1 How DNA Makes Stuff

It's all well and good to say that DNA is the blueprint for your body, but after you've said it, what does it mean? In order to build a house, a blueprint is read by a builder, who then uses it to decide what to nail into what. A cell is often said to use its DNA in the same way, although the DNA is more like a list of construction materials than a set of instructions.

As we've seen in previous chapters, proteins are what you're made of, and DNA contains the recipes for making them. There are two steps to the process. The DNA is first used to make another nucleic acid called RNA which comes in several varieties, and the different kinds of RNA are then used to make proteins. This way the DNA can stay safe in its nucleus and the RNA copies of it can be transported to the protein-making factories in the cell, the ribosomes.

Transcription is easier to say than "the RNA-making process," so scientists use that name instead. And **translation** is less of a mouthful than "the protein-making process," so we use that name for the second step. Of course each of these steps is itself a pretty complicated process, so we'll look at each one closer now.

Transcription - Copying the DNA

It can be misleading to think of DNA as a blueprint, but it works sometimes, like now. (But only just for now—see the box on the next page.) You don't bring the master plans for a building to the job site, where things are messy and windy. You make a copy of the plans, and bring them instead. (And if you copy them with the cheap ferroprussiate process, the prints come out blue, oddly enough.)

DNA is the master plan for a cell, and the cell wants to keep it safe. So the DNA stays inside the nucleus of the cell, and a copy must be transcribed into a form that can be taken away, edited, copied again, destroyed, munched, and otherwise used for real work. This form is called **RNA**, or **ribonucleic acid**.

RNA is very similar to DNA in some respects. Chemically, it's close, but with two important differences. The first is that it has an extra oxygen atom hanging off the side (this is why the DNA is "deoxy-"), and instead of the T base, RNA uses uracil, a close cousin. The other three bases (A, C, and G) are the same in RNA as in DNA. Uracil matches with the adenine base in the same way as thymine does in DNA, and the cytosine and the guanine go together, just like in DNA. Another important difference is that it usually comes in a single strand, unlike the DNA double helix.

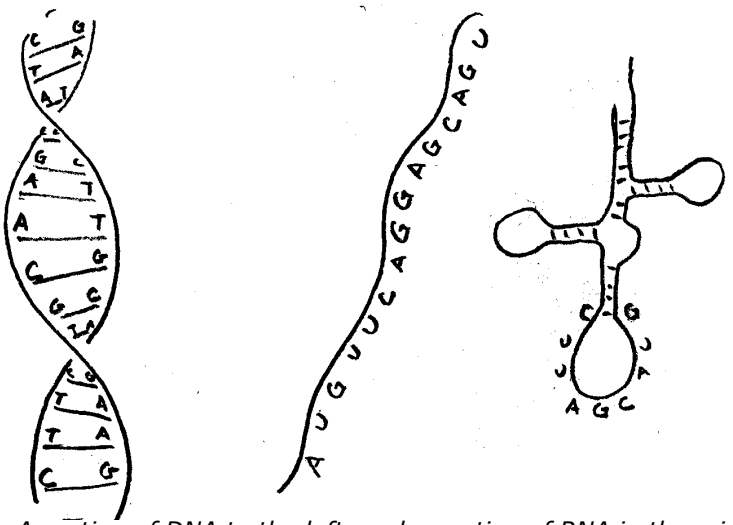

Figure 1-1. A section of DNA to the left, and a section of RNA in the middle. RNA comes in a single strand, and uses uracil (U) instead of thymine (T). RNA is sometimes present as a long strand, but often it comes in little pieces, or curled into funny shapes. On the right is one of the entertaining shapes that RNA can take. In this case, it's a piece of "transfer RNA", which is folded back on itself to expose three nucleotides on the bottom loop.

Other differences between DNA and RNA

In addition to the chemical differences—the oxygen molecule and the number of strands—there are several other important differences between DNA and RNA.

- RNA's single strand can be wound up in some fairly peculiar ways, more like a complicated protein than the tidy double helix of DNA. The complementary bases on the RNA are sometimes used to hold these complicated curls together and sometimes to interact with some other nucleic acid.
- RNA comes in several varieties, which have different functions. They're all relatively small, sometimes encoding an entire protein, and sometimes much much smaller than that. Scientists are still completing the catalog of the different kinds of RNA.
- RNA is *active*. It was once thought that RNA was just a pattern, like DNA, or like a blueprint. But it turns out that many varieties of RNA function like enzymes. Instead of being just another form of a blueprint, these types of RNA are more like a guy with a hammer. This means that your body uses them to do stuff, like bringing parts of proteins together, and trimming and splicing genes.

We'll meet the different varieties of RNA one at a time, as we step through the transcription process (and its aftermath) in the rest of this chapter.

Protein Factors - Getting Ready

The first step in transcription is for a bunch of different transcription "factors" to collect near the DNA segment to be copied. These are a group of proteins whose job is to glom onto the DNA strand near the gene to be transcribed. You can think of them as a guide or a jig for the enzyme complex that does the job of transcription.

Most genes have one or more **promoters**, short stretches of DNA just before (in the 3' direction) the beginning of the gene. There are promoters right near the gene, and some further away. Both of them provide a place for an assortment of different transcription factors to hang on. The factors, in turn, determine how fast the gene gets transcribed, or whether transcription happens at all.

Kinds of Promoters

There are a couple of different classes of promoters. The first group, the **basal promoters**, are usually around 30 base pairs from where transcription begins. They almost always have a particular genetic sequence: TATAAAA, called the **Tata box**. The others, the **upstream promoters**, can be up to a few hundred base pairs away. The factors binding to the upstream promoters are different for different genes. How exactly factors so far upstream affect the transcription of a gene is not known, but some think that the factors bind to each other, and pull the DNA into loops, allowing part of it to come near the gene, where it could interfere with the other transcription factors.

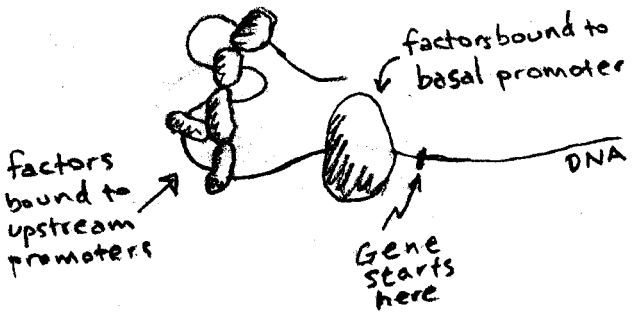

Figure 1-2. Here's a segment of DNA right at the start of a gene. The gene goes off to the right, and to the left there's a glob of transcription factors bound to the basal promoter.

In addition to the transcription factors that bind onto promoter sites near the gene, there are also factors that bind to regions of DNA thousands of base pairs away from the gene. These factors are called **enhancers**, and they act to speed up transcription. It's not completely clear how they do it, since they're so far from the gene, but some think that they also act to bend the DNA strand around into a loop and connect with the transcription

factors closer to the gene. The opposite of an enhancer is called a **silencer**. This is a protein that acts to slow or halt the gene transcription.

Hormones and Promoters

Many hormones do their hormonal thing by acting on DNA promoters, as part of the mix of transcription factors, attaching to a cell's DNA to command it to make some substance, or suppress some other substance (often both). Growth hormones, for example, act by promoting the creation of the protein building blocks of your body.

The action of the promoters, transcription factors, enhancers, and silencers is called gene regulation, and it's how your body decides how much of the protein from some gene should be produced. The mechanics of regulation are amazingly variable—you'd be tempted to say that 50 different genes can have 50 different arrangements of transcription factors, but that would be an understatement. It's more like 50 different genes having hundreds of different arrangements, since many genes can be triggered in several ways.

Gene regulation is an incredibly complex (and controversial) part of modern genetics, but you don't have to understand it all in order to get the basics of transcription. Which is good, because there aren't any scientists who understand it all, either. We'll have more to say about regulation in chapter XX. For now we'll just leave it that each gene has a bunch of transcription factors and assorted other features that make the next step possible.

RNA Polymerase - Making the Copy

Once the transcription factors are in place, transcription can begin. The workhorse for this process is a collection of enzymes called **RNA polymerase**. There are a few of these, but the one most intimately connected with the process of making proteins is called RNA polymerase II (also called RNAP II or pol II). The DNA is all twisted up, of course, so an enzyme that plays an important role here is **helicase**, whose job it is to unwind the DNA helix a little bit so the transcription factors can get at it.

Using the transcription factors and the helicase, the RNA polymerase unzips a little of the relevant part of the DNA and begins to move down it in the 3' to 5' direction (which you might remember from chapter XX). As it moves, it assembles a chain of RNA to match the corresponding parts of the DNA. In a petri dish, the RNAP II spins the DNA as it slides down the helix, like some kind of toy. We don't know which one spins when they're in a live cell. About 50 bases get copied in a second, so it can take a while, since small genes can still have hundreds of thousands of bases.[1]

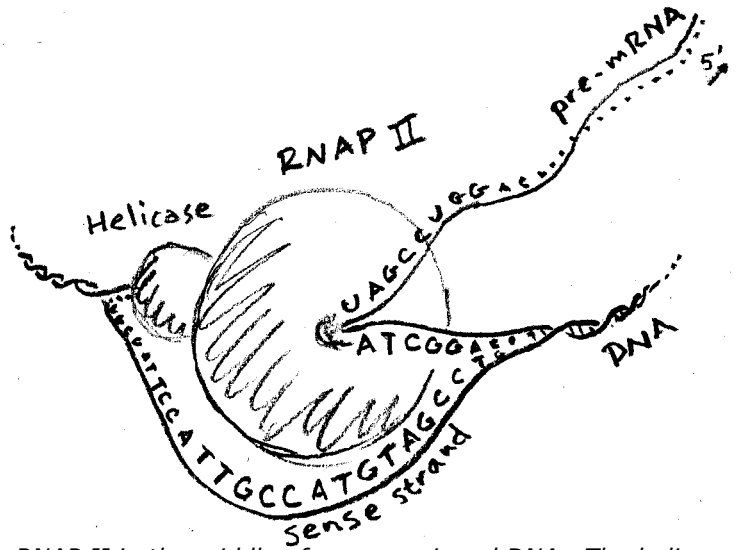

Figure 1-3. RNAP II in the middle of some unzipped DNA. The helicase unwinds the DNA strands, and the RNAP II takes the "antisense" strand and assembles a strand of mRNA to match it. The whole thing must be swimming in a soup of U's, A's, C's, and G's, to give the RNAP II material to work with, but drawing these in would have gotten in the way.

1. We're talking about these processes like they happen one step at a time, but that's not the way nature usually works. In any living cell, these processes are happenening at the same time, almost on top of each other. Any gene can have lots of different RNA copies being made at any moment. A chromosome can have thousands happening at the same time.

Here's a place to mention an aspect of DNA structure you haven't heard yet. You may have wondered which strand of the DNA carries the message. The answer is that it varies, which seems to be the universal answer to these kinds of biology questions. One strand carries most of the message, but occasionally the other has it, and very occasionally, they both do. At whatever point you're looking, the side that carries the message is called the **sense** strand, and the other is called the **antisense** strand. Just to be confusing, it's the *antisense* strand that gets copied by the RNAP II, since the resulting copy is the inverse of the original. This way, the resulting RNA is an exact match of the sense strand except that all the T's are changed to U's. (In the sketch, you can see that the mRNA sequence matches the visible bit of the sense strand.)

What about RNAP I and III?

There are a RNA polymerase I and III, and the part they play in this drama is that they are used in the synthesis of ribosomes, the little machines that actually make the proteins from the pattern in the DNA. We'll hear more about the ribosomes in a couple of pages.

The RNAP II proceeds down the DNA, collecting RNA bases, and attaching them to the strand in progress. Each C on the DNA strand attracts a G on the RNA, each G gets a C, each T gets an A, but each A gets a U, since there is no T in RNA.

The copying proceeds until the RNAP II encounters one of the stop codons. At this point, the RNA strand is complete and almost ready to be sent to a ribosome to use for some protein. That is, it is almost what is called messenger RNA, or mRNA, but it must be edited a little bit.

But isn't the DNA wound around into chromosomes?

For organisms whose DNA is arranged in chromosomes—most of us who aren't bacteria—there's an extra step. As the RNAP II moves down the DNA, the nucleosomes, the plugs of protein around which the DNA is wound, must be removed and replaced behind it. There is another complex of proteins that works with the RNAP II for this job.

Editing - Fixing the Copy

Once the chain of RNA is completed, it must have a "cap" stuck onto its beginning, and a "tail" stuck on its end. The cap is a guanine molecule modified so that the enzymes that break up used RNA won't attack it, and the tail is a chain of adenines.

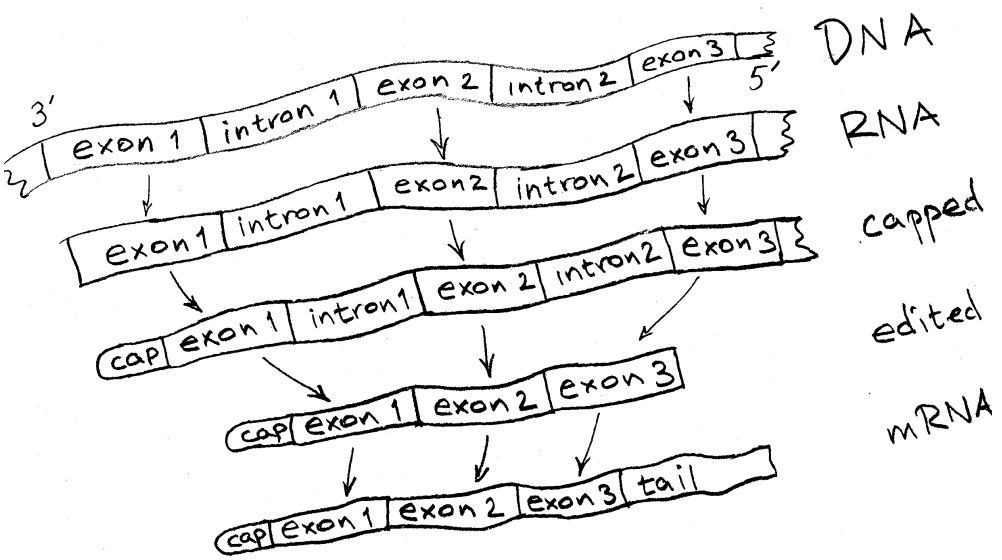

Figure 1-4. DNA is copied to RNA, which is capped, edited, and "polyadenylated" to make mRNA

In eukaryotes, the important editing step is the removal of the **introns**, and piecing together the **exons**. Introns are parts of DNA that don't code for any amino acids, and so don't specify any part of the protein to be constructed. So they have to be removed. (Prokaryotes like bacteria don't usually have this problem, since they have few introns.) Genes coding for some protein can be split up into dozens of separate exons. The gene for dystrophin, which is mutated in boys with muscular dystrophy, has 79 exons, for example. An average human exon is around 40-50 codons long (140-150 nucleotides), while some introns can contain hundreds of thousands of nucleotides.

Exons

Exons were named when it became clear that some RNA left the nucleus. Scientists noticed that not all the genes left, so they gave the name to the particles that *ex*ited. Introns are the parts that stay behind. It's thought that whatever function the introns have—assuming they have any at all—must take place in the cell nucleus.

Special enzymes called collectively called the **spliceosome** are made out of units of small nuclear RNA (snRNA) combined with proteins into a small nuclear riboprotein (snRNP, pronounced “snurp”), edit the introns out. A team of five snurps comes along and works together (with some protein factors) to snip out the unnecessary parts of the RNA chain. The parts that get snipped don't have the right cap and tail, so other enzymes lying around disassemble them into single bases, leaving them ready to be linked together in some other RNA construction project.

The RNA left behind after this process is the messenger RNA, and it is ejected from the nucleus out into the cell's cytoplasm, to be “translated” into a chain of amino acids, which will be turned into some protein. We’ll look at that step next.

Summary

Protein-encoding genes have the following parts:

- Exons - The parts of the gene that define a protein. In eukaryotes, most genes are made from several exons (often dozens).
- Introns - The opposite of an exon. These are the interspersed parts of the gene that do not encode for a protein. They must be removed from the messenger RNA (mRNA) before it is used for translation.
- Start - A mark that indicates where the gene starts
- Promoters - Sites onto which proteins bind to promote the transcription of the gene. These are both "basal" promoters near the gene, and "upstream" promoters, up to a few hundred base pairs away.
- Enhancers and Silencers - Sites onto which proteins bind to speed up or slow down (or stop) the transcription. These may also be quite close to the gene in question, or thousands of base pairs away.

Messenger RNA (mRNA) comes in a single strand, and has the following features:

- Cap - A modified guanine molecule that marks the beginning of the RNA chain.
- Exons - These coding stretches of the base sequence are all that remains, and the introns, the non-coding bits, have been removed.
- Tail - A series of adenine molecules marks the end of the RNA chain. Biologists say the chain with a new tail has been "polyadenylated," largely as a way to show off their command of Greek.

Ribosomes - Protein Machines

The mRNA specifying some chain of amino acids is made and is pushed out of the nucleus into the rest of the cell. What then? Well, the first step is that it has to float over near a ribosome, which is where the protein-making action happens. It may do more than "float," but—like quite a number of the details in this process—the transport within a cell is still kind of mysterious.

However it works, eventually the mRNA gets itself over to some ribosome, or a ribosome gets over to it. Compared to a strand of mRNA, a ribosome is a pretty big thing, but they're not big compared to other organelles, like mitochondria or chloroplasts. A cell can contain lots of ribosomes: thousands in regular cells, and as many as 200,000 times more in egg cells.

A ribosome is composed of proteins, and another kind of RNA called **ribosomal RNA** (rRNA). Molecules of rRNA are relatively short chains, wound into a loop or two, and used to hold the mRNA in place while it's being copied.

Ribosomal RNA

The rRNA seems to come in several forms. They have funny names like 18S, 28S, 5.8S and 5S. The "S" stands for "Svedberg units", which is just a way to identify where things wind up after they've been mixed up and put in a centrifuge for a while. The low numbers are the lightest, and wind up on the top. The kinds of rRNA are named for the bands where they were first observed.

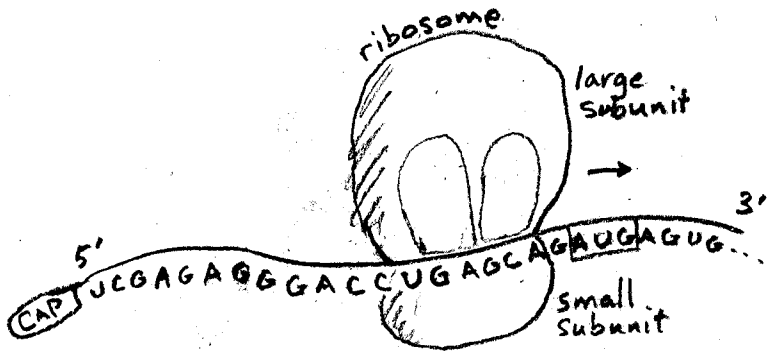

Figure 1-5. A ribosome holding a strand of mRNA. You can see two indentations. These represent the sites waiting for loaded pieces of tRNA to land on them. The ribosome is moving to the left, looking for the start (AUG) where it will wait.

A ribosome is made up of two subunits, and for some reason they're called the "small subunit" and the "large subunit." The small subunit is made of about 30 different proteins,

and a strand of 18S rRNA. The large subunit consists of around 45 different proteins, in combination with one molecule each of 28S, 5.8S or 5S rRNA.

When the mRNA floats by, the ribosome divides itself into its large and small subunits, the mRNA attaches itself to the small subunit of a ribosome, on the 5' side of the START codon. The ribosome moves itself down the molecule (in the 3' direction) until it gets to the start, where it waits for an appropriate piece of transfer RNA to wander by. The start of the gene isn't necessarily the start of the mRNA. The start of the mRNA was signalled by the position of the transcription factors, which mostly weren't looking at the sequence of the gene.

Where does a gene start?

The start of a gene is a nebulous concept. You could consider the start of a gene the position marked by the TATA box, which is where the RNAP II begins transcribing the DNA. Or you could say that the beginning is at the "start" codon (AUG), which is where the ribosome begins translating. (Or you could also include all the transcription factor sites in your definition of the gene, which is what some people do.) Either way, there is some mRNA before the start codon that gets ignored in translation.

Transfer RNA - Decoding the Gene

Hanging around the ribosome is a cloud of small molecules called transfer RNA (tRNA). These are chains of about 60-80 base pairs, arranged in three loops.[2] These are where the real work of decoding the mRNA takes place. One end of the tRNA has three active nucleotides that fit into a codon of the mRNA (this is called the amino acid's **anticodon**), while the other end catches the amino acid specific to that codon.

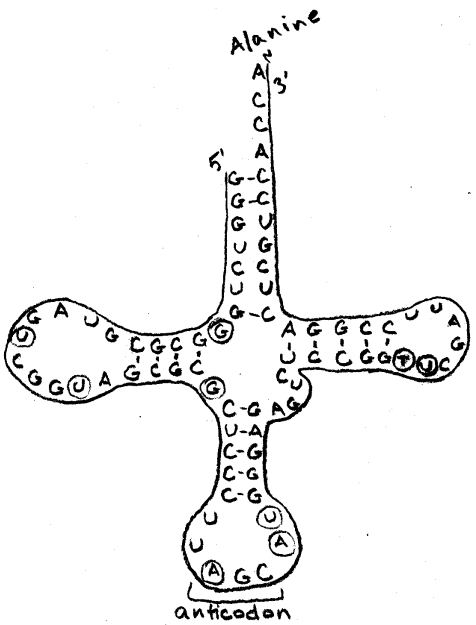

Figure 1-6. A schematic diagram of tRNA. The 3′ end attaches to a molecule of alanine, and the other end (the "anticodon") mates with the RNA sequence for that amino acid: GCU. The circles around some of the bases indicate that they are modified. For example, the A in the anticodon part is changed enough so it won't really matter in the mating with the mRNA, allowing the anticodon to mate with GCU, GCC, GCA, and GCG.

For example, the codon GCU means alanine, so there is a tRNA molecule that bonds to proline at one end, and has a CGA at the other end. (It's flipped around to read AGC in the figure, but it's the same thing.) The CGA matches with the GCU, and so delivers the alanine to the growing end of the protein chain under construction. The table here shows the RNA code.

2. We draw the tRNA laid out flat, in goofy-looking, but relatively neat loops. But inside a cell, remember that this molecule, like any protein, is coiled up into a shape that, while a little on the lumpy side, isn't nearly as ungainly looking as the loopy cross in the figure.

Table 1-1. RNA Codons - In case you're wondering, the code looks like this. Notice that there are four non-coding codons. One signals the start of the sequence to be translated, and three of them signal the end to translation. (The "Start" codon can also be used for methionine.)

	U	C	A	G	
U	UUU Phenylanine (Phe)	UCU Serine (Ser)	UAU Tyrosine (Tyr)	UGU Cysteine (Cys)	U
	UUC Phe	UCU Ser	UAC Tyr	UGC Cys	C
	UUA Leucine (Leu)	UCA Ser	UAA STOP	UGA STOP	A
	UUG Leu	UCG Ser	UAG STOP	UGG Tryptophan (Trp)	G
C	CUU Leu	CCU Proline (Pro)	CAU Histidine (His)	CGU Arginine (Arg)	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Glutamine (Gln)	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Isoleucine (Ile)	ACU Threonine (Thr)	AAU Asparagine (Asn)	AGU Serine (Ser)	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lysine (Lys)	AGA Arginine (Arg)	A
	AUG Methionine (Met) or START	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Valine (Val)	GCU Alanine (Ala)	GAU Aspartic Acid (Asp)	GGU Glycine (Gly)	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glutamic Acid (Glu)	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

When last we left our ribosome, its small subunit had advanced to the start codon, and was patiently waiting there, for the combination of the large subunit of the ribosome, and a piece of initiator tRNA. This is a piece of tRNA that will match the AUG of the start codon. In eukaryotes, the initiator tRNA carries a methionine molecule, which sits at the beginning of the protein chain to be built. In prokaryotes, the initiator carries a modified methionine, labeled fMET, which doesn't.

Elongation - Building the Chain

Now there's a ribosome, a piece of initiator tRNA, and the mRNA all together. The spot on the ribosome where the initiator sits is called the "P" site. Bound together like this, they will bind to another piece of tRNA (in the ribosome's "A-site"), only under two conditions:

* The new piece of tRNA must be loaded with its characteristic amino acid, and

* The new piece of tRNA must match the next codon in the 3' direction on the mRNA.

When this new piece of tRNA settles into place (aided by more protein factor molecules), the partially-built protein chain grabs onto the amino acid on its back, and lets go of the tRNA it was attached to. That tRNA now falls off the ribosome, and the new tRNA moves into its place, opening up its spot for another piece of loaded tRNA.

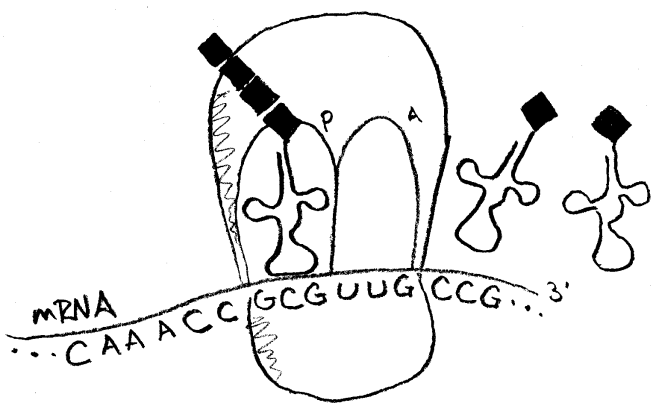

Figure 1-1. The chain elongation three-step: a loaded tRNA moves into the ribosome's A-site (it's the kind that carries leucine, to match the UUG below). The P-site already has a tRNA holding onto the end of a chain-in-progress.

Here's another way to look at it. Suppose you're an Alanine tRNA. You float about the cell, and when you find a free molecule of alanine, you grab it and then look for a ribosome with an A-site open over a matching codon (this would be anything that starts with GC). You move into that A-site, and the occupant of the neighboring P-site hands you its partly completed peptide chain. You stick it onto your alanine, and then move over to the P-site, pushing the empty tRNA out of the ribosome. Now together, you and the ribosome wait for another tRNA to float into the A-site, when the process begins again.

The process continues, with the chain of amino acids lengthening (this is also called a "peptide chain"), until the ribosome encounters one of the three STOP codons (UAA, UAG, or UGA). There are no tRNA that can bind to one of these codons, but there is yet

another protein factor that acts to release the mRNA, split the ribosome back into its sub-units, and free the new peptide chain. And now the process can begin again, lengthening the chain until the ribosome hits one of the stop codons, when the chain breaks free.

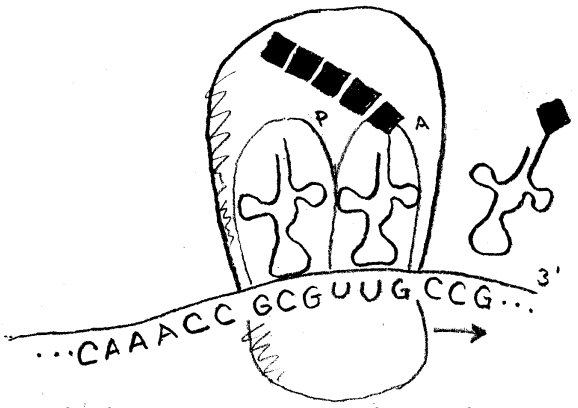

Figure 1-2. The peptide chain-in-progress attaches to the amino acid brought by the new tRNA, and the whole ribosome moves one codon down the chain.

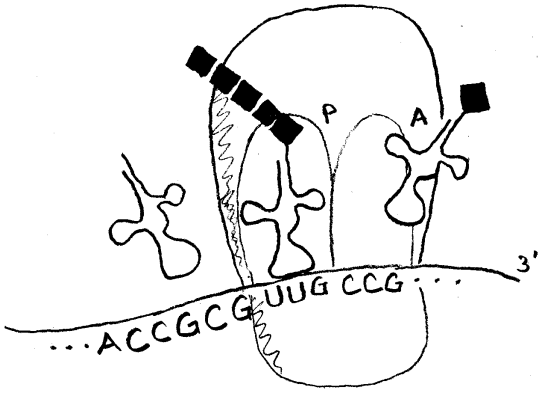

Figure 1-3. The now-unloaded tRNA gets ejected from the ribosome, and the A-site is cleared for a new tRNA loaded with proline (matches CCG).

Ribosome P's and A's

The two spots on the ribosome are called the "P" and the "A" sites. The P site can only be occupied by a tRNA attached to the end of an peptide chain-in-progress, which is how it got its "P" (except for the initiator tRNA, which gets a special pass), and the A site can only be occupied by a tRNA carrying a single amino acid, which is how it got its "A."

Timing is everything

After the third step in the elongation of the peptide, we said, “the process can begin again.” Actually, it's more likely that the process has already begun again. It’s easier to describe these processes as if they are like steps in a recipe. In a way they are, but it’s a huge team of impetuous cooks who follow those recipes: always getting on top of each other, grabbing each other’s ingredients, and starting again before the process is even half over.

For cxample, we just described translation, as if each mRNA was processed by a single ribosome. In reality, the usual case is that many ribosomes are busy at the same time on a single piece of mRNA, looking like beads on a long string. The group all together, of a single mRNA with a bunch of ribosomes working along it, and peptide chains of various lengths hanging off the side, is called a polysome. This was once thought to be a single fabulously complex kind of molecule, before people realized that it was just a phase in a process.

The accompanying figure shows an E. Coli chromosome in the process of being transcribed into RNA, and the RNA being translated into proteins. In the picture you can see that it’s all happening at the same time.

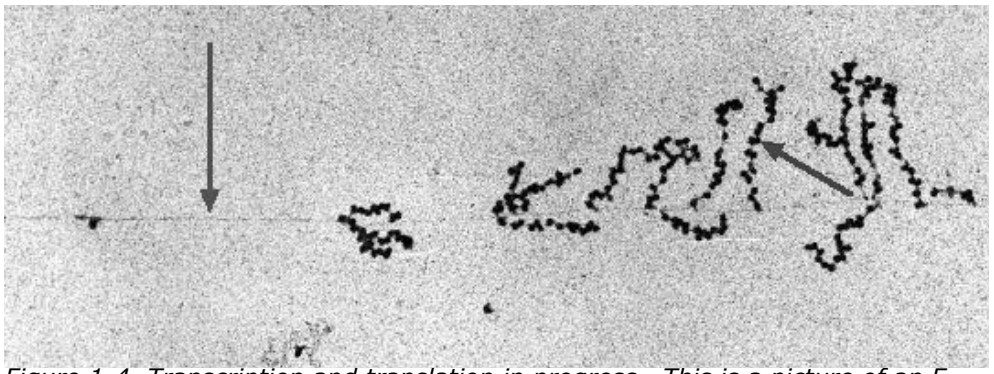

Figure 1-4. Transcription and translation in progress. This is a picture of an E. Coli chromosome (the thin horizontal line indicated by an arrow) in the process of being transcribed. The strings of beads to the right are ribosomes on not-yet completed pieces of RNA (polysomes). That is, the RNA is in the process of being translated before it’s finished. (Miller, et al.)

Summary

Transfer RNA (tRNA) has:

- A series of around 60-80 RNA bases,
- A cross-shape, with three loops,
- An anti-codon to match the mRNA on one end and
- A site for an amino acid to bind on the other end.

Ribosomes have the following features:

- They have two parts called the “large subunit” and a “small subunit.”
- The combination of ribosomes and mRNA has two sites that fit tRNA molecules. Called the “P-site” and the “A-site,” both of them can accommodate a tRNA molecule. tRNA enter on the A-site, and move to the P-site when the P-site occupant leaves.
- In the making of an amino acid chain, the loaded tRNA moves into the ribosome, grabs the protein-in-progress from the tRNA already there and binds it to its own amino acid. It then pushes the now-unloaded tRNA out of the ribosome, and waits for the next loaded tRNA.
- Initiator tRNA is the only tRNA that can bind to a ribosome without a protein chain already in progress.
- tRNA needs an amino acid to enter the ribosome A-site.

What's Missing from this Story?

This process seems absurdly complex, as if it could barely work once, and yet it works zillions of times each day in you, and zillions of times in everyone else, too. (Not to mention all the E. Coli, groundhogs, plankton and poison ivy, too.) What's more amazing, though, is that this account really just scratches the surface of what's going on. For example, very little in the body just "happens" spontaneously. Usually there has to be some energy put into the system to get things to work, though sometimes it's not necessary. Here are some of the things we've left out. Many of them we'll get back to in the course of the rest of the book. To learn more about some, you may need to find a course in molecular biology. If you learn enough about some of them, you may earn a Nobel prize.

- Energy budgets. Many of the steps require energy to work.
- What happens when things go wrong? How does the system correct errors, and how does it keep from making them in the first place?
- Gene regulation. How does your body decide it needs to make this protein or that one? How is that decision passed down from a cell to its daughter cells or passed on at all, to any other cell?
- If enzymes are necessary for final protein assembly and for so many of the steps of the process, and if the DNA is used to make those enzymes, then where did the enzymes come from in the first place? If they didn't come from the DNA in the first place, how did their patterns get into the DNA?