Chapter 3: DNA and the Genetic Code

Introduction

Life's genetic code is written in the DNA molecule (aka deoxyribonucleic acid).[1] From the perspective of design, there is no human language that can match the simplicity and elegance of DNA. But from the perspective of implementation—how it is actually written and spoken in practice—DNA is a linguist's worst nightmare. DNA has four major functions: (1) it contains the blueprint for making proteins and enzymes; (2) it plays a role in regulating when the proteins and enzymes are made and when they are not made; (3) it carries this information when cells divide; and (4) it transmits this information from parental organisms to their offspring. In this chapter, we will explore the structure of DNA, its language, and how the blueprint becomes translated into a physical protein.

Physical structure of DNA

Few people in literate societies can avoid seeing a picture of DNA. Physically, DNA resembles a spiral staircase.[2] For our purposes here, imagine that we twist the staircase to remove the spiral so we are left with the ladder-like structure depicted in Figure 3.1. The two backbones to this ladder are composed of sugars (S) and phosphates (P); they need not concern us further. The whole action of DNA is in the rungs.

---

[1] Several types of virus of which the HIV is an important example are exceptions. Their code is written into a molecule called RNA (aka ribonucleic acid), DNA's "first cousin."

[2] Because scientists are prohibited from using highly technical terms like "spiral staircase," they refer to the structure of DNA as a double helix. A helix is a spiral structure.

[Insert Figure 3.1 about here]

Each rung of the ladder is composed of two chemicals, called *nucleotides* or *base pairs*, that are chemically bonded to each other. DNA has four and only four nucleotides: adenine, thymine, guanine and cytosine, usually abbreviated by the first letter of their names—A, T, G, and C. These four nucleotides are very important, so their names should be committed to memory.

Inspection of Figure 3.1 reveals that the nucleotides do not pair randomly with one another. Instead A always pairs with T and G always pairs with C.[3] This is the principle of *complementary base pairing* that is critical for understanding many aspects of DNA functioning.

Because of complementary base pairing, if we know one strand (i.e., helix) of the DNA, we will always know the other helix. Imagine that we sawed apart the DNA ladder in Figure 3.1 through the middle of each rung and threw away the entire right-hand side of the ladder. We would still be able to know the sequence of nucleotides on this missing piece because of the complementary base pairing. The sequence on the remaining left-hand piece starts with `ATGCTC`, so the missing right-hand side must begin with the sequence `TACGAG`.

DNA also has a particular orientation in space so that the "top" of a DNA sequence differs from its "bottom." The reasons for this are too complicated to consider here, but the lingo used by geneticists to denote the orientation is important. The "top"

of a DNA sequence is called the 5' end (read "five prime") and the bottom is the 3' ("three prime") end.[4] If DNA sequence number 1 lies between DNA sequence number 2 and the "top," then it is referred to as being *upstream* from DNA sequence 2. If it lies between sequence 2 and the 3' end, then it is *downstream* from locus 2.

DNA replication

Complementary base pairing also assists in the faithful reproduction of the DNA sequence, a process geneticists call DNA *replication*. When a cell divides, both of the daughter cells must contain the same genetic instructions. Consequently, DNA must be duplicated so that one copy ends up in one cell and the other in the second cell. Not only does the replication process have to be carried out, but it must be carried out with a high degree of fidelity. Most cells in our bodies—neurons being a notable exception—are constantly dying and being replenished with new cells. For example, the average life span of some skin cells is on the order of one to two days, so the skin that you and I had last month is not the same skin that we have today. By living into our eighties, we will have experienced well over 10,000 generations of skin cells! If this book were to be copied sequentially by 10,000 secretaries, one copying the output of another, the results would contain quite a lot of gibberish by the time the task was completed. DNA replication must be much more accurate than that.[5]

---

[3] I am eternally grateful to my parents for naming me Gregory Carey. Without the initials GC, I would never remember the pairing of DNA nucleotides.
[4] The terms 5' and 3' refer to the position of carbon atoms that link a nucleotide to the backbone.
[5] Of course, DNA does not replicate with 100% accuracy, and problems in replication may cause irregularities and even disease in cells. However, we do have the equivalent of DNA proofreading

[Insert Figure 3.2 about here]

The first step in DNA replication occurs when an enzyme (cannot get away from those enzymes, can we?) separates the rungs much as our mythical saw cut them right down the middle. The two strands of the DNA separate. Enzymes then grab on to nucleotides floating free in the cell, glue them on to their appropriate partners on the separated stands, and synthesize a new backbone. The situation is analogous to opening the zipper of your coat, but as the teeth of the zipper separate, new teeth appear. One set of new teeth binds to the freed teeth on the left hand side of the original zipper, while another set bind to the teeth on the original right hand side. When you get to the bottom, you are left with two completely closed zippers, one on the left and the other on the right of your jacket front.

DNA Packaging

To appreciate the way that DNA is packaged, we should first think of three types of objects—a very, very long piece of twine, several million jelly donuts, and the Eiffel tower. Taking the twine in one hand and a donut in the other, wrap the twine twice around the donut in the way shown in panel (a) of Figure 3.3. Leaving a few inches of twine free, grab another donut and loop the twine around it twice. Repeat this process until you have about six donuts with twine wrapped around them, giving a structure similar to that in panel (b). Arrange these donuts in a circle on the ground right next to the

---

mechanisms that serve two purposes—helping to insure that DNA is copied accurately and preventing DNA from becoming too damaged from environmental factors. A genetic defect in one proofreading

bottom of the Eiffel tower. Repeat this with another six donuts and place them, again in a circle, almost on top of the first six donuts. Continue repeating this process until there is a loop of these twine-donut complexes as depicted in panel (c).

[Insert Figure 3.3 about here]

Eventually, the pile of twine and donuts will become high enough to make it unstable and in danger of toppling over. To prevent this, have a few helpers take the pile of donuts and start circling it around a leg of the Eiffel tower to give it support, gluing it to the tower if necessary. Proceed with this strategy of looping twine around donuts, arranging circles of donuts, and snaking these piles in and around all the rigid structure of the tower. When you finally run out of string, jelly donuts, and tower space, you will have created a chromosome.

The twine in this procedure is the DNA molecule and the donuts are composed of small proteins[6]. Geneticists give special names to the twine-donut complex and to the loops of these complexes. Although the names are given in Figure 3.3, do not waste precious neuronal space memorizing them, except for one important term. Geneticists call this complicated, looping structure of DNA and proteins *chromatin*. The important thing to recognize is that DNA (or chromatin) is packaged as spirals within spirals within spirals.

---

mechanism leads to the disorder xeroderma pigmentosum which eventually results in death from skin cancer.

[6] These proteins are called *histones*.

The rigid structure of the Eiffel tower serves as a scaffold for the DNA-protein structures. To be truthful, the scaffold in the chromosomes is composed of proteins and not iron, and any physical resemblance between these proteins and the Parisian landmark is coincidental. Nevertheless, the DNA-protein spirals and loops bind with these scaffold proteins, albeit in ways that are not well understood.[7]

Chromosomes are discussed at length in chapter 5, but some preparatory words about them are necessary here. It is the chromosome, not the gene, that is literally the physical unit of genetic inheritance. Genes are sequences of DNA that contain the blueprint for the stuff that makes up proteins and enzymes, so thousands of genes can be physically located within a single chromosome. Cells do not carry the genes as physically independent snippets of DNA. Rather, when cells beget cells, even in the sperm and egg that carry the genetic material to the next generation, they do so with the DNA packaged into chromosomes.

RNA: Ribonucleic Acid

Before discussing the major role of DNA, it is important to discuss DNA's first cousin, *ribonucleic acid* or RNA. Besides its chemical composition, RNA has important similarities and differences with DNA. First, like DNA, RNA has four and only four nucleotides. But unlike DNA, RNA uses the nucleotide *uracil* (abbreviated as `U`) in place

[7] In addition to physically supporting and organizing DNA, proteins in the chromosomes may also have a functional role in gene expression. The general topic of gene regulation is discussed in Chapter 4.

of thymine (T). Thus, the four RNA nucleotides are adenine (A), cytosine (C), guanine (G), and uracil (U).

Second, the nucleotides in RNA also exhibit complementary base pairing. The RNA nucleotides may pair with either DNA or other RNA molecules. When RNA pairs with DNA, G always pairs with C,[8] T in DNA always pairs with A in RNA, but A in DNA pairs with U in RNA. When RNA pairs with RNA, then G pairs with C and A pairs with U.

Third, RNA is single-stranded while DNA is double-stranded. That is, RNA does not have the ladder-like structure of the DNA in Figure 3.1. Instead, RNA would look like Figure 3.1 after the ladder was sawed down the middle and one half of it discarded (with, of course, the added proviso that U would substitute for T in the remaining half).

Fourth, while there is one type of DNA, there are several different types of RNA, each of which perform different duties in the cell. Think of DNA as the monarch of the cell, giving all the orders. Unlike human monarchs, however, king DNA is unable to leave the throne room (i.e., the cell's nucleus) and hence, can never execute his own orders. The different types of RNA correspond to the various types of henchmen who carry out the King's orders. Some occupy buildings in outlying districts (ribosomal RNA), others transport material to strategic locations (transfer RNA), while yet others act as messengers to give instructions on what to build (messenger RNA). As we will see, the

[8] Another *Doe gratis* to mom and dad.

common language of the realm is the genetic code and it is communicated by the way of complementary base pairing.

The genetic code: A general perspective

DNA is a blueprint. It does not physically construct anything. Before discussing how the information in the DNA results in the manufacture of a concrete molecule, it is it important to obtain an overall perspective on the genetic code.

It is convenient to view the genome for any species as a book with the genetic code as the language common to the books of all life forms. The "alphabet" for this language has four and only four letters given by four nucleotides in DNA (A, T, C, and G) or RNA (A, U, G and C). In contrast to human language, where a word is composed of any number of letters, a genetic "word" consists of three and only three letters. Each genetic word symbolizes an amino acid. (We will define an amino acid later.) For example, the nucleotide sequence AAG is "DNAese" for the amino acid phenylalanine, the sequence GTC denotes the amino acid glutamine, and the sequence AGT stands for the amino acid serine. Like natural language, DNA has synonyms. That is, there is more than one triplet nucleotide sequence symbolizing the same amino acid. For example, ATA and ATG both denote the amino acid tyrosine.

The sentence in the DNA language is a series of words that gives a sequence of amino acids. For example, the DNA sentence AACGTATCGCAT would be read as a polypeptide chain composed of the amino acids leucine-histidine-serine-valine. Because

of the triplet nature of the DNA language, it is not necessary to put spaces between the words. Given the correct starting position, the language will translate with 100% fidelity.

Like natural written language, part of the DNA language consists of punctuation marks. For example, the nucleotide DNA triplets `ATT`, `ATC`, and `ACT` are analogous to a period (.) in ending a sentence—all three signal the end of a polypeptide chain. Other punctuation marks denote the start of the amino acid sequence for the peptide. Unlike the triplet nature of the DNA words for amino acids, some DNA punctuation marks may be more or less than three nucleotides.

Finally, DNA, just like a book, is organized into chapters. The chapters correspond to the chromosomes, so their number will vary from one species to the next. The book for humans consists of 23 different chapters or chromosomes.[9] The book for other species may contain fewer or more chapters with little correlation between the number of chapters and the complexity of the life form.

The differences between natural human language and DNAese are as important as the similarities. All differences reduce to the fact that human language is coherent while DNA is the most muddled and disorganized communication system ever developed. First, the chapters in a human language book are arranged to tell a coherent story. There is no such ordering to chromosomes.

Second, sentences in English physically follow one another with one sentence qualifying, embellishing, or adding information to another in order to complete a coherent

line of thought. The genetic language rarely, if ever, has a logical sequence. Metaphorically, one DNA sentence might describe the weather, the next give two ingredients for a chili recipe, and the third could be a political aphorism.

Third, whereas it is absurd to write an English compound sentence with a paragraph or two interspersed between the two independent clauses, DNA frequently places independent clauses of the same sentence in entirely different chapters.

Fourth, no English book would be published where most sentences are interrupted with what appears to be the musings of a chimpanzee randomly striking a keyboard. A single DNA sentence may be perforated with over a dozen long sequences of such apparent nonsense.

Fifth, with natural language it is considered bad rhetoric to repeat the same thought in adjacent sentences, let alone in the same words. With DNA repetition is the norm, not the exception. Not only does DNA continuously stutter, stammer, and hem and haw, but it also contains numerous nonsensical passages that are repeated thousands of times, sometimes in the same chapter.

Finally, the size of the DNA "book" for any mammalian species far exceeds that of any book written by a human. With eighty-some characters per line and thirty-some lines on a page, a 500 page book contains about 1,500,000 English letters. It would take over 2,000 such books to contain the DNA book of homo sapiens. And almost 90% of the characters in these 2,000 volumes have no apparent meaning!

---

[9] We humans actually have 23 *pairs* of chromosomes, one pair inherited from our fathers and the other

Protein Synthesis

Proteins and Enzymes Revisited

We now examine the specifics of how blueprint in the DNA guides the manufacture of a protein. Although we have already spoken of proteins and enzymes, we must now take a closer look at these molecules. The basic building block for any protein or enzyme is the *amino acid*.[10] There are twenty amino acids used in constructing proteins, most of which contain the suffix "ine," e.g., phenylalanine, serine, tyrosine. Amino acids are frequently abbreviated by three letters, usually the first three letters of the name—e.g., phe for phenylalanine, tyr for tyrosine. There are three major sources for the amino acids in our bodies. First, the cells in our bodies can manufacture amino acids from other, more basic compounds (or, as the case may be, from other amino acids). Second, proteins and enzymes within a cell are constantly being broken down into amino acids. Finally, we can obtain amino acids from diet. When we eat a juicy steak, the protein in the meat is broken down into its amino acids by enzymes in our stomach and intestine. These amino acids are then transported by the blood to other cells in the body.

A series of amino acids physically linked together is called a *polypeptide chain*. For now, think of a polypeptide chain as a linear series of boxcars coupled together. The boxcars are the amino acids and their couplings the chemical bonds holding them together.

---

pair inherited from our mothers. Hence, the total number of chromosomes is 46.

[10] Although proteins are composed of amino acids, amino acids can perform other functions than simply being chained together to make a protein or enzyme. Some are used as substrates from which important substances are synthesized. For example, in Chapter 2, we saw how the amino acid tyrosine was used to manufacture the neurotransmitters dopamine and norepinepherine. Several amino acids also operate as neurotransmitters in their own right.

The series is linear in the sense that it does not branch into a Y-like structure. The notion of a polypeptide chain is absolutely crucial for proper understanding of genes, so permit some latitude to digress into terminology.

Unfortunately, there are no written conventions for the language used to describe polypeptide chains, so terminology can be confusing to the novice. Typically, the word *peptide* is used to describe a chain of linked amino acids when the number of amino acids is small, say, a few hundred or less.[11] The word *peptide* is also used as an adjective and suffix to describe a substance that is composed of amino acids. For example, a *peptide hormone* is a hormone that is made up of linked amino acids[12], and a *neuropeptide* is a series of linked amino acids in a neuron. The phrases *polypeptide chain, polypeptide*, or *peptide chain* usually refer to a longer series of coupled amino acids, sometimes numbering in the thousands. Be wary, however. One can always find exceptions to this usage.

We are now ready to define our old friend the protein. A *protein* is one or more polypeptide chains physically joined together and taking on a three dimensional configuration. The polypeptide chain(s) comprising a protein will bend, fold back upon themselves, and bond at various spots to give a molecule that is no longer a simple linear structure. An example is hemoglobin, a protein in the red blood cells that carries oxygen. It is composed of four polypeptide chains that bend and bond and join together. Some

---

[11] To add to the confusion, the term peptide is used in chemistry to define a particular type of bond between two substances, and the term peptide, as it is used here, derives from the nature of the chemical bonds (i.e., the boxcar couplings) between the amino acids. These are peptide bonds in which a nitrogen atom of one amino acids binds to a carboxyl group of the next amino acid.

[12] A hormone is any substance that is released directly from a cell and not through the duct of a gland. Hormones enter the bloodstream and are diffused throughout the body.

proteins contain chemicals other than amino acids.[13] A particularly important class of proteins is the receptor protein that resides on the cell membrane (but sometimes within the cell) and is responsible for "communication" between the cell and extracellular "messengers."

Finally, we must recall the definition of an enzyme. An *enzyme* is a particular class of protein responsible for metabolism.

With these definitions in mind, we can now present one definition of a gene. *A gene is a sequence of DNA that contains the blueprint for the manufacture of a peptide or a polypeptide chain*. Such genes are sometimes qualified by calling them *structural genes* or *coding regions*.[14] A synonym for gene is *locus* (plural = *loci*), the Latin word meaning site, place, or location.

Protein Synthesis: the Process

We can now look at the actual "manufacturing process" whereby the information on the DNA blueprint eventually becomes translated into a physical molecule. There are five steps in this process—(1) transcription; (2) editing (or post transcriptional modification); (3) transportation; (4) translation; and (5) spit and polishing and final assembly (or post translational modification).

---

[13] For example, the hemoglobin molecule contains a heme group, and a whole class of proteins (called lipoproteins) contain lipids (i.e., fats).

[14] This somewhat restrictive definition for a gene is often encountered among molecular biologists. Many human geneticists will define a gene in a looser sense as simply a section of DNA, regardless of whether or not it contains the code for a peptide or polypeptide.

(1) *Transcription*. Depicted in Figure 3.4, transcription is the processes whereby a section of DNA gets "read" and copied into a molecule of RNA. In the first step of transcription, the bonds joining the two nucleotides that comprise a rung of the DNA ladder are broken by an enzyme. Then, a chain of RNA is synthesized from one strand of the DNA using the principles of the complementary pairing of nucleotides. For example, if the DNA sequence is `GCTAGA`, then the RNA sequence that is synthesized will read `CGAUCU`. In this way, the information in the DNA is faithfully preserved in the RNA, albeit in the genetic equivalent of a "mirror image."

[Insert Figure 3.4 about here]

The process of transcription does not occur everywhere on the DNA. Instead, punctuation marks in certain sections of DNA act as recognition sites to initiate the transcription of RNA. Other punctuation marks act as stop signs to terminate transcription. Later on, we will see how these sites play an important role in genetic regulation. For now, let us just recognize that only 10% to 15% of human DNA ever becomes transcribed into RNA. Of the rest, about 1% consists of punctuation marks and the remaining 99% has no discernible function.

(2) *Editing* (or *post transcriptional modification*)[15]. After transcription, the RNA contains three different types of information.[16] The first of these is, of course, the information about the amino acid sequence for the peptide chain. These sections of RNA

[15] The term "editing" describes the process quite well, but is naturally not pedantic enough for most molecular geneticists who prefer the phrase "post transcriptional modification."

are called *exons* and contain the actual blueprint for peptide synthesis. The second type of information is the punctuation mark. These marks are the biological equivalents of saying "the information on making this polypeptide starts here" or "stops here." The third type is called an *intron*. In terms of blueprint information, introns are literally junk.[17] That is, they do not contain code for the amino acid sequence of the polypeptide. Although all RNA that is transcribed from DNA begins with a series of punctuation marks, there can be many different exons and introns in a single molecule of transcripted RNA.

Editing is a cut and paste process in which the junk is eliminated from the RNA, leaving only the important punctuation marks and the message. In more precise terms, the introns are physically cut out of the RNA transcript and the exons are spliced together. Figure 3.5 gives an example of this process. After this process is complete, the resulting RNA is termed *messenger RNA* and is abbreviated as mRNA.[18]

[Insert Figure 3.5 about here]

One important term for mRNA is the *codon*. A *codon* is a series of three nucleotides that contain the message for a specific amino acid.[19]

(3) *Transportation*. King DNA is imprisoned in the nucleus of the cell, but the actual synthesis of a polypeptide chain takes place in the cytoplasm. In transportation,

---

[16] There are some other RNA nucleotides that comprise a "header," a "tail," and other things, but they need not concern us.
[17] Introns may, however, play other important parts in gene transcription and regulation.
[18] Some microbiologists refer to the RNA molecule after editing as *mature* messenger RNA.
[19] Although the term codon most often refers to a three nucleotide sequence in mRNA, it may also refer to the three nucleotides in DNA.

the mRNA exits the nucleus, enters the cytoplasm, and attaches to an almost forgotten old friend, a ribosome. Recall that the ribosome is a "protein factory," in the sense that it is the physical location where the polypeptide chain is created. Ribosomes are composed of RNA (to which the mRNA attaches) and various proteins and enzymes that are the tools used in the manufacturing process.

(4) *Translation*. In translation, the message on the mRNA is read and a peptide chain is synthesized. To examine this step, let us begin with another class of RNA called transfer RNA or tRNA, depicted in Figure 3.6. In addition to ribonucleic acid, one molecule of tRNA carries one amino acid. Each tRNA also contains an *anticodon*—a series of three nucleotides that acts as a label identifying the specific amino acid attached to the tRNA. For example, the anticodon `AAG` means that the tRNA carries phenylalanine and the anticodon `ACG`, shown in Figure 3.5, denotes that the tRNA molecule carries tryptophan.

The process of translation is illustrated in Figure 3.6. The mRNA molecule moves through the ribosome until a punctuation mark is encountered that signifies "the next three nucleotides comprise the first codon, so start the synthesis of the polypeptide chain here." The first codon on mRNA then moves through the ribosome and a tRNA molecule that has the appropriate anticodon to the codon is attached to the mRNA. In Figure 3.6, the first mRNA codon is `UUU` (denoting phenylalanine or `phe`), so only a tRNA molecule with the anticodon `AAA` (and, hence, carrying phenylalanine) will pair with it.

Consequently, the first amino acid for the polypeptide chain will be phe or phenylalanine.

[Insert Figure 3.6 about here]

The second codon is then "read," the appropriate tRNA binds to it, and the amino acid carried by the tRNA is physically bound to the first amino acid. In Figure 3.6, the second codon is ACG so the tRNA molecule that binds with it will be UGC, carrying the amino acid threonine (thr). The nascent polypeptide chain now consists of the amino acid sequence phe—thr.

The mRNA then moves through the ribosome, the next codon is "read," the appropriate tRNA molecule is attached, and the amino acid that it carries is physically joined to the previous amino acid. This is depicted on the right hand side of Figure 3.6. The codon is CGG, and the tRNA molecule has the anticodon GCC. Hence, the amino acid arg (arginine) is added to the polypeptide chain. (Note how the tRNA molecule associated with the first amino acid has now dissociated from it.) This process continues—one codon moving through the ribosome, having a tRNA molecule attach to it, and then having the amino acid cleaved and joined to the polypeptide chain—until a punctuation mark on the mRNA signifies termination of the message.

(5) *Final assembly* (or *post translational modification*). In some cases, the polypeptide produced after translation is a perfectly fine biological molecule. Here, the polypeptide chain folds, bends, and binds upon itself to take on a three dimensional configuration, and the protein or enzyme is complete. In other cases, the polypeptide is a

raw product that requires further processing at the finishing table. Sometimes the polypeptide chain may be cleaved in one or more places to give a smaller, but biologically functional, peptide—the biological equivalent to removing spurs from a metal casting. More often, two or more polypeptide chains are joined together to make a functioning protein or enzyme. The process of assembly is too complicated and protein-specific to describe here, so we merely state that intracellular processes will join two or more polypeptide chains together to form the final protein. An important twist on the assembly process is that some molecules other than polypeptide chains may be added to generate a protein complex. For example, a chain of lipids (i.e., fat) may be added, giving a lipoprotein complex.

Hemoglobin: An example of the genetic code and its organization

The hemoglobin protein will figure prominently in several different sections of this book, so it will be used here to illustrate the genetic code and the organization of the genome. It will also help us to practice the genetic lingo we have learned in this chapter.

When we breathe in air, a series of chemical reactions in our lungs extracts oxygen atoms and implants them into the hemoglobin protein in our red blood cells. The red blood cells pulse through our arteries and eventually reach tiny capillaries in body tissues (e.g., liver cells, pancreas cells, muscle cells, neurons, etc.) where the hemoglobin releases

the oxygen atoms. In humans over five months of age, hemoglobin is composed of four polypeptide chains, two α chains and two β chains.[20]

Figure 3.7 depicts the DNA segment containing the gene for the β polypeptide. This long section of DNA section is located on chromosome 11 and is about 60,000 nucleotides long (or 60 kb, where kb denotes a kilobase or 1,000 base pairs). Only the tiny box with the label β contains the blueprint for the β peptide chain. (For the moment, ignore the boxes labeled ε, Gγ, Aγ, and δ.)[21] The boxes labeled $\psi\beta_1$ and $\psi\beta_2$ are called *pseudogenes* for the β locus. A pseudogene is a nucleotide sequence highly similar to a functional gene but its DNA is not transcribed and/or translated.

The middle section of Figure 3.7 gives the structure of the β locus, including the "punctuation marks." This locus is roughly 1,600 base pairs long and includes three exons. The first exon is composed of the 90 nucleotides that have the code for the first 30 amino acids in the peptide chain, the second exon codes for the 31st through 104th amino acids, and the last for the remaining 40. Hence, of the 1,600 base pairs only 438 contain blueprint material. If we add to these about 60 nucleotides worth of punctuation marks, then less than a third of the whole β locus contains the actual blueprint and processing information for the polypeptide chain. The final section of Figure 3.7 gives the actual nucleotide sequence for the beginning of exon 1.

---

[20] I am lying here. There is another chain called the δ chain. But only 3% of hemoglobin molecules contain this chain, so we can safely ignore it in order to make things simple.

[21] Three of the other four genes—ε, Gγ, and Aγ—are expressed in the embryo, fetus, and neonate. They will be described in a later chapter. The δ gene produces a δ hemoglobin chain that also form functional adult hemoglobin with the α chain. However, this locus is not strongly expressed, so only about 3% of adult hemoglobin is composed of δ chains.

[Insert Figure 3.7 about here]

Figure 3.8 depicts the DNA region for the α chains. This is located on an entirely different chromosome from the β cluster, chromosome 16, and is roughly 30kb in length. The boxes labeled α_1 and α_2 both contain the blueprint for the α peptide chain. This is an example of a *gene duplication*—the DNA for both of these loci is transcribed, edited and translated into the same α chains. Like the β chain, there is also a pseudogene for the α locus, denoted in Figure 3.8 by the box labeled $\psi\alpha_1$. (Once again, ignore the boxes labeled ζ_1, and ζ_2.[22]) The actual structure for the two α loci is very similar to that of the β locus—they too have three exons—and is not depicted in Figure 3.8.

[Insert Figure 3.8 about here]

Adult hemoglobin is composed of two polypeptide chains coded for by the α loci and two chains coded for by the β locus. These are joined together in the assembly stage of protein synthesis. Two heme groups are added to the protein. (A heme group contains an iron atom that permits oxygen to bind when the hemoglobin is carried through the lungs.) The structure of the hemoglobin molecule is given in Figure 3.8.

Hemoglobin is a good example of the capricious organization of our human genome. Only about 10% of the DNA in both clusters is actually used, and the two clusters are on entirely different chromosomes! In addition, the fact that there are two α loci but only one β locus requires an elaborate genetic control mechanism to insure that an equal number of α and β polypeptide chains is manufactured. A human engineer

would clearly arrange things quite differently by, for example, placing the α and β locus close together so they are transcribed as a unit. This would insure that an equal number of α and β chains is produced, avoiding the mess of regulatory mechanisms. If any intelligent being created the human genome, then he (or she, of course) must have had a few beers before starting the project and considerably more than a few while removing the bugs!

[22] Loci ζ_1 and ζ_2 are expressed in the early embryo and are discussed in a later chapter.

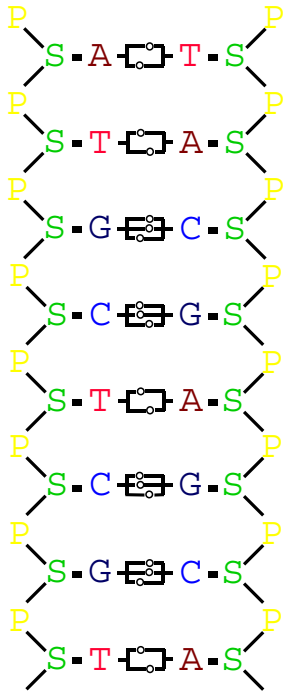

Figure 3.1. The physical structure of DNA. P = phosphate, S = sugar, A = adenine, T =Thymine, C = Cytosine, G = Guanine

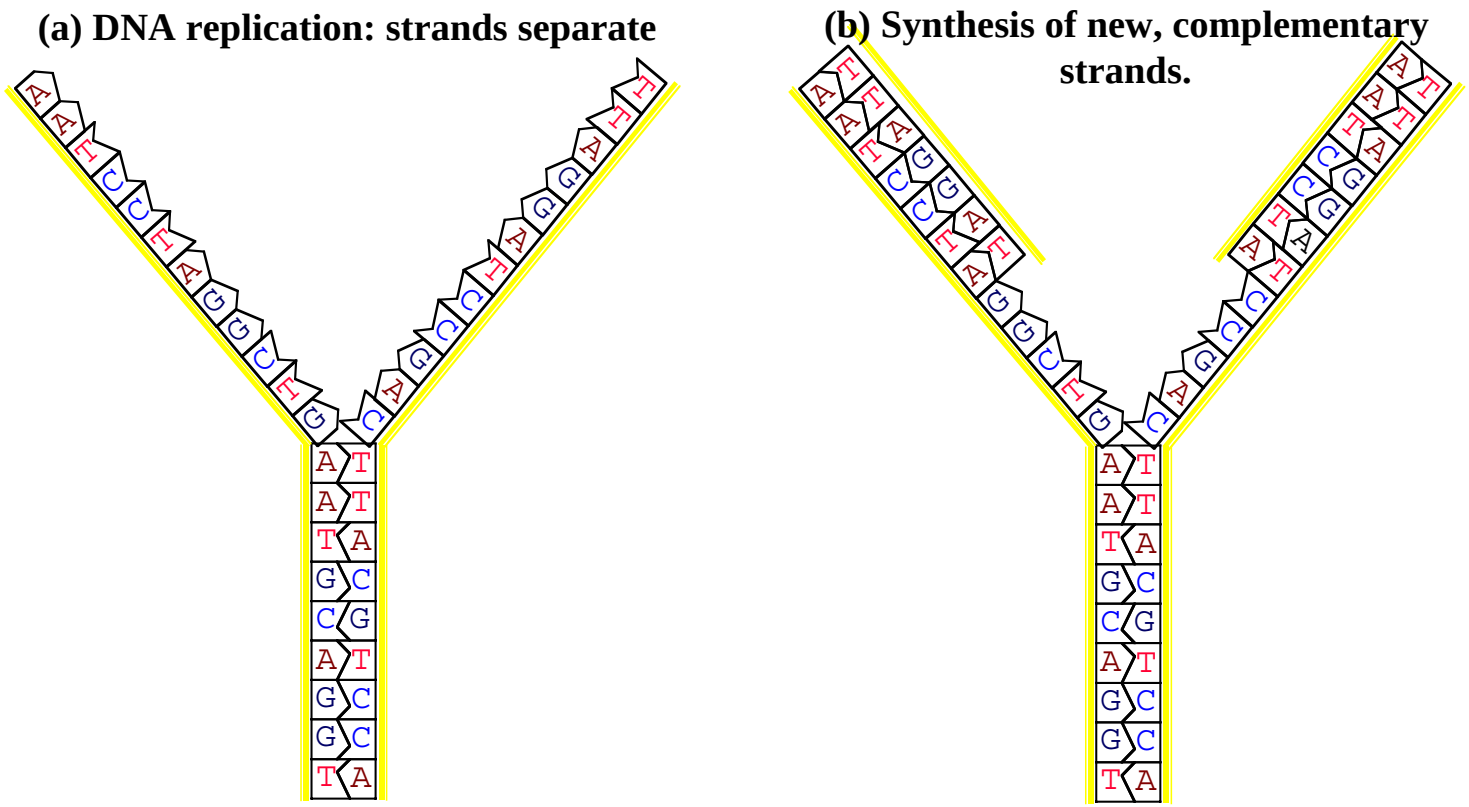

Figure 3.2. DNA replication. (a) Enzymes separate the strands. (b) other enzymes attach to the single-stranded DNA and attach free nucleotides according to complementary base pairing.

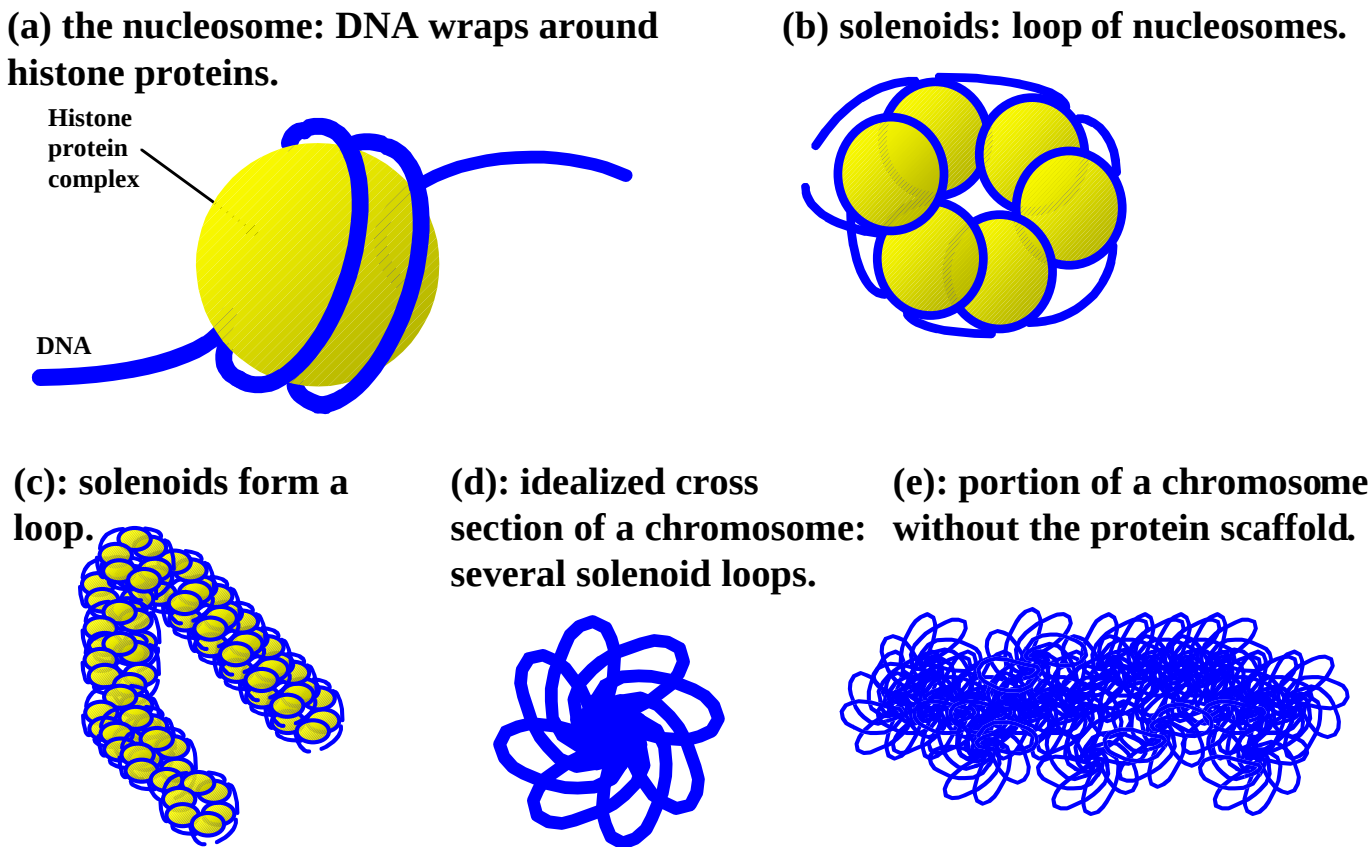

Figure 3.3. How DNA and protein scaffolding creates chromosomes.

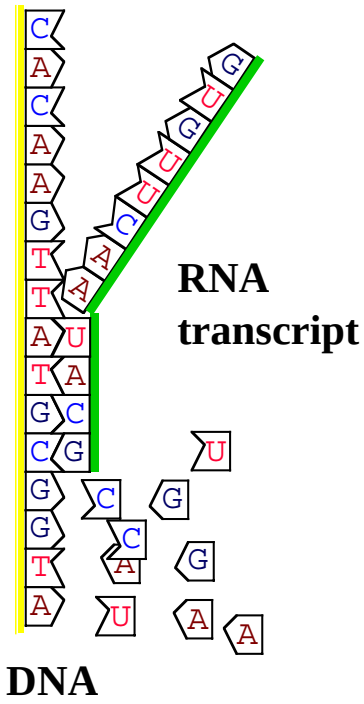

Figure 3.4. Transcription. Enzymes attach to DNA, curt the hydrogen bonds and make the DNA single-stranded. Other enzymes attach to the DNA strand and synthesize a chain of RNA using free nucleotides and complementary base pairing.

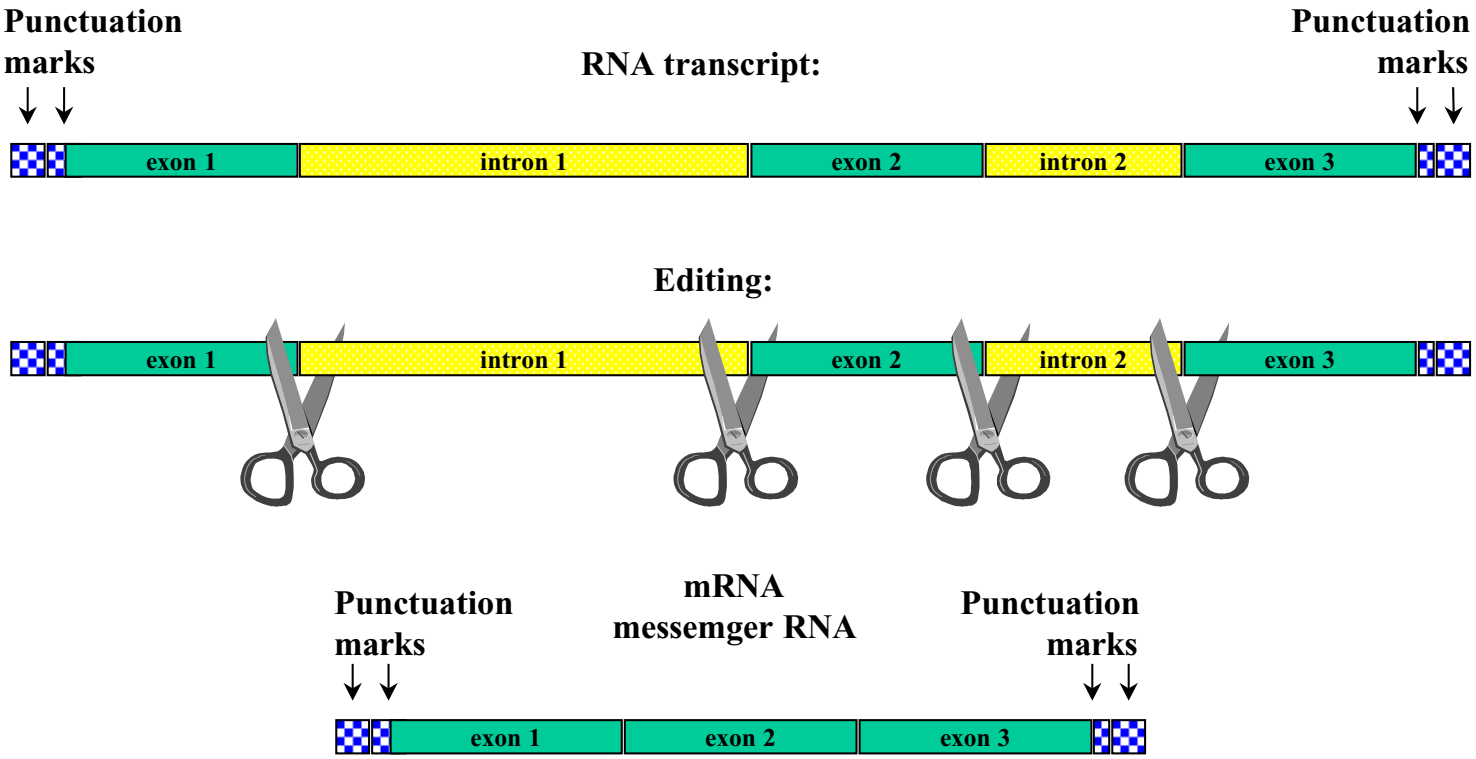

Figure 3.5. Editing. Enzymes cut out the junk (introns) and splice the message (exons) together.

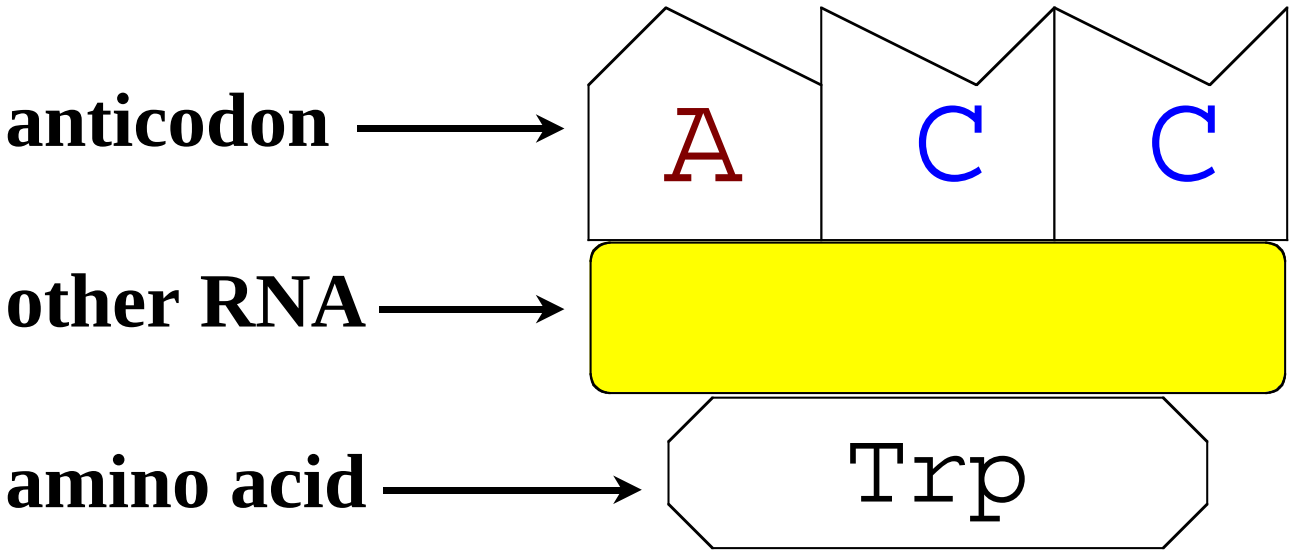

Figure 3.6. Schematic of a transfer RNA (tRNA) molecule. The anticodon acts as a "bar code" signaling the specific amino acid that the molecule is carrying.

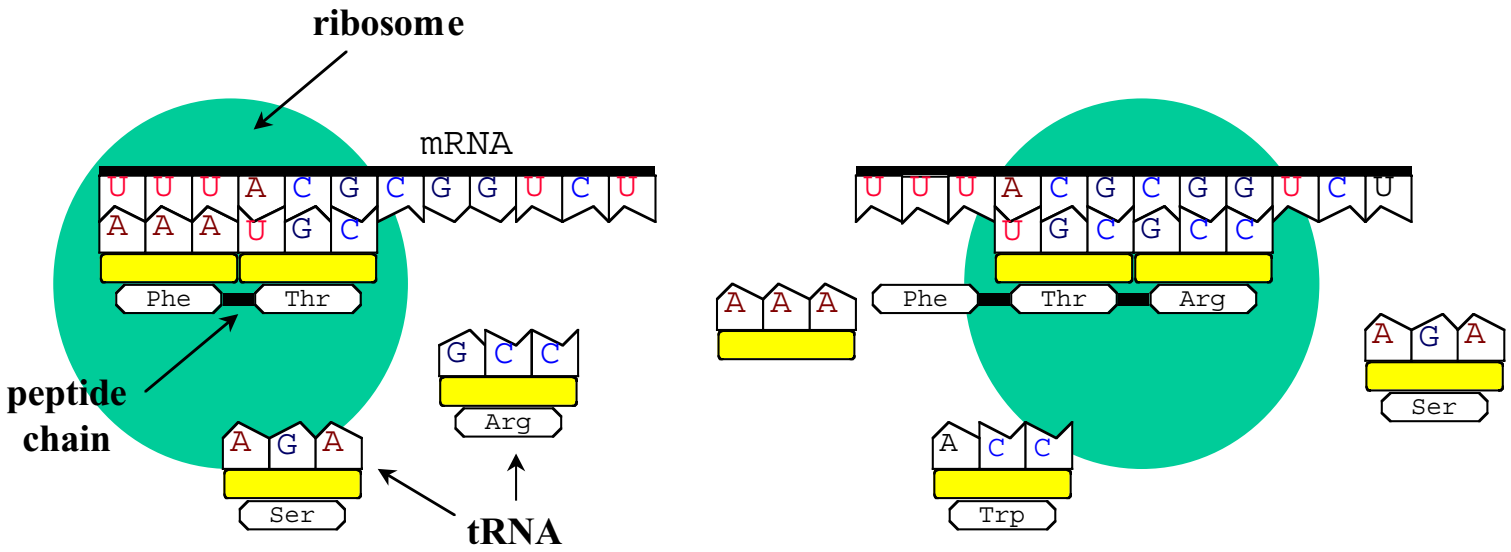

Figure 3.7. Translation. The mRNA strand attaches to a ribosome. Left panel: Two codons then attract their complementary anticodons using the principles of base pairing; enzymes cleave off the two amino acids from the tRNA molecules and join them together. Right panel: The mRNA moves through the ribosome like a ticker tape; the next tRNA molecule binds with the mRNA codon, the amino acid is released and then attached to the growing polypeptide chain.

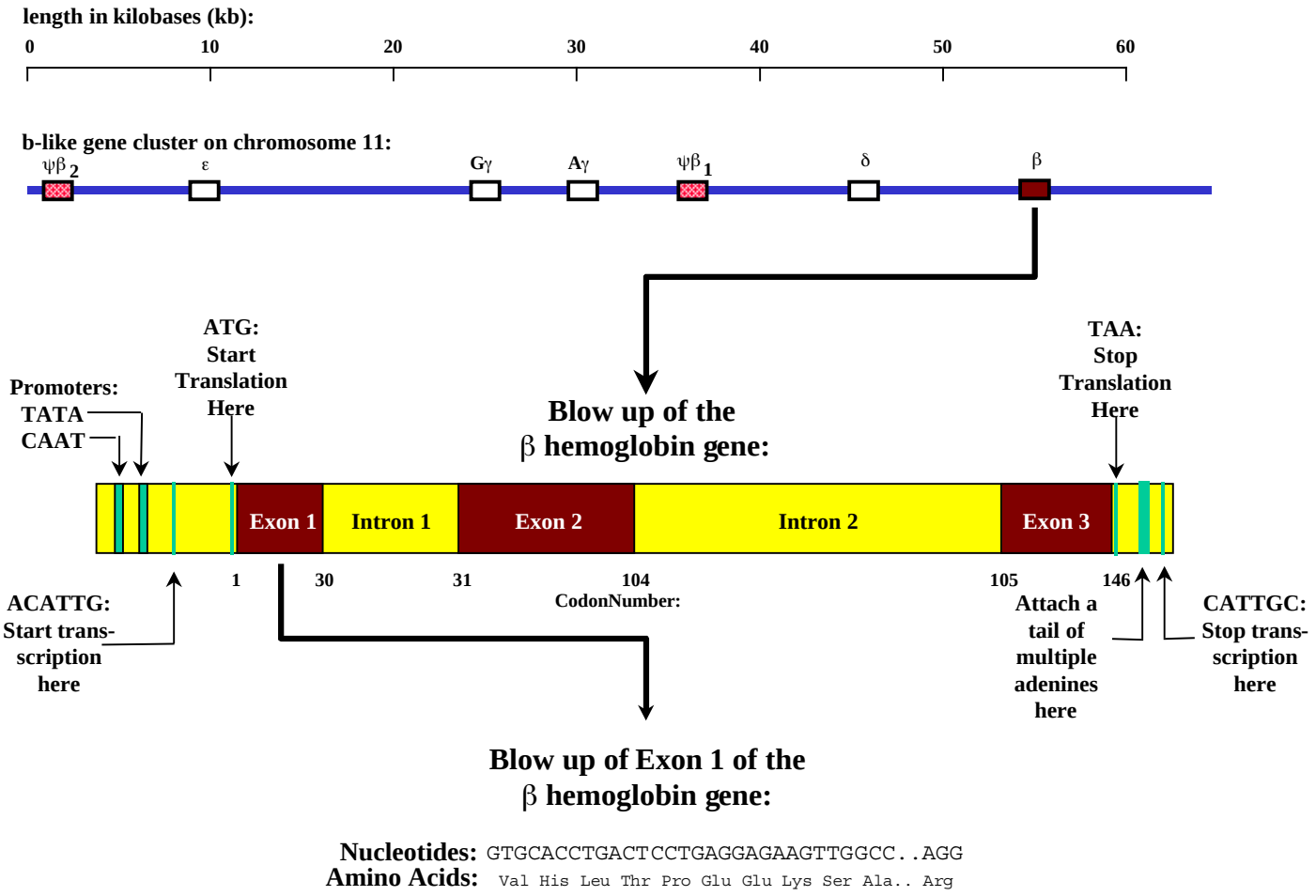

Figure 3.8. Schematic of the DNA sequence that codes for the β polypeptide chain of the hemoglobin protein.

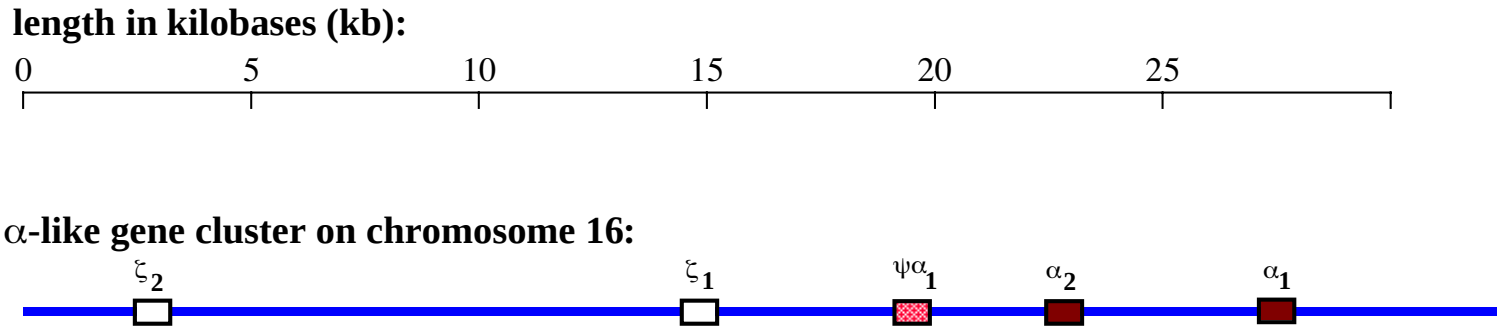

Figure 3.9. The structure of the DNA sequence coding the for α polypeptide chain of the hemoglobin