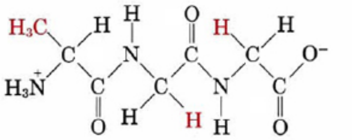
20.14: Primary Protein Structure

Proteins which occur in nature differ from each other primarily because their side chains are different. Partly this is a matter of composition. In wool, for example, 11 percent of the side chains are cysteine, while no cysteine occurs in silk at all. To a much larger extent though, the differences between different proteins is a matter of the sequence in which the different side chains occur. This is especially true of globular proteins like enzymes. The sequence of side chains along the backbone of peptide bonds in a polypeptide is said to constitute its **primary structure**.

The 20 different amino acids permit construction of a tremendous variety of primary structures. Consider, for example, how many tripeptides similar to that shown in Eq. 1 on the page on polypeptide chains can be constructed from the 20 amino acids. In the example shown, the first amino acid in the chain is glycine, but it might just as well be proline or any other of the 20 amino acids. Thus there are 20 possibilities for the first place in the chain. Similarly there are 20 possibilities for the second place in the chain, making a total of 20 × 20 = 400 possible combinations. For each of these 400 structures we can again choose from among 20 amino acids for the third place in the chain, giving a grand total of 400 × 20 = 20^3 = 8000 possible structures for the tripeptide.

A general formula for the number of primary structures for a polypeptide containing *n* amino acid units is 20^n—a very large number indeed when you consider that most proteins contain at least 50 amino acid residues. [$20^{50} = (2 \times 10)^{50} = 2^{50} \times 10^{50} = 10^{15} \times 10^{50} = 10^{65}$] Primary structure is conventionally specified by writing the three-letter abbreviations for each amino acid, starting with the —NH_3^+ end of the polymer. In some cases, this is even shortened to the 1-letter abbreviation sequence. For example, the structure



which reading from the -NH_3^+ end is alanine, glycine, glycine would be specified as

Ala-Gly-Gly or **AGG**

> Note that Ala-Gly-Gly is not the same as Gly-Gly-Ala. In the latter case glycine rather than alanine has the free —NH_3^+ group. Because its ends are different, there is a directional character in the polypeptide chain.

Both three letter and single letter abbreviations are shown in Figure 20.14.1:

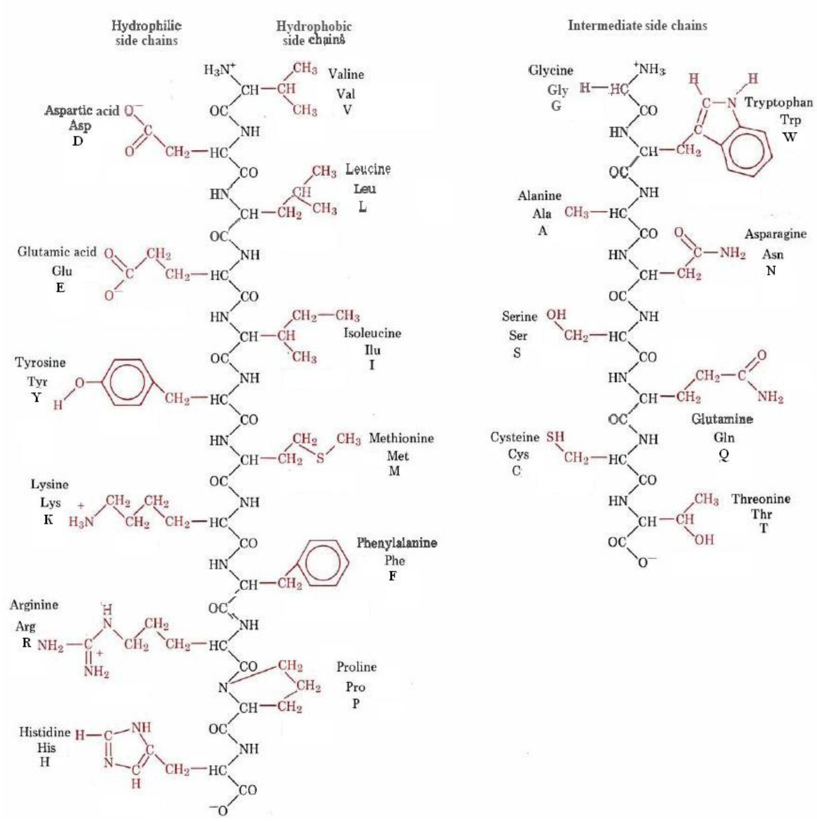
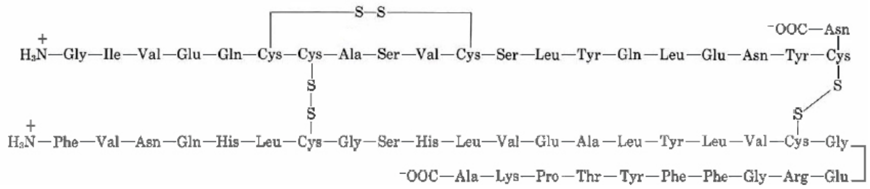


Figure 20.14.1: Structures of the 20 common amino acid side chains in proteins at pH 7. The three letter and single letter abbreviation for each is also included.

Determination of the primary structure of a protein is a difficult and complicated problem. It also is a rather important one—the sequence of amino acids governs the three-dimensional shape and ultimately the biological function of the protein. Consequently much effort has gone into methods by which primary structure can be elucidated.

Insulin was the first protein whose amino acid sequence was determined. This pioneering work, completed in 1953 after some 10 years of effort, earned a Nobel Prize for British biochemist Frederick Sanger (born 1918). He found the primary structure to be


H3N—Gly—Ile—Val—Glu—Gln—Cys—Cys—Ala—Ser—Val—Cys—Ser—Leu—Tyr—Gln—Leu—Glu—Asn—Tyr—Cys
⁻OOC—Asn
S—S
H3N—Phe—Val—Asn—Gln—His—Leu—Cys—Gly—Ser—His—Leu—Val—Glu—Ala—Leu—Tyr—Leu—Val—Cys—Gly
⁻OOC—Ala—Lys—Pro—Thr—Tyr—Phe—Phe—Gly—Arg—Glu

> Note how there are two chains in this structure, one with 21 side chains and the other with 30. These two chains are linked in two places with disulfide (—S—S—) bridges, each connecting two cysteine residues in different chains.

In order to determine the primary structure of a protein, a known mass of pure sample is first boiled in acid or base until it is completely hydrolyzed to individual amino acids. The amino acid mixture is then separated chromatographically and the exact amount of each amino acid determined. In this way, one can find that for every 3 mol serine in the insulin molecule, there are 6 mol leucine. The next step is to break down the protein into smaller fragments. Disulfide bridges are broken by oxidation after which the protein is selectively hydrolyzed by enzymes, called proteases, such as trypsin or chymotrypsin. In a favorable case one will then have several fragments each containing 10 or 20 amino acid residues. These can then be separated and analyzed individually.

Using Edman degradation, so named for Pehr Edman, the sequence of amino acids in one of these polypeptide fragments is usually determined using phenylisothiocyanate, $C_6H_5N{=}C{=}S$, which selectively attacks the —NH_3^+ end of the polypeptide chain. This reaction is carried out under basic conditions. Addition of acid then splits off the terminal amino acid, and it can be identified. Since the rest of the polypeptide chain is left intact, this process can be repeated, and each amino acid in the sequence can be attacked, removed, and identified. By snipping off amino acids one at a time, one eventually finds the complete sequence for the fragment. This whole process can be automated and hence sped up considerably. Once the fragments have been sequenced, it becomes a matter of ordering them correctly. Since different proteases hydrolyze peptide bonds at different places in the amino acid sequence, different fragmentation patterns can be used to determine the sequence for the whole protein. The end of a fragment from a trypsin digest will be in the middle of a fragment from a chymotrypsin digest, for instance. This provides a relatively quick means to sequence unknown protein sequences.

Edman degradation is not the only method for which proteins are sequenced nowadays. Mass spectrometry can sequence polypeptides of 20 to 30 amino acids in length, using a technique called tandem mass spectrometry. In this method, a polypeptide is sent through one mass spectrometer, which ionizes the polypeptide. The charged peptide then enters a collision chamber, causing the peptide to fragment at different peptide bonds. The resulting fragments are then measured by a second mass spectrometer. The resultant spectrum can determine the peptide sequence by differences in mass of the fragments.

With the advent of DNA sequencing methods and the success of the Human Genome Project, many protein sequences are now determined indirectly, through the genetic code. When the DNA sequence for a protein is known, this can be used to determine the protein sequence. By the same token, a known protein sequence can be used to determine the gene coding for that protein. Thus there are many different ways to determine protein sequences, all of which compliment each other.[1]

As methods for determining primary structure have become more advanced, a great many proteins have been sequenced, and some interesting comparisons can be made. A particularly intriguing example is that of cytochrome c, an electron carrier which is found in all organisms that use oxygen for respiration. When samples of cytochrome c from different organisms are compared, it is found that the amino acid sequence is usually different in each case. Moreover, the more widely separated two species are in their macroscopic features, the greater the degree of difference in their protein sequences. When horse cytochrome c is compared with that of yeast, 45 out of 104 residues are different. Only two substitutions are found between chicken and duck, and cytochrome c is identical in the pig, cow, and sheep. The magnitudes of these changes coincide quite well with biological taxonomy based on macroscopically observable differences. Cytochrome c can be used to trace biological evolution from unicellular organisms to today’s diverse species, and even to estimate times at which branching occurred in the family tree of life. This makes molecular methods a powerful tool for the evolutionary biologist as well.

References

1. ↑ Nelson, D.L., Cox, M.M. Lehninger Principles of Biochemistry(5th ed). New York: W.H. Freeman and Company, 2008. pp. 96-100.