C. THE COVALENT STRUCTURE OF PROTEINS

As just about any biochemistry textbook will tell you, the word protein comes from the Greek word "προτειν" which means first. Well, that's not strictly true. So far as we know, RNA preceded protein on the evolutionary scene. Even today, the central dogma of molecular biology places protein last in the chain of information from DNA to RNA to protein. However, if you wanted to pick a class of molecules that were of the first importance in producing life as we know it today, proteins would certainly top the list. Later chapters will focus on how the structures of certain proteins contribute to their functions, but for now we're going to be content to take a bottom up approach in explaining how the structure of proteins can be tied to the structure of their building blocks, the amino acids. There are twenty genetically encoded amino acids[1] that are incorporated into growing protein chains on the **ribosome**, the cell's protein synthesis factory. The diversity of the amino acid building blocks, layered on top of some of their common structural features, is responsible for the diversity in protein structure itself, so that's where we'll start.

The Twenty Amino Acids

Basic Structure

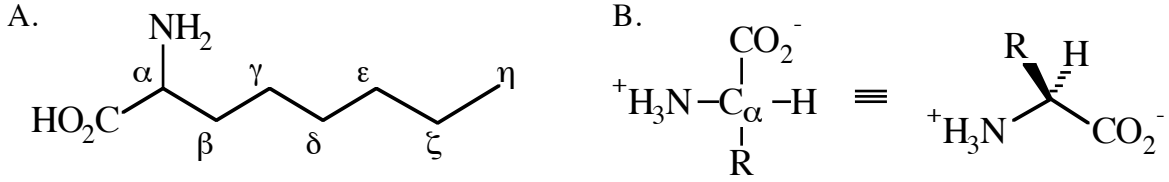

Figure C.1. (A) The alphabetic (Greek) labeling of carbons in an alkanoic acid, showing an amino group at the C_α. (B) A Fischer projection showing the L-configuration of the naturally occurring amino acids, converted to the standard 3D projection on a 2D surface.

The name "amino acid" fundamentally describes the chemical nature of these molecules; each contains an carboxylic acid function and an amino function (Figure C.1A). Furthermore, each of the 20 is an α–amino acid, which refers to the position of substitution of the amino group with respect to the carboxylic acid functionality. As shown in Figure C.1A, each of the carbons in an alkanoic acid is given a label from the Greek alphabet, denoting that carbon's distance from the carboxyl carbon. The α-carbon (C_α) of an amino acid is directly adjacent to the carboxyl group, and is the position of attachment for the α–amino group. Among the twenty, there is an additional "R" group, or side-chain, attached at C_α that renders it a chiral center. So we add on another label, and specify the naturally occurring amino acids as α-L-amino acids. The "L" appellation for these amino acids refers to a specific chiral configuration according to Fischer's nomenclature, which is shown in

[1]Actually, there are 21 if you include selenocysteine, but that's a separate topic entirely.

Figure C.1B.[2] The common chiral configuration of the twenty amino acids is essential for providing regular structural features as we'll see, but the reason for the choice of "L-" is a mystery, if one even exists. Amino acids isolated from carbonaceous meteorites (which are thought to be abiotic in origin) are racemic, so somewhere along the road, a (perhaps fortuitous) choice was made by the first common ancestor of all existing life forms to use this set of stereoisomers.

The structures of the twenty amino acids are shown in Figure C.2. It would be hard to overemphasize the importance of these structures in understanding structural biochemistry. Just as the 26 letters of the alphabet are required to construct the variety of words in the English language, the 20 amino acids provide the fundamental alphabet for protein biochemistry. Included in Figure C.2 are three and one letter codes for each amino acid, which along with the molecular structures, **must be committed to memory**. The categories used here to segregate the twenty by their chemical character is somewhat arbitrary, since in a few instances, one amino acid might be placed in any of a number of the boxes. Tyrosine, for example, which has a 4-hydroxybenzyl side chain is generally described as a non-polar aromatic amino acid - which is in part true. The phenyl ring contains a lot of non-polar surface area which is important to the role of Tyr in many proteins. However, it is also a polar non-charged amino acid, in that the hydroxyl group at Cζ is capable of acting as a hydrogen bond donor or acceptor. And finally, tyrosine might also be classified as an acidic amino acid. The phenolic hydroxyl group has a pK_a of about 10, which means that the side chain could act as an acid under certain conditions. That's why the "classic" categories can be misleading. Often a given side chain may have several different characteristics, of which one or more may contribute to its function in a protein. In the following sections, we'll individually address each of the broad characteristics by which amino acids are judged.

Hydrophilicity/Hydrophobicity

Previously, we looked at electrostatic and entropic contributions to intermolecular complexes. In comparing the basic amino acid structures (Figure C.2), one of the chief means of categorizing the twenty amino acids relies on their principal mode of interactions with other molecules. For example, the character of the side chains of amino acids like serine, asparagine, and aspartic acid is dominated by their polar functional groups: the hydroxyl, amide and carboxylate groups respectively. On the other hand, amino acids like valine, leucine and phenylalanine have non-polar aliphatic and aromatic side chains, which restricts their enthalpic contribution in intermolecular contacts to relatively weak vdW forces, but provides opportunities for strong interactions through the hydrophobic effect. However, most amino acids' side chains have both hydrophobic and hydrophilic character - they are **amphiphilic**. For example, threonine has both a hydrophilic hydroxyl functionality and a hydrophobic methyl group, both in the γ position. Likewise, lysine, which is categorized as a basic amino acid, has four methylene groups in its side chain, which is equivalent in non-polar surface area to amino acids like leucine and methionine. Also note that two

[2]An observant eye will note that Figure 2.1B shows the amino acid in a doubly ionized form, reflecting the predominant protonation states of the basic amine and acidic carboxylic groups at pH 7. This ionization state is refered to as a **zwitterion**, indicating that it's a neutral species, with paired and opposing charges within the molecule. As we'll discuss shortly, the acid-base chemistry of the amino acids is one of the keys to their structural and functional roles.

of the three aromatic amino acids, tyrosine and tryptophan, have polar functionalities combined with large non-polar side chains.

So, can one definitively categorize an amino acid side chain as hydrophilic or hydrophobic? No. Instead, the issue is typically addressed by assessing the *relative* solubilities of the amino acid in water and some non-polar solvent, like octanol or cyclohexane. Table C.1 presents some statistics on the relative contributions of polar and non-polar groups to the total surface area of each amino acid as well as the change in free energy associated with transfer of a side chain analog (where the C_α atom is replaced by a hydrogen atom) from water to cyclohexane.

Table C.1. The total surface area and polar surface area of the sidechains of the twenty amino acids are compared. $\Delta G_{transfer}$ refers to the free energy of transfer from water to cylcohexane. Note that the greater the polar surface area of the sidechain, the more positive the free energy of transfer.

Amino Acid	**Total Surf. Area of sidechain ($Å^2$)**	**Polar Surf. Area in the Sidechain ($Å^2$)**	**$\Delta G_{transfer}$ of sidechain (kcal/mol)**
Alanine	113	-	-0.87
Arginine	241	107	15.93
Asparagine	158	69	5.22
Aspartate	151	58	9.71
Cysteine	140	69	-0.34
Glutamine	189	91	6.51
Glutamate	183	77	7.78
Glycine	85	-	0
Histidine	194	49	5.63
Isoleucine	182	-	-4.00
Leucine	180	-	-4.00
Lysine	211	48	6.52
Methionine	204	43	-1.42
Phenylalanine	218	-	-2.05
Proline	143	-	
Serine	122	36	4.36
Threonine	146	28	3.53
Tryptophan	259	27	-1.40
Tyrosine	229	43	1.09
Valine	160	-	-3.11

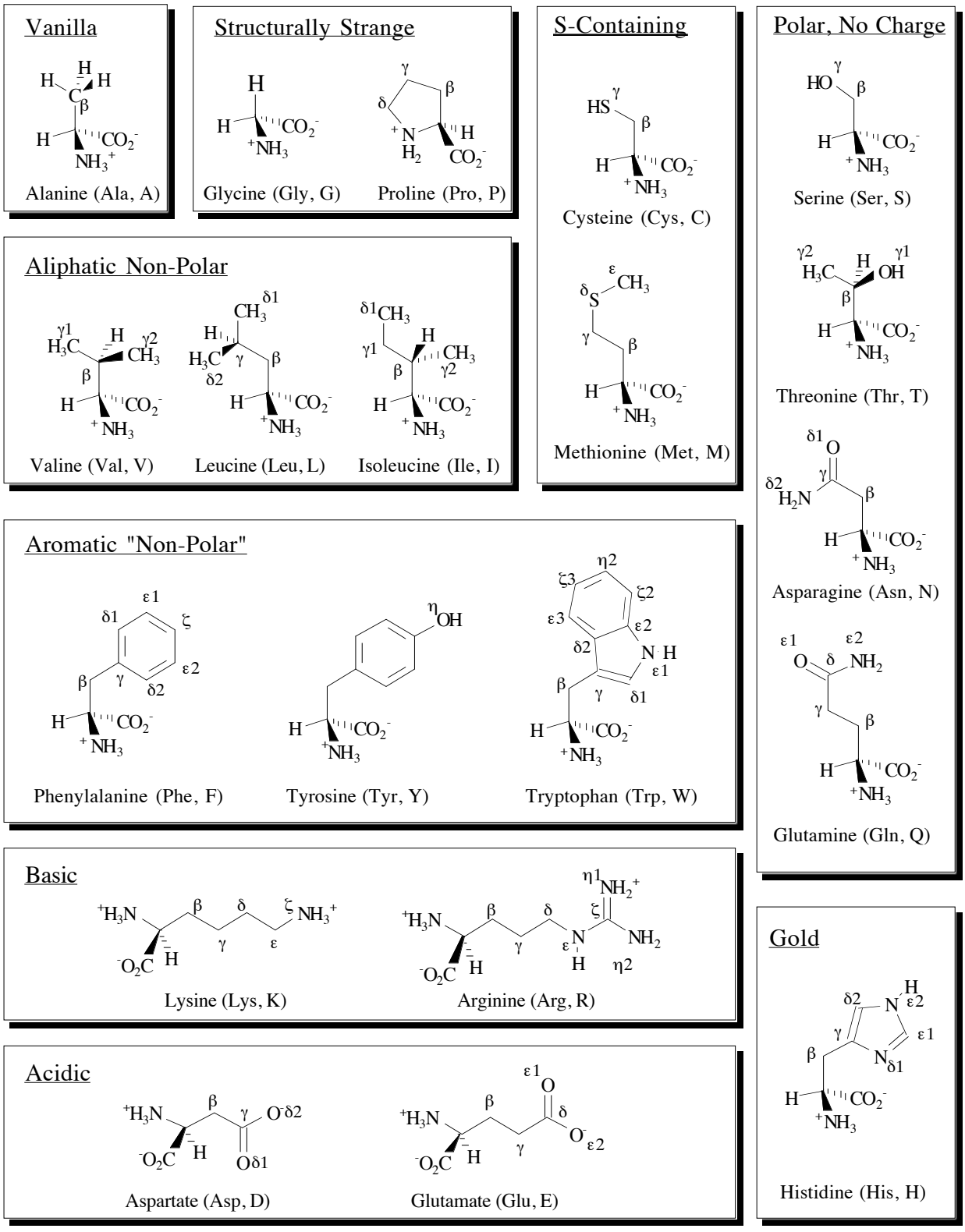

Figure C.2. Structures of the 20 amino acids.

Acid-Base Chemistry

Another of the most fundamental aspects of amino acid chemistry relates to their acid-base chemistry. Equations C.1-C.4 provide a brief review of some fundamental descriptors used in describing the properties of acids and bases.

$$K_a = \frac{[H^+][A^-]}{[HA]} \qquad \text{(Eq. C.1)}$$

$$pK_a = -\log_{10} K_a \qquad \text{(Eq. C.2)}$$

$$pH = -\log_{10}[H^+] \qquad \text{(Eq. C.3)}$$

$$pH = pK_a + \log_{10} \frac{[A^-]}{[HA]} \qquad \text{(Eq. C.4)}$$

Most importantly, the transfer of a proton between donors and acceptors is an equilibrium process (Eq. C.1), and each acid has a defined acid dissociation constant (K_a) that corresponds to the transfer of a proton from the acid (HA) to water. Defining pH and pK_a in equations C.2 and C.3, one obtains the Henderson-Hasselbalch equation (Eq. C.4), which provides a simple relationship between pH and pK_a. Consider the following situation: a weak acid of pK_a 4.0 is dissolved in a buffer at pH 7.0. Solving for the ratio of $[A^-]/[HA]$, one finds that there is a 1000 fold higher concentration of the conjugate base, A^-, in solution than the conjugate acid HA. The higher the pH of the solution, the less of the conjugate acid than will be found relative to the conjugate base. At neutral pH, acids with pK_a's below 7 will be found predominantly in their conjugate base form, and acids with pK_a's above 7 will largely be found in the protonated (conjugate acid) form.

Table C.2 The pK_a's of some functional groups relevant to the acid base chemistry of amino acids.

Molecule	**Functionality**	**pK_a**
Acetic acid	Carboxylic acid	4.7
Methylammonium ion	Ammonium group	10.6
Glycine	α–Ammonium group	9.6
Glycine	α–Carboxylic acid	2.3
Aspartic acid	γ–Carboxylic acid	4.0
Glutamic acid	δ–Carboxylic acid	4.4
Histidine	Imidazolium group	6.8
Cysteine	Sulfhydryl group	8.0
Tyrosine	Phenolic hydroxyl group	10.2
Lysine	ε-Ammonium group	10.7
Arginine	Guanidinium group	12.0

From Figure C.1B, it should be apparent that amino acids readily react as acids and bases. In aqueous solution, only a miniscule fraction of the total number of dissolved molecules in fact exist as "amino acids". Perhaps a more appropriate name might be "ammonium carboxylates". The α-amino group is basic; its conjugate acid, the ammonium group, has a pK_a of 9.6 in free glycine, while the pK_a of the carboxylic acid is C.3. From the Henderson-Hasselbalch equation (Eq. C.4), it can be calculated that at pH 7, less than one percent of the amine group will be neutral, and only one part in 5000 will be in the acid form. Similarly, many of the other amino acid sidechains are ionized at neutral pH (Table C.2). In most instances, the pK_a's of the sidechains dictate that only one protonation state will predominate at pH 7. For example, the "acidic" amino acids, Asp and Glu, are typically found as their conjugate bases, whereas the the basic amino acids, Arg and Lys, typically exist as their conjugate acids.

Histidine, on the other hand, is of interest because its pK_a (6.8) dictates that roughly equal fractions of the side chain will exist in the conjugate acid and base forms simultaneously. This makes histidine a particularly valuable amino acid (hence "gold") in that its conjugate base form[3] represents the strongest base likely to be found in any abundance at neutral pH, while its conjugate acid is similarly the strongest acid to be found at high concentration at pH 7. Thus, while His turns out to be the least common amino acid found in proteins, it proves to be a frequent contributor to their function given its unusual properties. It may be used sparingly, but it contributes importantly to protein function in many instances.

Something that you may have noticed is that, despite a common structure, the carboxylic acid groups of glycine and acetic, aspartic and glutamic acids have a wide range of pK_a values, between 2.3 and 4.7. This is the result of differences in chemical environment. For example, the ammonium group of glycine acts as an electron withdrawing group which decreases the pK_a of the carboxylic acid, just as trifluoroacetic acid is a much stronger acid that plain old acetic acid. (In addition, the ammonium group provides enthalpic stabilization of the carboxylate form of the acid via an ion-ion interaction.) This effect shifts the equilibrium in favor of the carboxylate, yielding a relatively low pK_a of 2.3. Similar arguments can be made for the lower sidechain pK_a of the α–ammonium group of glycine relative to methylammonium ion. It's also worth pointing out that some chemical environments favor the neutral species. For example, the pK_a of acetic acid changes dramatically depending on the solvent in which it is dissolved (Table C.2). In solvents that are less polar than water, the conjugate base, acetate, receives less enthalpic stabilization and so is higher in free energy relative to the acid. Thus, the pK_a of acetic acid is raised in methanol relative to H_2O.

[3]Note that, like carboxylic acids, the imidazole side chain of histidine can exist in two tautomeric forms, in which the N-H bond can be on either the δ or ε ring nitrogen.

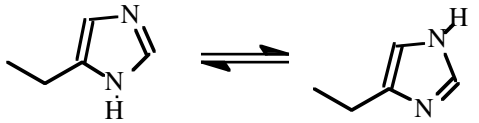

Table C.3 The pK_a of acetic acid in various solvents.

Solvent	**pK_a**
Water	4.7
Methanol	9.6
Dimethylsulfoxide	12.6

All of this is merely a means of warning you that the pK_a of a given functional group is not an absolute. It varies substantially with its immediate chemical environment.

Redox Chemistry

While it is not a predominant feature of amino acid chemistry, redox behavior is associated with the two sulfur-containing amino acids, cysteine and methionine. From a biochemist's point of view, this possibility is more often an obstacle in working with a protein than a feature that reveals inherent protein function (though it is that too). The interior of the cell has a reducing environment, assuring that sulfur containing amino acids will remain in a reduced state. However, when a protein is isolated from cells it is exposed to the oxidizing environment of the atmosphere, which is 18% oxygen, unless precautions are taken. The oxidation of cysteine or methionine can be disastrous for a protein, since the resulting sidechains are more polar than those of the starting materials (Figure C.3). The cell is able to maintain strong reducing conditions by keeping high concentrations of glutathione, a molecule that contains a cysteine sidechain. These free thiol groups can be used to scavenge oxygen, and even to reduce oxidized amino acids. In solution, β-mercaptoethanol and dithiothreitol, which each contain free sulfhydryl groups, are added for the same purpose.

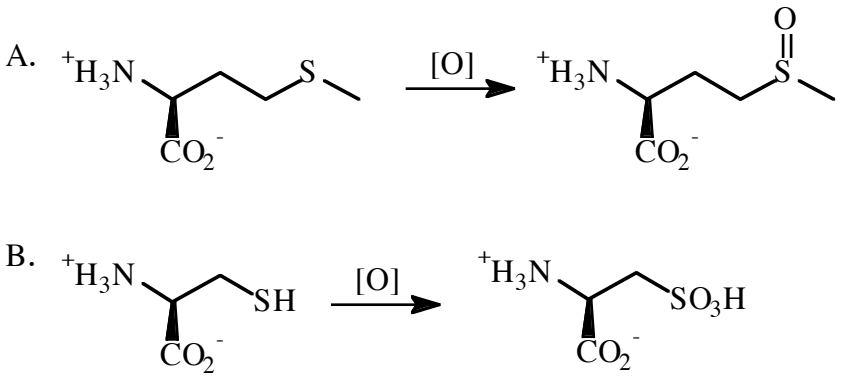

Figure C.3 (A) The oxidation of methionine to its sulfoxide form. (B) The oxidation of the cysteine thiol to a sulfonic acid. Note that at pH 7, the sulfonate form would predominate.

On the other hand, the redox chemistry of cysteine does occasionally play a significant role in protein structure and function, usually through the oxidative formation of a disulfide bridge with another cysteine sidechain. The resulting molecule is called **cystine**, which possesses a covalent linkage between the two sulfur atoms (Figure C.4). Cystine turns out to be an important contributor to a number of extracellular proteins, since it can be used to cross-link different protein molecules together and even to provide crosslinking within a protein. As with acid-base chemistry, redox chemistry is a reversible process with an equilibrium constant that varies depending on local conditions. The oxidizing half-reaction for the production of cystine from cysteine is sufficiently favorable that it can be coupled to the reduction of a variety of other molecules. For example, Hg^{2+} is reduced by cysteine sidechains to produce Hg(l) and cystine.

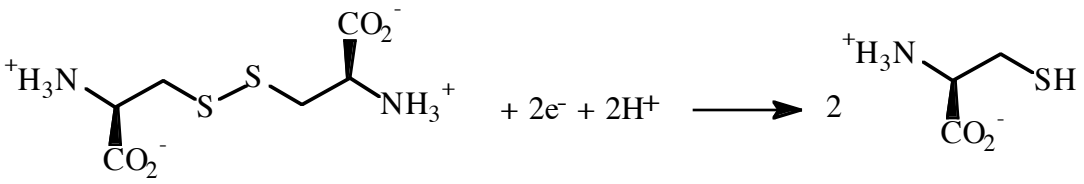

Figure C.4. The half reaction corresponding to the reduction of cystine to cysteine. The standard reduction potential at pH 7 is -0.22 V.

Spectroscopic Properties of the Aromatic Amino Acids

Although the UV spectra of Phe, Tyr and Trp aren't of any significance to the function of amino acids in proteins, they are a useful diagnostic tool to be used in handling proteins. Some spectroscopic data for the three aromatic amino acids are given in Table C.4. Of primary interest are the absorbance data. Tryptophan and tyrosine both absorb strongly near 280 nm, while phenylalanine absorbs more weakly with a maximum near 260 nm. Commonly, protein is identified by its absorbance at 280 nm thanks to the contributions of Trp and Tyr, and estimates of protein concentration can be made based on the absorbtivity of protein solutions. As will be seen in Chapter 3, the fluorescence properties of Trp and Tyr are important in structural studies. Both absorbance and emission spectra can be used to learn about the chemical environment of an amino acid side chain.

Table C.4. Spectroscopic properties of the aromatic amino acids.[a]

Amino Acid	**UV Absorbance**		**Fluorescence Emission**	
	λ_{max}	**ε ($M^{-1}cm^{-1}$)**	**λ_{max}**	**Quant. Yld.**
Phe	257 nm	197	282	0.04
Tyr	275 nm	1420	303	0.21
Trp	280 nm	5600	348	0.20

Covalent Structure in Proteins

Peptide Linkage and Primary Structure

Amino acids are the building blocks of proteins. This simple analogy was used earlier in this chapter to rationalize the study of amino acid chemistry. Now we need to look at how proteins are constructed from amino acids. Proteins are linear polymers with amino acids acting as the monomers that combine to form the chain. The chemical linkage that holds the protein together occurs between the carbonyl carbon of one amino acid and the α-amino group of an adjacent amino acid - an amide bond. In proteins, the amide linkage is refered to as a **peptide bond** (Figure C.5A). In organic chemistry, an amide is formed by the reaction of an amine and an activated carboxyl derivative, such as an ester or an acyl halide. In the cell, the peptide bond is synthesized by the ribosome by reacting an ester of one amino acid with the α-amino group of a second (Figure C.5A). This occurs sequentially, until the full protein is synthesized as per the directions of the DNA sequence encoding the protein (much more on this later). The resulting polymer of amino acid **residues** (note that they are no longer amino acids, since both the amino and acid groups have been lost to other functionalities) is sometimes referred to as a **polypeptide chain**. And since this is a paragraph of jargon, let it be noted that a **peptide** (or oligopeptide) is a term that's used to describe chains of 50 (roughly) or fewer amino acid residues. In addition, a Greek numerical prefix can be used to specify exactly how many residues. For example, a tripeptide has three amino acid residues.

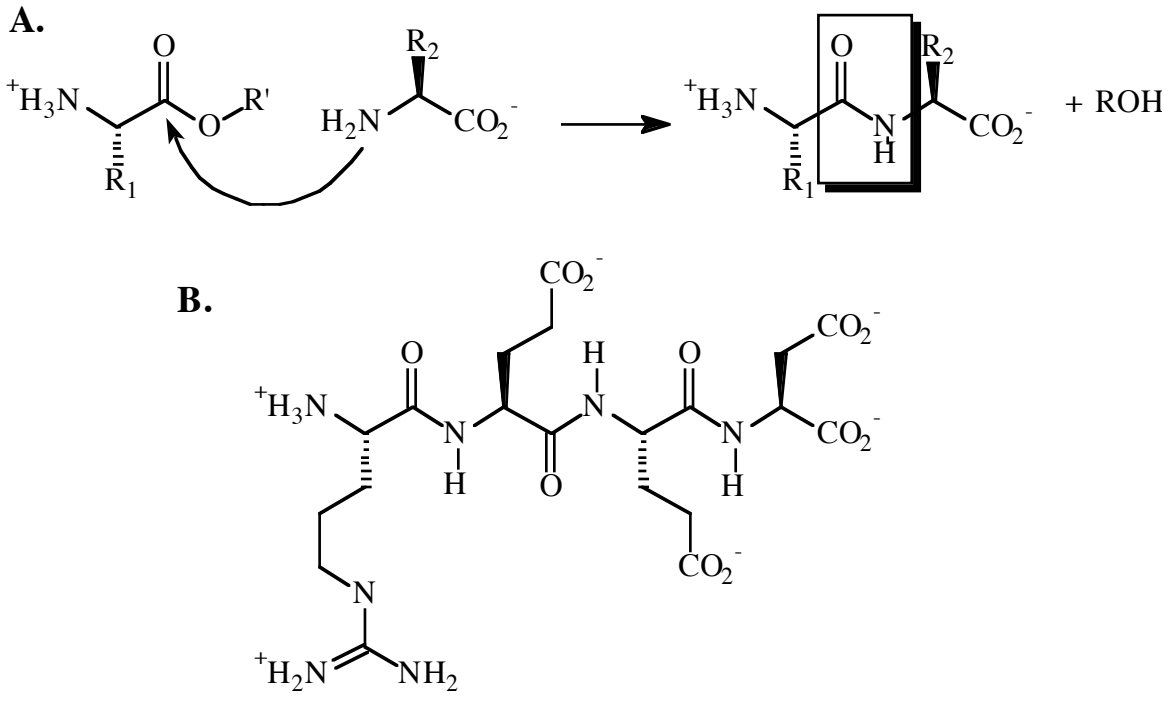

Figure C.5 (A) The α-amino group of one amino acid condensing with an ester of the the α-carboxylate of a second yields a peptide linkage (R' is the 3' hydroxyl of a tRNA molecule – coming later). (B) This tetrapeptide has the sequence REED.

We will frequently talk about the **sequence**, or **primary structure** of a specific peptide or protein. Any given protein, such as lysozyme or the α chain of hemoglobin, is a chemically defined compound that is most conveniently described by providing the names of the amino acid residues in order as they appear in the chain. By convention, the sequence is read from the amino acid residue with a free α-amino group towards the residue that has a free carboxylate - from the **N-terminus** to the **C-terminus**.[4] For example, in Figure C.5B the tetrapeptide's sequence is ArginylGlutamylGlutamylAspartatic acid, abbreviated ArgGluGluAsp, or more succinctly, REED. Their are numerous strategies for determining the sequence of a polypeptide, including genetic and chemical techniques, that won't be covered here, but the information is available in most Biochemistry textbooks. To learn more about the biological and chemical synthesis of peptides, see the appendix.

The Stability and Structure of the Peptide Bond

The integrity of the protein depends upon the stability of the peptide bond. Otherwise, the information contained in the primary structure of the protein would be lost upon hydrolysis of the polypeptide to the thermodynamically more stable amino acids (the carboxylate and ammonium groups are prefered over the peptide bond). As it turns out, the peptide bond is extremely resistant to hydrolysis . It is thermodynamically unstable, but kinetically inert. The change in free energy for hydrolysis of the amide bond is -5 kcal/mol, reflecting the stability of the free carboxylate group in comparison to the amide derivative. Despite this instability, the half life of an amide bond is 7 years at pH 7. Typically, the cell will use catalysts called proteases to degrade a protein long before hydrolysis of the polypeptide has become a problem.

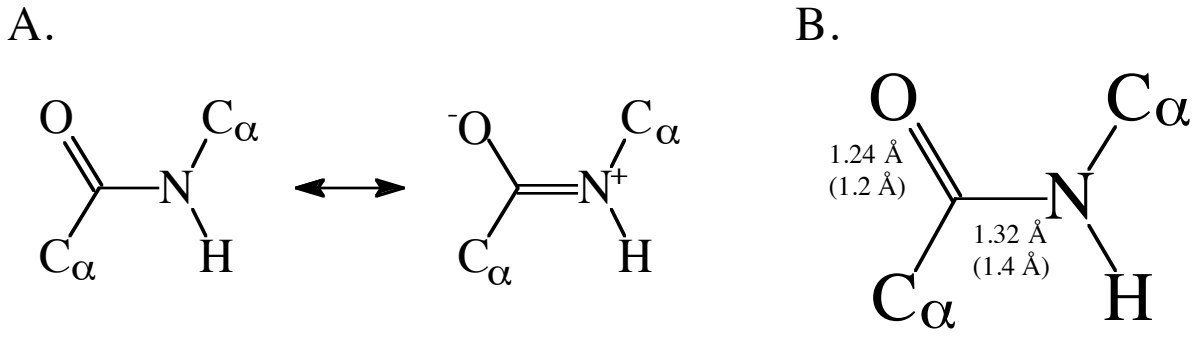

Figure C.6 (A) The resonance structures associated with the peptide bond. The C-N bond has about 40% double bond character. (B) The dimensions of the amide functionality. The carbon oxygen bond is somewhat longer than normal for a carbonyl group (1.24 Å vs. 1.20 Å) and the C-N bond is somewhat shorter than normal (1.32 Å vs. 1.4 Å).

Aside from its unusual resistance to hydrolysis, the amide bond is notable for possessing substantial double bond character between the carbonyl carbon and amide nitrogen. Meanwhile, the bond between the carbonyl carbon and oxygen is unusually long, reflecting weakened double bond

[4]This isn't totally arbitrary. The N-terminal amino acid is the first one to be put in place during protein synthesis on the ribosome and the C-terminal residue is the last to be added before the polypeptide chain leaves the ribosome.

character (Figure C.6). The simple explanation for this phenomenon is that the lone pair on the amide nitrogen can participate in resonance with the carbonyl group, leading to a three center conjugated π system. Two other significant results derive from this phenomenon.

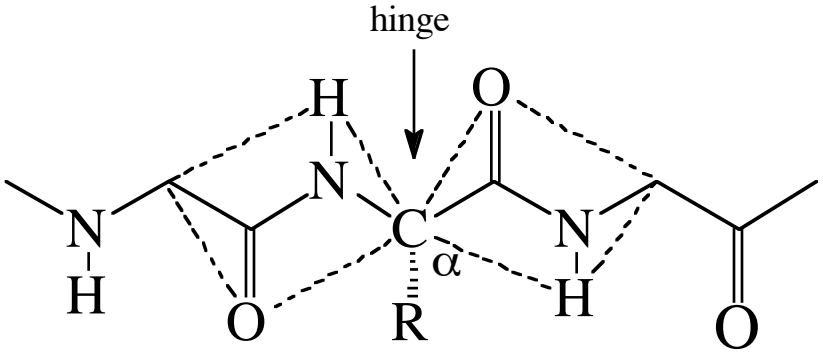

Figure C.7 The dashed lines connect two groups of six atoms flanking the central α-carbon which are held fixed in two "plates" that are hinged at the α-carbon.

The peptide bond is planar. The sp^2 hybridized amide nitrogen fixes its two substituents, the α-carbon of the second residue and a hydrogen, in the same plane as the substituents on the carbonyl carbon - namely the carbonyl oxygen and the α-carbon of the first residue (Figure C.7). This planarity reduces the conformational flexibility at each amino acid's α-carbon (see below), and has been described as creating a chain of plates on hinges. The partial double bond character between the amide nitrogen and carbonyl carbon provides an energetic barrier to the free rotation about the peptide bond of 20 kcal/mol (compared to 90 kcal/mol for a carbon-carbon double bond). Two conformations for the linkage are thus available for a dipeptide. The ***trans*** conformation (Figure C.8A) places the connected α-carbons opposite to each other, while the ***cis*** conformation (Figure C.8B) has the α–carbons placed in close proximity to one another. Because of the increased steric conflict between α-carbons in the *cis* conformation, it is rarely found in proteins, except in peptide linkages where proline, with its tertiary amide, is the C-terminal residue (Figure C.8C). For these peptide bonds, the *trans* isomer is favored by 2 kcal/mol, a ratio of thirty to one.

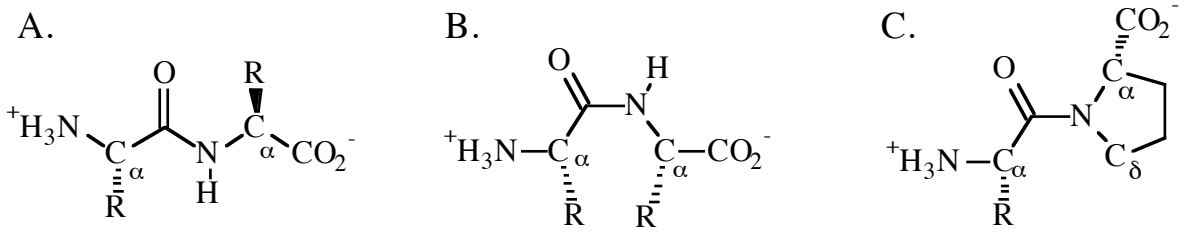

Figure C.8 (A) A *trans*peptide bond, placing the α-carbons 180° apart. (B) A *cis*peptide bond. Note that the C_α's are directly adjacent to one another. (C) A *trans*peptide bond involving proline. Since C_δ of proline is adjacent to the α-carbon of the N-terminal residue, instead of the usual N-H group, the relative stability of the *trans* conformation to the *cis* conformation is diminished with proline C-terminal to a peptide bond.

The peptide bond is polar. The second resonance structure of the amide bond (shown in Figure C.6A) not only contributes to the planarity of the linkage, but also to its polarity. There are unusually large partial charges on the carbonyl oxygen and the hydrogen attached to the amide nitrogen. The dipole associated with this charge difference is 3.5 D (Figure 1.2D). The peptide bond functionalilty therefore makes a large electrostatic contribution to the energy of interactions with other hydrogen bond donors and acceptors.

Disulfide Bonds

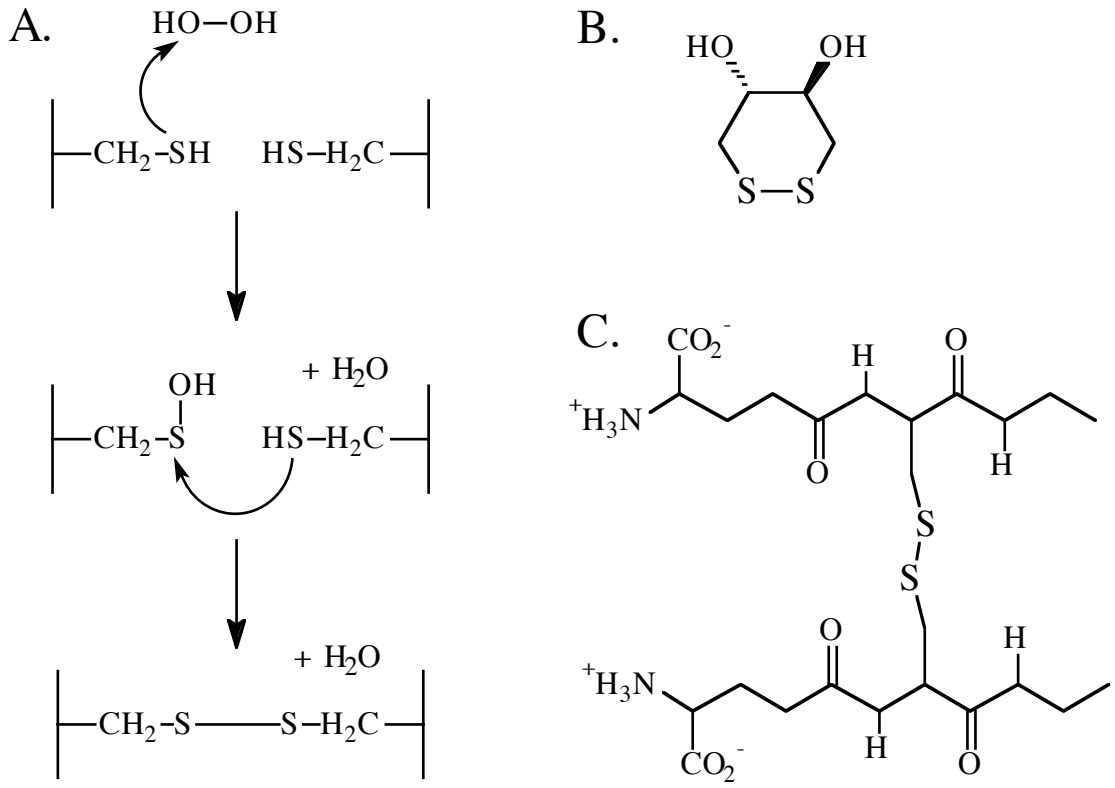

Figure C.9 (A) The oxidation of two cysteines to a cystine disulfide bridge by hydrogen peroxide. (B) The oxidized form of dithiothreitol. (C) Oxidized glutathione. Both dithiothreitol and glutathione can oxidize other thiols to disulfide linkages.

Proteins are linear polymers of amino acids linked by amide bonds. There is, however, one additional source of covalent structure in polypeptides. Two cysteine residues may be oxidized to a cystine disulfide bridge, providing a covalent bond between residues that are distant from each other in sequence. The reducing environment of the cytosol (about -0.27 V in *E. coli*) means that most intracellular proteins will contain only reduced cysteine residues. However, disulfide bridges are particularly important in stabilizing the tertiary structures of small monomeric extracellular proteins, such as lysozyme and ribonuclease, which each contain four disulfides. The "hydrogen atom" of

protein biochemistry - bovine pancreatic trypsin inhibitor, another secreted protein - contains three disulfides in a 60 residue polypeptide chain.

The formation of disulfides between reduced cysteines requires the participation of an oxidizing agent. For example, the OxyR regulatory protein of *E. coli*[5] contains a redox active pair of cysteine residues that are selectively oxidized to cystine by hydrogen peroxide (Figure C.9A). However, the oxidizing agent can be any of a number of species, though other disulfides, such as oxidized dithiothreitol (Figure C.9 B) or oxidized glutathione (Figure C.9C) are most commonly used *in vivo* and *in vitro*, respectively.

Summary

- It is important to memorize the 20 amino acids!
- Amino acids can be classified on several grounds. In particular, we looked at the differing intermolecular contacts that could be made by various sidechains, acid-base chemistry and redox chemistry.
- The covalent structure of proteins is dominated by the planar peptide bond. As a linear polymer, the polypeptide has a specified sequence of amino acid residues that is read from the amino terminus to the carboxy terminus.
- Disulfide bridges between cysteines provide covalent crosslinks between regions of the polypeptide that are distant from each other in sequence.

Further Reading

Thomas E. Creighton (1993) Proteins: Structure and Molecular Properties.

G. A. Petsko & D. Ringe (2004) Protein Structure and Function

[5]Zheng, M., Åslund, F. &Storz, G. (1998). Activation of the OxyR Transcription Factor by Reversible Disulfide Bond Formation. *Science* **279**, 1718-1721.

APPENDIX. WHERE POLYPEPTIDES COME FROM

Biological Synthesis of Polypeptides

For now we will only sketch the mechanism by which proteins are synthesized in the cell. The central dogma of (molecular) biology (Figure CA.2) indicates that the instructions for making a protein of a specified sequence of amino acids lies in the sequence of nucleotides of stretch of DNA. Those instructions are **transcribed** to messenger RNA (mRNA), which possesses the same nucleotide sequence as in the DNA (absent a methyl group on one particular type of nucleotide), and that will in turn be **translated** to protein via the action of the ribosome.

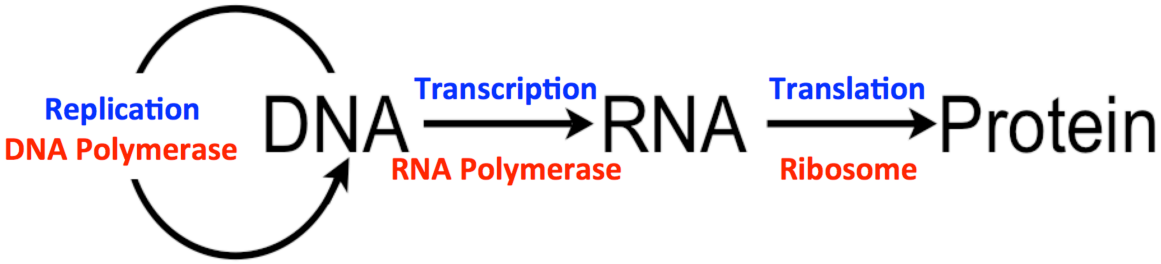

Figure CA.1. The Central Dogma of Molecular Biology. Note that DNA possesses the ability to specify the structure of its copies (via the action of DNA polymerase), while also specifying the structure of RNA in transcription (via the action of RNA polymerase), which in turn specifies the structure of protein via translation (via the action of the ribosome).

The process of translation has the complex task of converting sequence information in one chemical form to another and will be covered in detail later. The basic chemistry, however, is simple. It is the reaction of an amine group on one amino acid with the carboxylic acid (carboxylate) group of a second (Figure CA.2). At pH 7, however, this chemistry is not favorable and proceeds with a change in free energy of -5 kcal/mol. So, as in a round bottom flask, the carboxylate needs to be activated to a more reactive form. Translation actually passes the carboxylate through two reactive species, first a mixed acid anhydride and secondly an ester. The mixed acid anhydride forms from the reaction between ATP and an amino acid, and the ester forms when the amino acyl~AMP mixed acid anhydride reacts with a transfer RNA (tRNA) molecule to form an amino acyl~tRNA ester ("~" is used to denote a high energy chemical linkage). Note that amino acids are added to the growing polypeptide chain and the carboxyl group. As a result the polypeptide chain grows from the first amino acid (which retains a free ammonium group). Hence the definition of protein sequence from the N-terminus to the C-terminus.

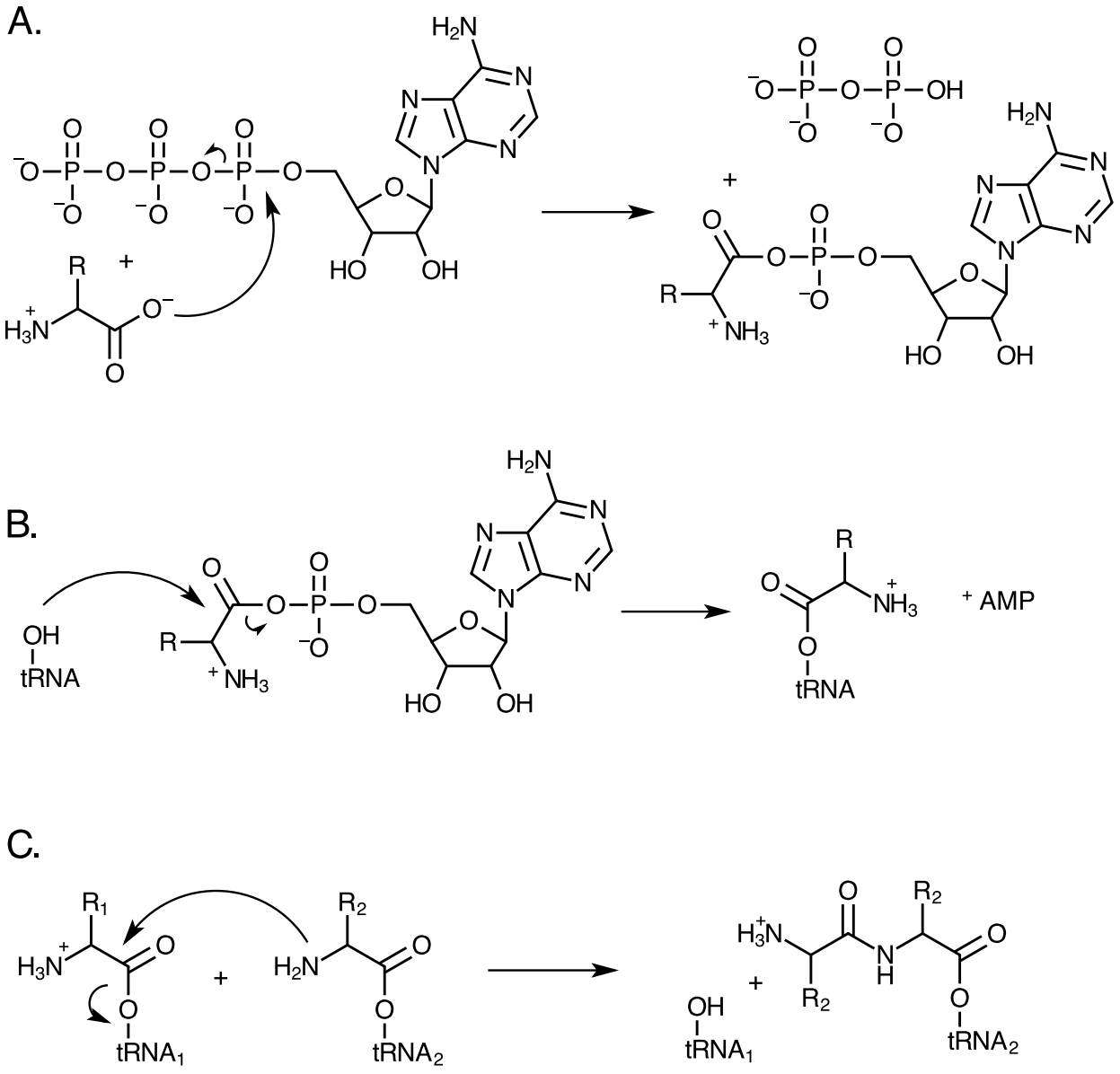

Figure CA.2. (A) Activation of amino acid's carboxylate group to a mixed acid anhydride with AMP via the sacrifice of ATP. (B) The amino acyl~AMP transfers the activated carboxylate to a tRNA molecule. An RNA hydroxyl group forms an ester with the carboxylate. (C) Amino acyl~tRNA molecules are the feedstock for protein synthesis. The tRNA is the leaving group and the amino group of the second amino acid in the growing chain attacks the first. I am being fast and lose with protons here, but you get the drift.

Chemical Synthesis of Polypeptides

In many biochemical studies it is advantageous to use peptides that are not derived from biological sources. Short segments are often difficult to obtain from biological sources. In addition, biological synthesis effectively limits a person to the 20 regularly encoded L-α-amino acids (with some non-trivial expansion possible). Chemical synthesis allows myriad variations in structure, limited only by the bounds of imagination.

The difficulty of using chemical synthesis is in moderating the efficiency of each chemical step. Imagine that synthesizing a 31-residue peptide is achieved by 30 sequential amide-forming reactions (in reality, there is more to it). If there is 90% efficiency at each step, then the end product will be achieved with 4% overall yield (0.9^{30}). Ninety percent is ridiculously high by most solution techniques, and even so will fail to provide much material.

To achieve improved yields, solid-phase chemical synthesis is used. Developed by Bruce Merrifield in the early 1960's, solid-phase techniques immobilize the starting material and each intermediate on a resin bead. In organic chemistry, a lot of material is lost of work-up; the process of quenching a reaction mixture and separating products from reagents, intermediates and impurities. By performing the reaction on a starting material attached to a solid support, reagents can be added and washed away at each step, while retaining the desired material in a small column containing the support.

A general scheme for solid-phase synthesis is shown in figure CA.3. A brief description of the steps follows:

1. The process starts by adding the C-terminal residue to an amine group linked to a protecting group linked to the resin. At some future point, the amine can be readily released from the protecting group, causing the peptide to fall off the support. Note that the added amino acid has a protected amino group itself (the Fmoc group; fluorenylmethyloxycarbonyl). That way the amine of the added amino acid will not react with its partners. The carboxylate of the amino acid is activated by a leaving group so that the resin amine can successfully do addition/elimination chemistry at the carbonyl carbon of that amino acid. Linkage is make and the first (C-terminal) amino acid residue is now attached to the residue. Excess materials are washed away.

2. The C-terminal amino acid is now ready to have the penultimate added. However, the first step is to deprotect the amino group of the resin-bound residue. The Fmoc group may be cleaved by the addition of piperidine, a weak base.

3. Now the chemistry in step 1 is repeated. A new amino acid is added, its carboxylate is activated and the next peptide bond is formed. Unreacted reagents are washed away.

4. Repeat step 2, and deprotect the amino group and get ready to cycle between steps three and four (is this a failure in computer logic?), until…

5. After all the residues have been added and the complete peptide is attached to the resin, it is time to release the peptide and deprotect the side chains (I haven't mentioned that yet – but it is an

issue. Side chains must be protected to avoid unwanted reactivity at those positions). The linkage to the resin is acid-labile, as are the protecting groups on the side chains. So, you just add acid (typically trifluoroacetic acid), wash the product from the column and purify the peptide by chromatography. Typical impurities are partially synthesized peptides, either missing an internal residue or lacking some stretch of the final residues to be added.

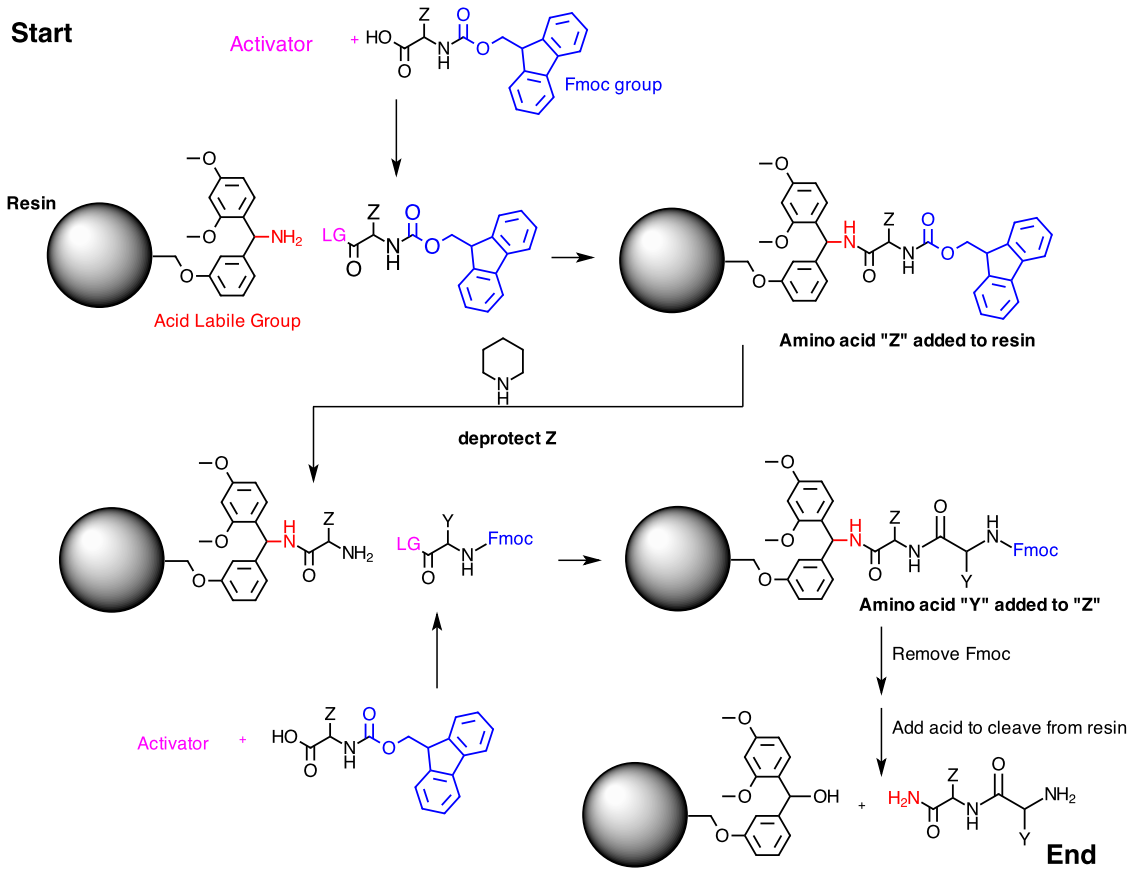

Figure CA.3. Scheme for solid phase synthesis of a two residue peptide (sequence "YZ"). Starting at upper left, an Fmoc-protected amino acid "Z" is activated with a leaving group. That allows attack of the resin-attached amine upon the carbonyl carbon of "Z", which is added to the resin. Deprotection by piperidine ($C_5H_{11}N$) creates a naked amine on "Z", which then attacks the activated carbonyl of amino acid "Y". Now the dipeptide is attached to the resin. More amino acids could be added sequentially, but here we deprotect the amine of "Y" with piperidine and then cleave from the resin with acid. Note that the C-terminus has a carboxamide functionality, not carboxylate. This is usually tolerated, but can be circumvented with added effort and expense.