence, can be illustrated by considering the synthesis of a dipeptide such as Ala-Gly. If we simply mix equal quantities of glycine and alanine and run a DCC coupling reaction, we will get glycines reacting

with glycines to give Gly-Gly, alanines reacting with alanines to give Ala-Ala, and glycines reacting with alanines in two ways to give Ala-Gly and Gly-Ala. This is a mess and it would be better to develop a more specific approach.

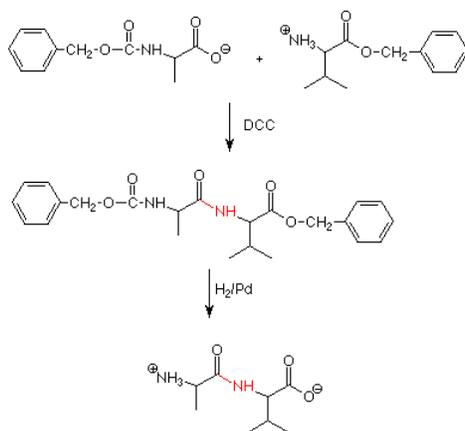
To do this we need to arrange things so that only one of the two carboxyl groups and only one of the two amino groups are free to engage in the coupling reaction. This is done by the use of protecting groups. If we make an amino group into an amide, it is much less reactive as a nucleophile. This can be done by reaction with an acid chloride.



Carboxyl groups are normally protected by conversion to a benzyl ester. This reaction is a Fischer esterification



After coupling, the protecting groups can be removed by hydrogenation.



This is a specific reaction in which only the bonds between the benzyl groups and the oxygens are broken, so the amide bond we have just made by coupling is not affected. (The carbonyl group on the formerly protected amino group is lost as CO₂.) More sophisticated protection schemes in which either the N protection or the O protection can be selectively removed have been developed.

In practice, proteins are now synthesized by molecular biological techniques in which the gene which encodes the sequence of amino acids is isolated and used to direct the synthesis of the protein by a bacterium or a yeast.

Secondary, Tertiary and Quaternary Structure

When we are thinking of a peptide's sequence it is convenient to think of it as a chain which is stretched out, peptides are more commonly coiled (alpha helix, Brown p 518) or folded (beta sheet, Brown p 519). These shapes are called a peptide or protein's *secondary structure* and they are held in place primarily by hydrogen bonds. Recall that hydrogen bonds are much weaker than covalent bonds, but strong enough to resist rupture by mild temperatures. Hydrogen bonds are attractive interactions between the positive end of dipoles like the N-H and O-H bonds and negatively charged locations such as the unshared electron pairs on atoms like oxygen or nitrogen. In peptides it is commonly the N-H bonds of an amine and the oxygens of the carbonyl groups which participate in hydrogen bonds.

Regions of alpha helix or beta sheet are often combined by further folding patterns which make up a protein's *tertiary structure*. Structural proteins such as keratin or fibroin often have large regions of alpha-helix and make fibers. Enzymes are more often globular with hydrophilic amino acids on the outside and hydrophobic amino acids folded in towards the middle.

Individual proteins are often combined into clusters, which may include non-protein molecules such as heme (Fig 18.14, p 522 in Brown). Often such combinations are necessary so that the protein can carry out its biological function. Such clusters constitute the

protein's *quaternary structure*.

Experiments have established that a protein's primary structure is enough, by itself, to determine how it will fold and combine with other proteins to make the appropriate secondary, tertiary and quaternary structures. It is not clear how this happens, and this is an area of active study.

This page titled [1.22: Proteins](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by [Kirk McMichael](#).