I. Primary Structure and the Flow of Information in Biology

Introduction

Polymers in Biochemistry

There are three important polymers in biochemistry: deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and protein. Unlike the simple polymers described in organic chemistry such as polyethylene and polystyrene, which are **homopolymers** constructed from one building block, the biological polymers are **heteropolymers** constructed from a set group of structurally related subunits (**Figure I.1**).

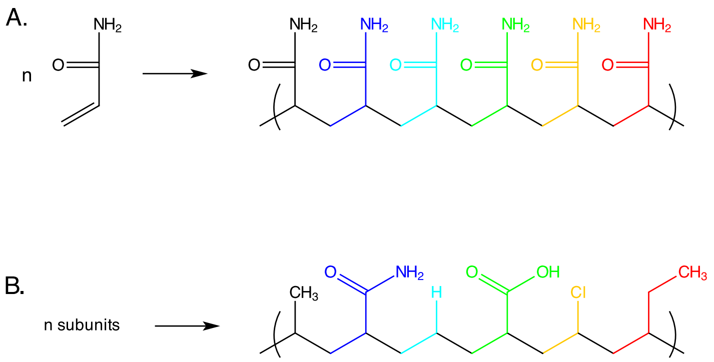


Figure I.1 (A) A homopolymer, polyacrylamide, is constructed from many copies of one subunit, acrylamide. (B) A heteropolymer, such as this one, is potentially constructed from many iterations of many different types of subunits.

Heteropolymers can be constructed randomly, with the different subunits distributed along the polymer chain with no reproducible order, or they can be constructed with a defined **sequence**, in which each polymer strand in a sample possesses the same ordering of subunits along the strand. The sequence is often described as the **primary structure** of an ordered polymer; it defines the covalent structure of the molecule. An important theme of this course is that **structure determines function**.[1]

Sequence and Heredity: The Central Dogma

Heredity might be defined as the capacity of an organism to pass on the structure of its molecules to its progeny. This is one of the fundamental chemical problems of life. How does an organism

[1] In organic chemistry, structure merely determined properties like melting point, boiling point, etc. Biological molecules perform a specific function in the organism, promoting survival of the individual and propagation of the species.

ensure that the structures of its three chief polymeric molecule classes are inherited by its offspring? Only one solution to that problem is in common usage among all self-supporting life forms (I'm not including viruses). It and is captured in the **Central Dogma of Molecular Biology** (Figure I.2).[2] Only DNA (among free-living organisms) provides a mechanism for self-replication of its covalent structure, and it is therefore the molecule that must be distributed by parents to their progeny. The covalent structure of RNA and protein can then be derived from the DNA structure. Before we address the chemical mechanisms of propagating structure, we need to define the polymeric structure of DNA, RNA and protein.

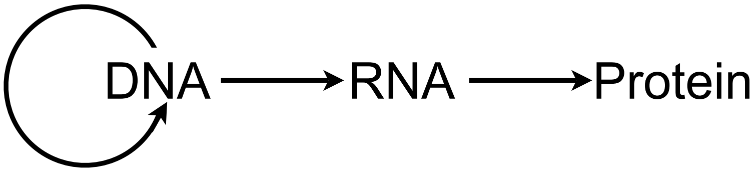


Figure I.2. The Central Dogma of Molecular Biology. Note that DNA possesses the ability to specify the structure of its copies, while also specifying the structure of RNA, which in turn specifies the structure of protein.

The Nucleic Acids: DNA and RNA

As noted above, DNA and RNA are both polynucleotides and are built from structurally similar subunits. Nucleotides are compounds that contain three components: a sugar, a phosphate and a nitrogenous base. The chief difference between DNA and RNA is the sugar (Figure I.1), which is ribose in ribonucleotides (RNA) and deoxyribose in deoxyribonucleotides (DNA). Both DNA and RNA commonly are composed of four different subunits, which vary in the identity of the nitrogenous base. Figure I.3 shows the nucleosides (sugar and base only) for DNA and RNA. The four deoxyribonucleosides are dA, dC, dG and dT and RNA is similarly composed of rA, rC, rG and rU (the "d" and "r" are often omitted if it is clear which type of nucleoside is being discussed). Note that the sole difference in bases between DNA and RNA is the additional methyl group on dT vs. dU.

Polynucleotides are formed through phosphodiester linkages between subunits. The sugar molecules have carbons numbered from 1' to 5' (1' is where the nitrogenous base is attached). The phosphodiester linkage is formed between the hydroxyl of the 3' carbon and the hydroxyl of the 5' carbon (Figure I.4). Short polymers can be defined with numerical prefixes (dinucleotides, trinucleotides, etc.) that indicate the number of subunits. These may also classified as oligonucleotides (oligo = few). As the polymer length grows, at some point they stop being oligonucleotides and simply are referred to as polynucleotides (usually around 50-100 subunits in length).

The sequence of a polynucleotide is defined as the ordering of nucleotides along the chain. By convention, the *first* nucleotide in the sequence has no additional nucleotide attached to its 5'

[2] The experimental work that led Francis Crick to define the "Central Dogma" is beautifully captured in "The Eighth Day of Creation" by Horace Freeland Judson. It should be high on your reading list for some break to come.

carbon and the *last* has no additional nucleotides attached to the 3’ end. Typically the sequence is given without the “d” and “r” since the presence of T’s and U’s make it clear that you have either DNA (T) or RNA (U).

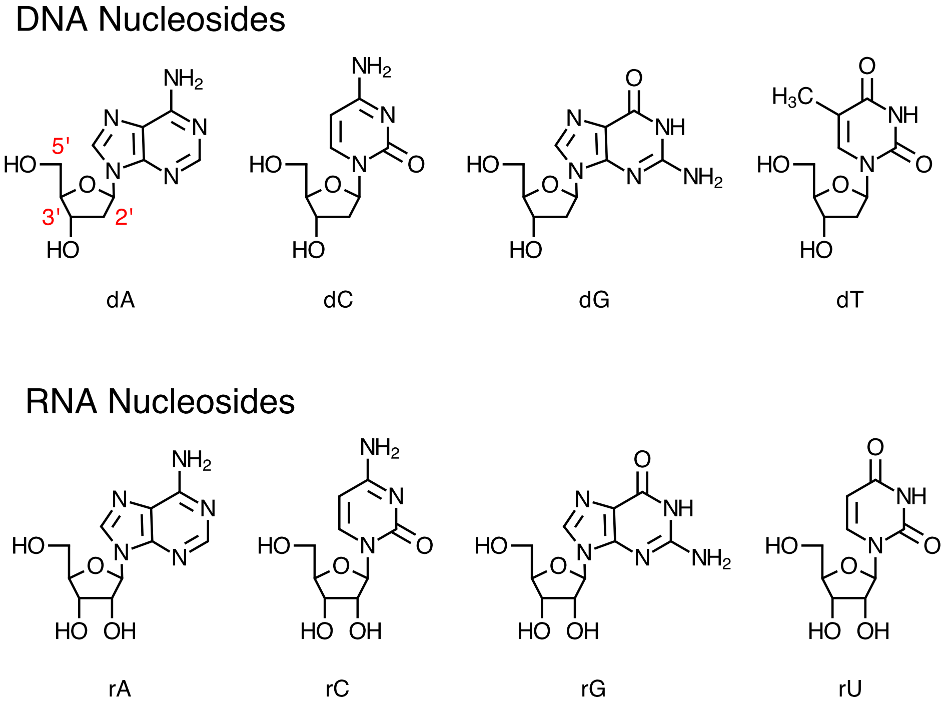


Figure I.3. Structures of the DNA nucleosides and RNA nucleosides. Sugar carbons on dA are labeled. Note the absence of a 2’-hydroxy group on the sugar of the DNA nucleosides and the difference of a ring methyl group between dT and rU.

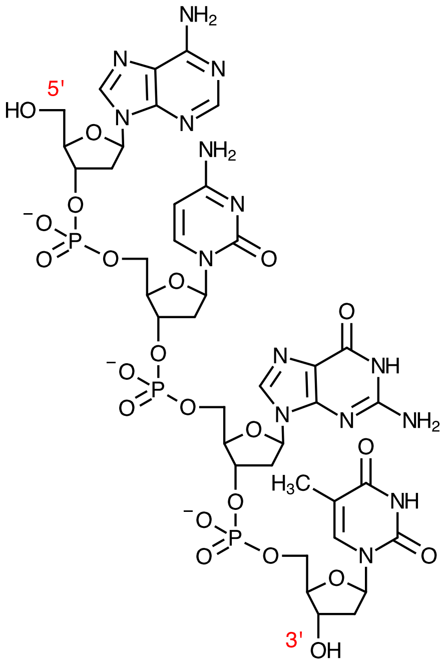


Figure I.4. A tetranucleotide of DNA. Note the phosphodiester linkages between nucleosides. The sequence of this tetramer is 5’-ACGT.

Proteins and Amino Acids

Structure of the 20 Amino Acids

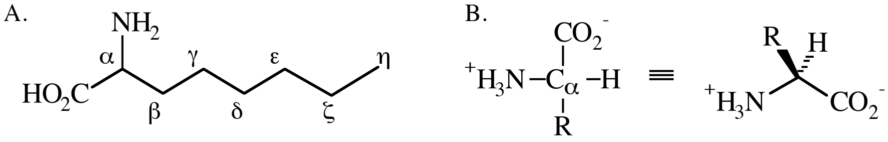


Figure I.5. (A) The alphabetic (Greek) labeling of carbons in an alkanoic acid, showing an amino group at the C_α. (B) A Fischer projection showing the L-configuration of the naturally occurring amino acids, converted to the standard 3D projection on a 2D surface.

The name "amino acid" describes the chemical nature of these molecules; each contains an carboxylic acid function and an amino function (Figure I.5A). 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 2.1A, 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 Figure I.5B.[3] 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 prevalence of one set of stereoisomers 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 “L” amino acids in building proteins.

The structures of the twenty amino acids are shown in Figure I.6. 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 2.2 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

[3] 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.

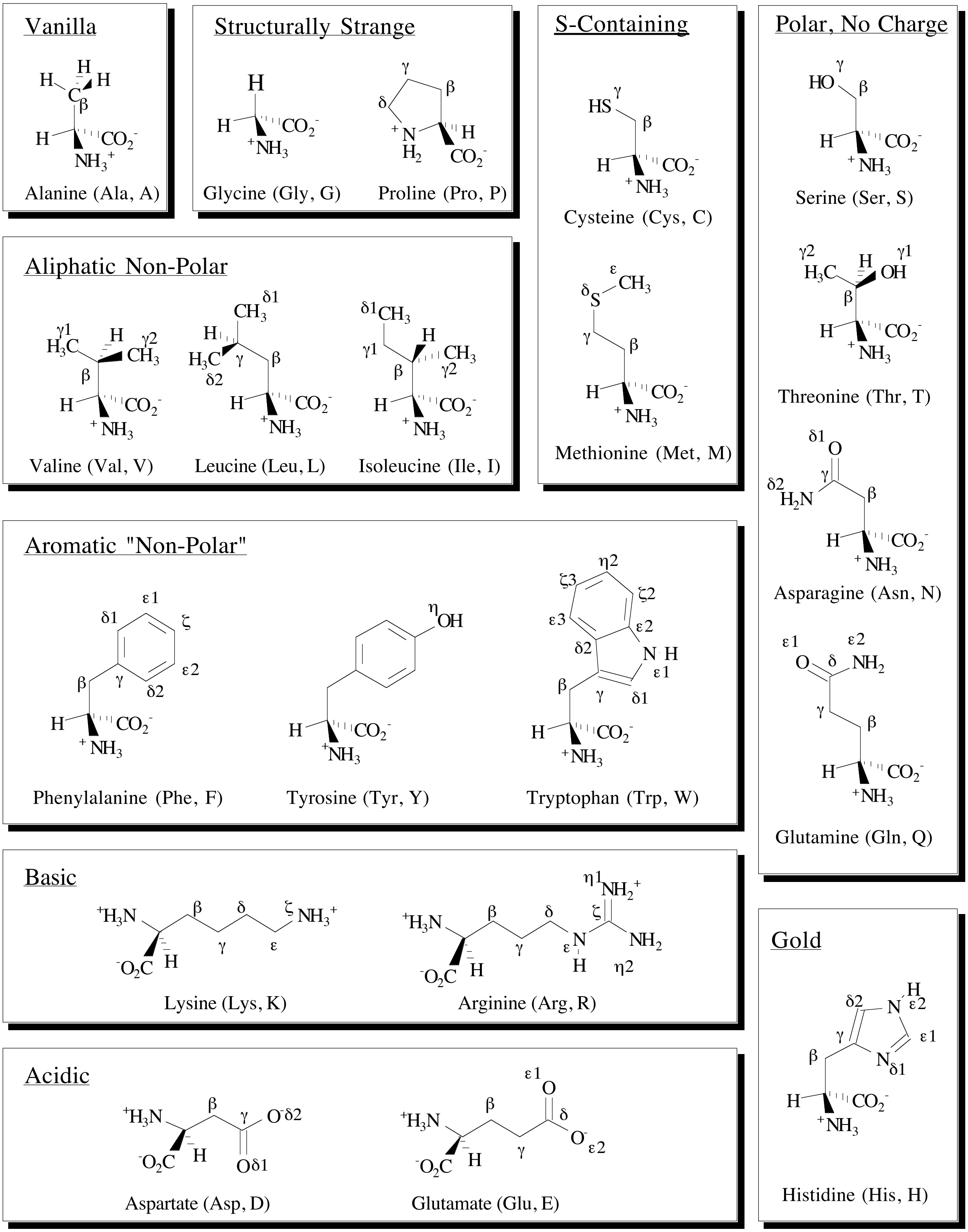


Figure I.6. Structures of the 20 amino acids.

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.

Acid-Base Chemistry

This is an aside to the current discussion, but is important in defining the structure of amino acids in neutral aqueous solution. The most fundamental aspect of amino acid chemistry relates to acid-base chemistry. Equations I.1-I.4 provide a brief review of some descriptors used in describing the properties of acids and bases.

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

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

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

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

Importantly, the transfer of a proton between donors and acceptors is an equilibrium process (Eq. I.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 I.2 and I.3, one obtains the Henderson-Hasselbalch equation (Eq. I.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.

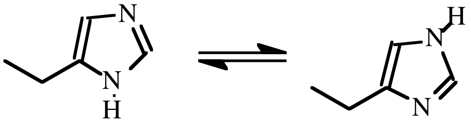
Table I.1 The pK_a's of some functional groups relevant to the acid base chemistry of amino acids.

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 I.1B, it is apparent that amino acids 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 2.3. From the Henderson-Hasselbalch equation (Eq. I.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 I.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 is of particular 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[4] 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

[4] 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.



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 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 I.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.

Table I.2 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.

Covalent Structure in Proteins

As noted previously, 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 referred to as a **peptide bond** (Figure I.7A). 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 I.7). 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.

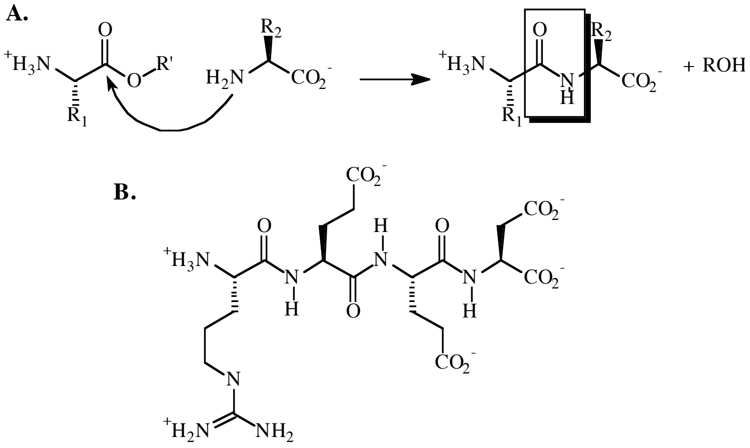


Figure I.7 (A) The α-amino group of one amino acid condensing with an ester of the the α-carboxylate of a second yields a peptide linkage (R' is the 3' hydroxyl of a tRNA molecule - see Ch. *). (B) This tetrapeptide has the sequence REED.

The preceding definitions are general references to the peptide bond. We can, however, get more detailed and start to 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**.[5] For example, in Figure I.7B 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.

How the Central Dogma Works

Now that the two classes of polymers have been described, we can look at how information in DNA is propagated to produce proteins of a given sequence via the Central Dogma.

Replication

The most important structural feature of DNA is its ability to act as a template for its own replication. The information in DNA is readily transmitted to create additional copies of that information because no DNA polymer exists in isolation (for long). The native structure of DNA

[5] 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.

is the double helix, in which two strands pair lengthwise in an antiparallel fashion (the 5'-3' orientations are opposite each other) and in a sequence specific fashion. The assembly of the DNA duplex is achieved through specific hydrogen bonding interactions that take place between bases in so-called Watson Crick base pairs (Figure ***). There are two classes of bases: purines, the two ring heterocycles (A and G) and pyrimidines, one ring heterocycles (C and T, or U in RNA). Base pairs form between a purine and a pyrimidine; A pairs with T or G pairs with C thanks to complementary hydrogen bonding interactions. With that rubric, you can predict the sequence of a second strand in a DNA duplex, knowing only the sequence of the first strand. For example, the tetradeoxyribonucleotide 5'-ACTG-3' with pair with 5'-CAGT-3' to form a duplex (Figure I.8).

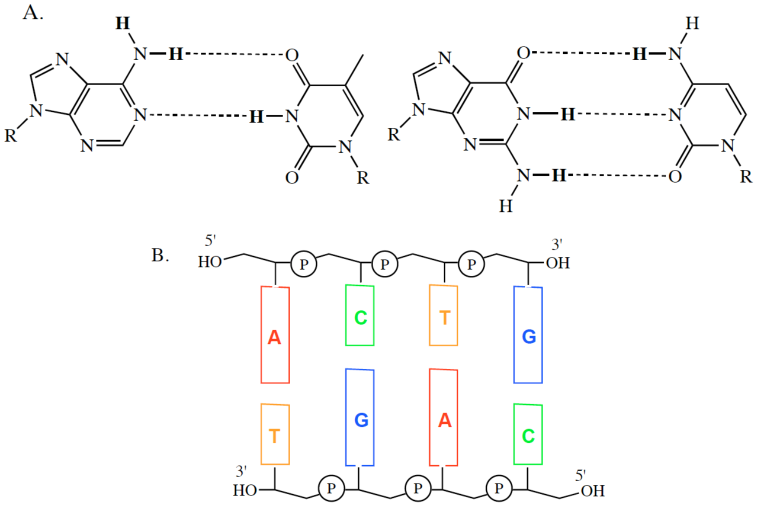


Figure I.8. (A) The Watson-Crick base pairing scheme for A:T (on left) and C:G (on right). R is the deoxyribose moiety. (B) Cartoon diagram of strand complementarity in duplex. Note that larger purine and smaller pyrimidine match in each base pair to create constant helix diameter.

Nature can do the same, as noted oh-so-subtly by Watson and Crick in the paper that first identified the hydrogen bonding scheme described above. "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material."[6] Indeed. That mechanism is called **semi-conservative replication** and involves using one strand as the information needed to generate its partner. In that manner, two DNA duplexes are generated from a parent, a process catalyzed by an **enzyme** called DNA polymerase (enzymes are proteins that catalyze reactions). Each "daughter" duplex contains one strand of the original duplex and one newly generated strand, but the fidelity of replication is assured thanks to the specificity of hydrogen bonding interactions (Figure I.9).

[6] J. Watson and F. Crick (1953) Nature 171, 737-738.

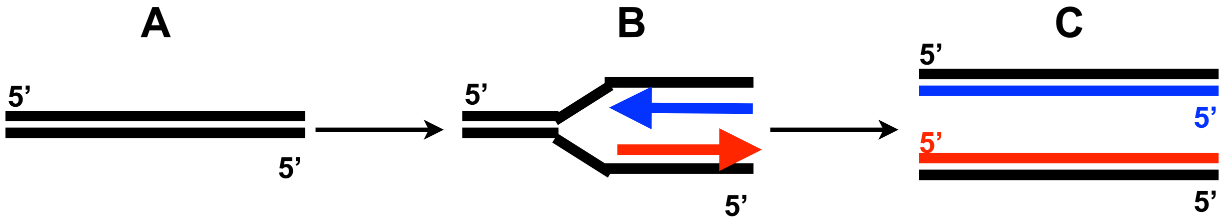


Figure I.9. Semi-conservative replication. The double-stranded DNA duplex (A), becomes the template for synthesis of two new strands (blue and red in B), which leads to two new duplexes, each containing one of the original strands and one of the newly synthesized strands (C).

Transcription

Transcription is literally the process of making a *copy* of an original text. In Biology, that meaning is subverted slightly to indicate the generation of an RNA strand that is a copy of a DNA strand. Because the polymers are slightly different (ribose vs. deoxyribose and U vs. T) the notion of copying isn’t quite perfect, but then in textual transcription hand-writing or font could vary as well, so maybe this is a good analogy.

In all events, and enzyme called RNA polymerase is responsible for binding to DNA and *transcribing* a gene (a section of DNA that encodes a protein) in to **messenger RNA (mRNA)**. The mRNA contains all the information of the gene, but can be transported away from the chromosome to act as instructions in protein synthesis.

There are two DNA strands but only one is transcribed to RNA – the sense, or coding, strand. To generate the copy of the DNA sense strand in RNA, the “anti-sense” strand must act as the template (Figure I.10). Transcription is a highly regulated activity in the cell and many intermediate stages exist between the gene, residing as in the DNA polymer, and the transcript, present as an RNA polymer, but those issues will be saved for later in the semester.

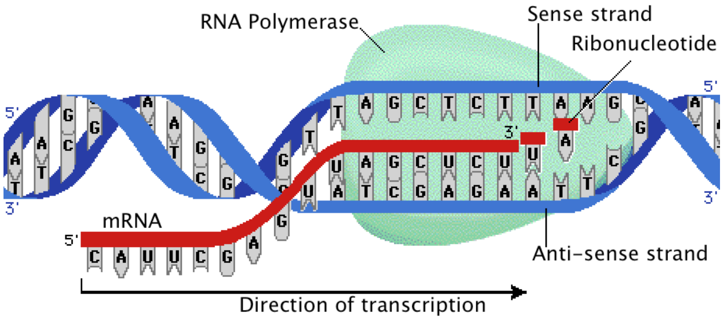


Figure I.10. Cartoon diagram of transcription. The anti-sense strand of DNA is used as a template to generate a copy of the sense strand as mRNA. (Figure adapted from http://www.phschool.com/science/biology_place/biocoach/transcription/tcproc.html without permission.)

Translation

While replication and transcription rely on the simple Watson-Crick hydrogen-bonding scheme for propagating genetic information, the process of translation – translating the sequence of a nucleic acid, mRNA, into protein – is more complicated. We will discuss the mechanism of the process in detail later this semester, but for now it is sufficient to indicate that the **ribosome** is the catalyst for translation, using mRNA as a template and a complex series of reactions including a second variety of RNA that helps correlate nucleic acid sequence to amino acid identity – transfer RNA (tRNA).

Second Base in Codon

First Base in Codon	U	C	A	G	Third Base in Codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Figure I.11. The Genetic Code. Nobel Prize to Marshall Nirenberg and H. Gobind Khorana (1968).

For the moment it is enough to indicate that there is a direct relationship between RNA sequence and amino acid sequence. Each three base sequence of RNA encodes a specific amino acid. Those three base sequences are referred to as "codons" (this was a time in Physics when all new particles were ???ons – bosons, fermions, muons, pions, etc.). Because there are four RNA bases, there are 64 (4^3) possible codon sequences for only 20 amino acids. Even with the addition of three "stop" codons (UAG, UAA, or UAG)[7], this abundance leads to a "degenerate"

[7] Note that there is also a "start" codon – AUG – that specifies methionine. As a result, the first residue in all proteins made on the ribosome is methionine. Not all AUG's are start codons however. The start codon is

code, where more than one codon can specify a particular amino acid. The intellectual achievement related to determining the **genetic code** (Figure I.11) can't be overstated. It is to molecular biology what the periodic table is to chemistry. To read Figure I.11 note that each amino acid lies on a row/column intersection of three bases. At the top left, Phe is found with the codon UUU (this was the first codon to be discovered). At the bottom right, GGG encodes glycine and in the middle, more or less, CAG encodes Gln.

Determination of Sequence

The genetic code provides the simplest route towards discovering the covalent structure of biological polymers. DNA sequencing has become routine and rapid. For about $5, various private and public companies will provide about 900 bp of DNA sequence overnight. Since DNA sequence determines RNA sequence, which in turn determines protein sequence, this is the most common route chosen for determining the covalent structure of proteins. By sequencing the gene encoding a protein of interest, one can obtain the sequence of that protein using the "key" provided by the genetic code. Alternate approaches to protein sequencing are available and will be presented, but the most important tool for protein sequence determination is DNA sequencing, which currently relies on **polymerase chain reaction** based techniques.

The Polymerase Chain Reaction

The polymerase chain reaction (PCR) is the most important technological tool in propagating and duplicating biochemical information. Starting with a single molecule of DNA (or RNA in an alternate technique) it is possible to generate billions of copies of that molecule within a couple of hours. PCR takes the replication process occurring in all organisms and puts it on steroids. The key ingredients to PCR are (1) template DNA, (2) a thermostable DNA polymerase, (3) feedstock nucleotides to be used by the polymerase (deoxyribonucleotide triphosphates; dNTP's) (4) short oligonucleotide primers that are complementary to the ends of the region of DNA to be duplicated, and (5) a thermal cycling machine.

As noted above, DNA polymerase catalyzes the synthesis of a copy of a template strand of DNA. In addition to the template, that reaction requires a "primer" that acts as a starting fragment of single stranded DNA to which nucleotides will be added in sequence. Without a starting fragment of primer DNA to build on, DNA polymerase can't do its job (Janis Shampay's research is built around the problems encountered when primers aren't available in the cell). The primer is elongated by the addition of nucleotide triphosphates complementary to the template DNA. The free –OH group at the 3' end of the primer performs nucleophilic attack on the α phosphate, displacing pyrophosphate (PP_i) and forming a new phosphodiester linkage. This reaction will proceed until there is no more template DNA to copy (Figure I.12).

distinctive because an additional sequence, AGGAGG, 8 bases to the 5' side (the Shine-Dalgarno sequence) specifies a particular AUG codon as a start codon. That is, AGGAGGnnnnnnnnAUG says "start making protein".

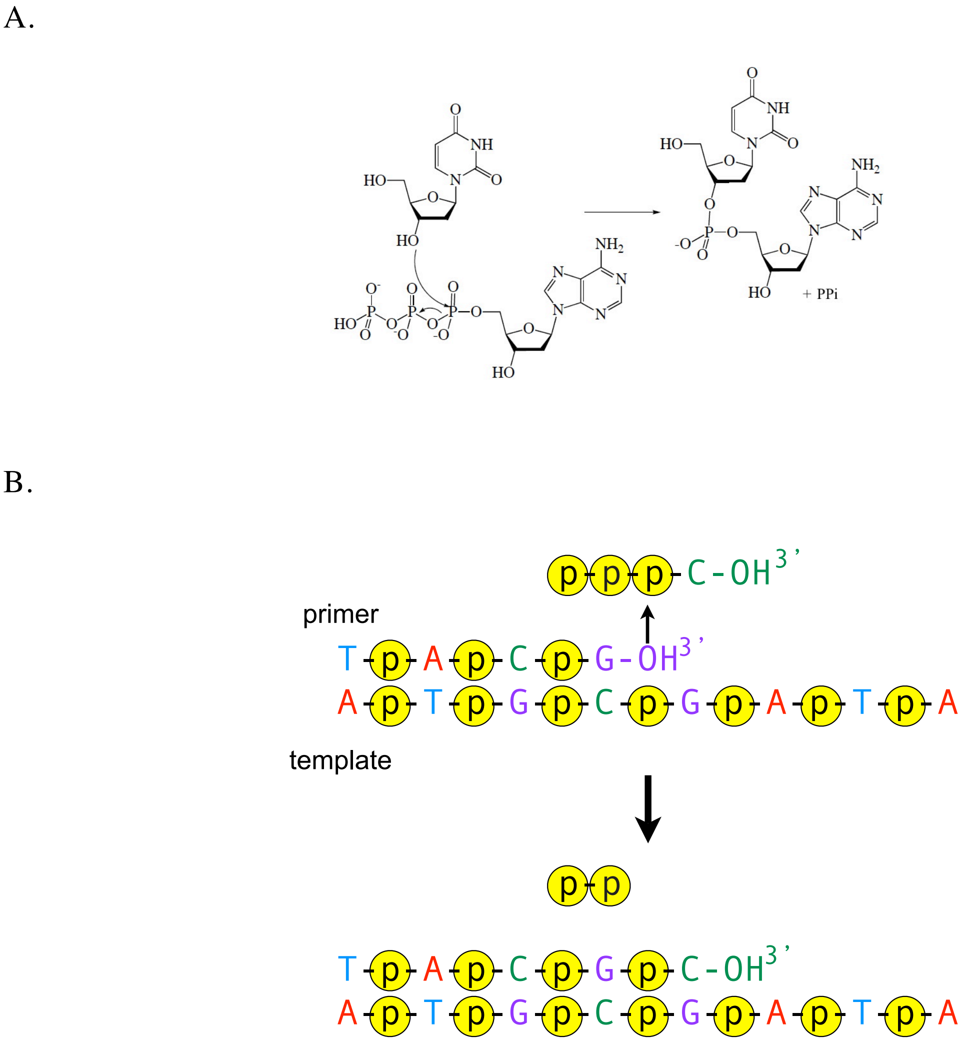


Figure I.12. The action of DNA polymerase. (A) Reaction diagram for the attack of the 3'-OH of a nucleoside onto the α-phosphate of dATP. (B) Schematic showing how primer is elongated by addition of dCTP onto the growing chain, opposite a G nucleotide on the template strand.

The novel aspect of PCR is the opportunity to continue to copy the template once an initial copy is made. Since the copy of the template creates a stable DNA duplex, the duplex must be "melted" by heating to a high temperature (typically above 95˚C). This creates two strands of DNA that can be used as templates in round two, but only if the DNA polymerase is still active as a catalyst. By using a thermostable enzyme and by cooling the mixture down to a temperature at which the primers will stably combine with the new templates (the annealing temperature), the reaction can start all over again (Figure I.13). Typically 20-30 cycles of this process are run on a single template, meaning 2^{20}-2^{30} copies of an original can be generated in a single experiment.

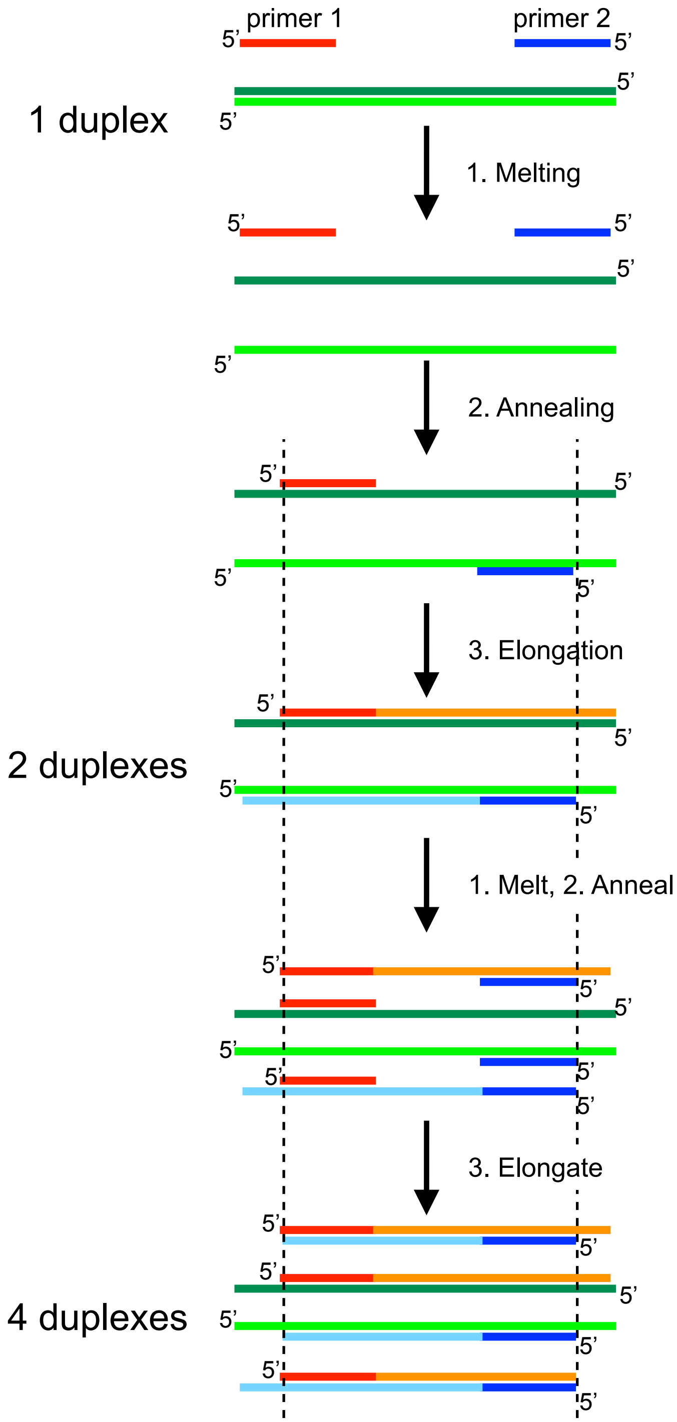


Figure I.13. Two cycles of PCR generate 2^2 (4) copies of the original duplex. Each cycle comprises 3 steps: (1) melt the existing duplex(es) at 95-98˚C. (2) cool rapidly with an excess of the two primers, so that some anneal with the template strands from the duplex, and (3) allow DNA polymerase to elongate from the primer to create a new strand. These two complete copies of the original duplex can be used in the second cycle and beyond.

DNA Sequencing

Standard DNA sequencing takes advantage of some of the features of PCR, but does not generate complete copies of both strands in a duplex. Instead it provides incomplete copies of only one strand. The modifications from Figure I.13 that achieve this are the use of a single primer, complementary to only one of the strands, and dideoxynucleotides (ddNTP) that poison the elongation step. One typically uses about 1 nmol of template and the method assumes that for each template being sequenced there is a reasonable chance of 500-1000 bases being copied before a ddNTP poisons further extension. Of course there is a similar chance that the copying will end after only a few bases have been added.

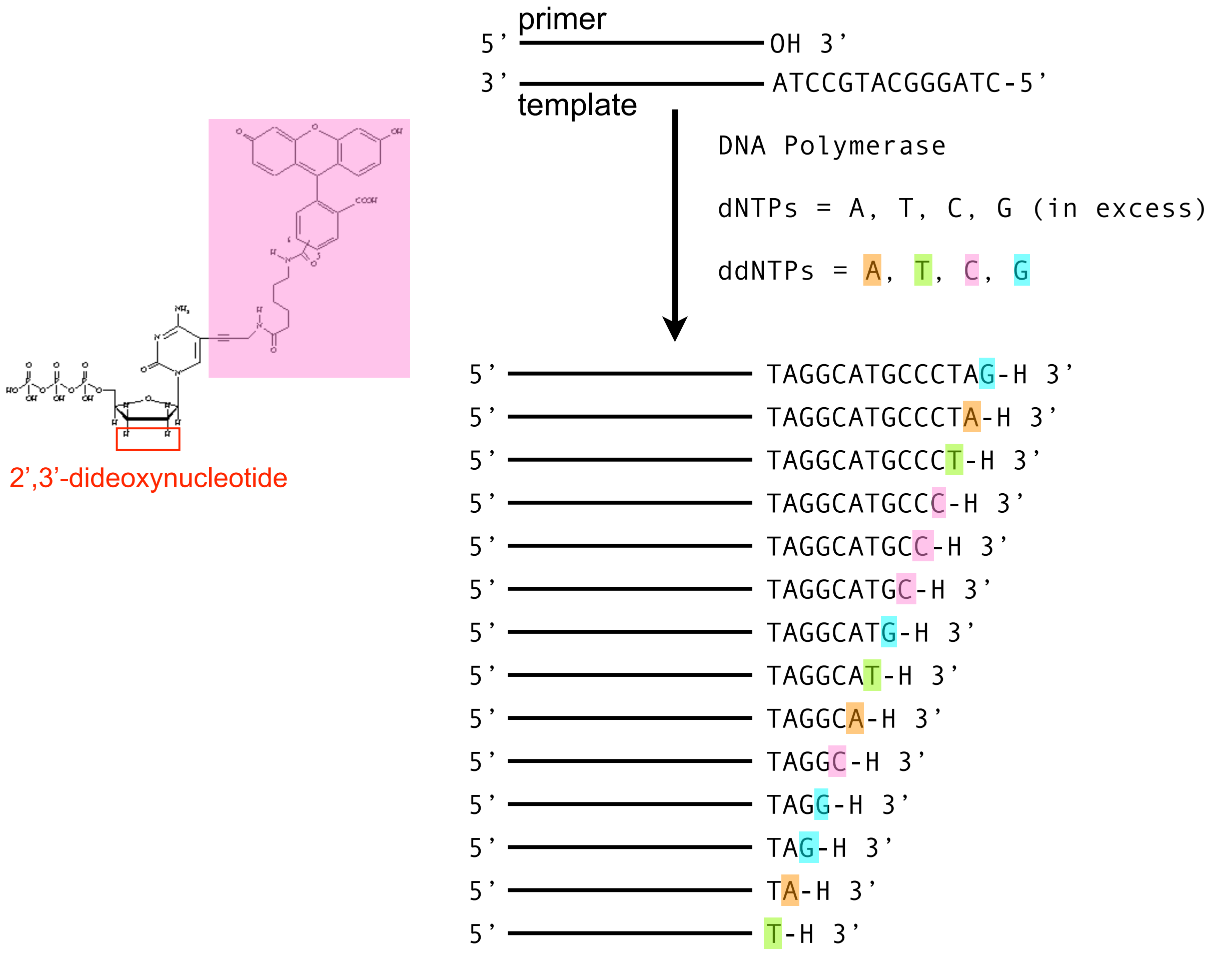


Figure I.14. DNA sequencing by chain termination. A primer complementary to the DNA to be sequenced (the template) is annealed, and then the primer is extended by the action of DNA polymerase. Occasionally, a fluorescently labeled dideoxynucleotide (ddNTP) is added to the growing strand. That simultaneously blocks further synthesis and provides a color for the strand related to the last nucleotide added. By sorting the many different terminated strands by size and monitoring color the sequence may be determined.

As a result - at the end of one cycle there will be a collection of copies of differing lengths, each capped with a dideoxynucleoside. To tell which nucleoside was the last one added, four different dyes are used – one for each base (Figure I.14). Because all fragments of a given length will end in the same dideoxynucleotide, all fragments of that length will be color-coded to that nucleotide. This process is repeated until there is a measurable population of each fragment size, and then the fragments are separated by capillary electrophoresis, a technique that separates molecules by size and charge (each added nucleotide adds one negative charge to a strand of DNA). Smaller fragments elute from the capillary more rapidly, and as each sized fragment elutes, the color of the attached dye is noted, allowing software to record the identity of each base added to increase fragment size by one. That generates the DNA sequence. From the DNA sequence one can then predict the mRNA sequence and the protein sequence. Of course, there may be a question of which piece of DNA is responsible for the protein that interests you, but geneticists are standing by to give you a hand.

Protein Sequencing

Traditionally, direct sequencing of proteins could be performed chemically by dissection of a protein into multiple fragments by either chemical or enzymatic digestion and then sequence determination of the individual peptides. Those individual sequences could be recompiled into the full protein sequence. This technique, referred to as **Edman degradation** in honor of Pehr Edman who developed it, is well-described in most biochemistry textbooks and on several on-line sources.

In the last decade, mass spectrometry has become the go-to technique for protein sequencing. Those methods are discussed in a separate set of notes.