Chapter 3: Information Storage and Transfer in Life

The trapped scientist examples are great for conceptual purposes, but they do not accurately model how information in life changes because they do not take into account the fact that amino acid changes are caused by changes in DNA. This chapter will explore how DNA stores information, and how this information is used to build proteins. It will also explore how mutations change this information.

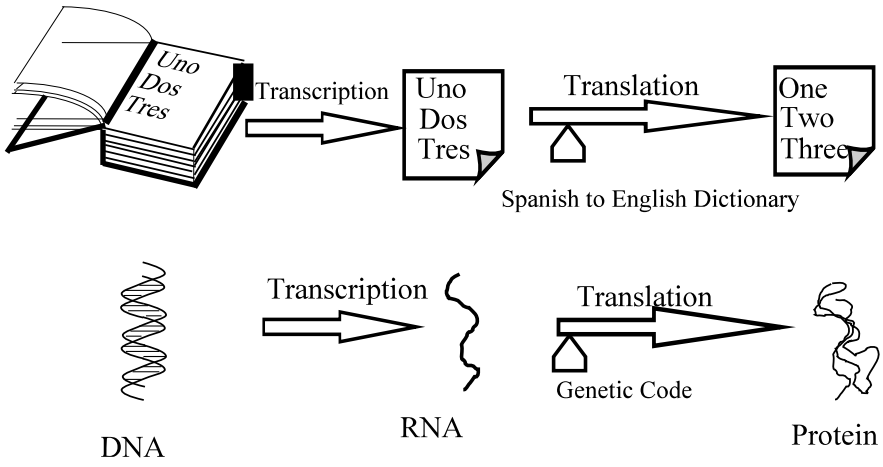
The language that life uses to store and transmit information is similar to human languages, but the rules of grammar and the vocabulary are much simpler. Only 20 words are used by life, so the vocabulary is very limited. Punctuation is limited to capitalization and periods. Every sentence must start with the same word.

This chapter will start with a simple system based on coin tosses and show how coin tosses can be used to store and transmit information. This simple example will then be improved by using a four sided coin. The trapped scientist will then be used with a new set of rules to show how information changes in life.

Transcription and Translation

Suppose that the phrase uno, dos, tres is found in a Spanish book. The process of transcription copies a page from this book onto a piece of paper. Translation then requires a person with a Spanish to English dictionary to copy the phrase onto another piece of paper in English (figure 3.1). Life uses both transcription and translation. The message in the DNA corresponds to the written words in a Spanish book (uno, dos, tres). This message is copied through transcription to create an intermediate message called RNA. RNA is analogous to the piece of paper with the phrase uno, dos, tres. The final piece of paper with the phrase one, two, three is analogous to a protein.

Figure 3.1: Transcription and Translation



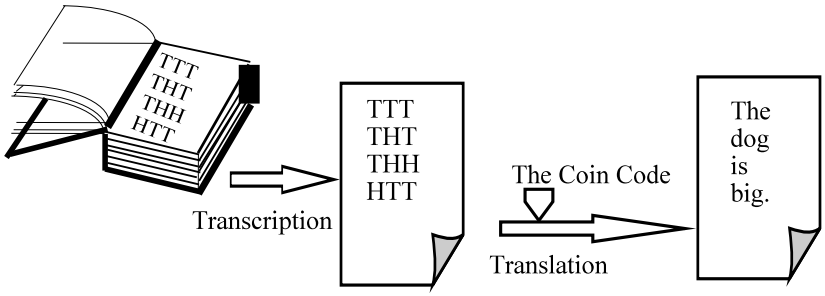
Information Using Coins

Coins have two possible outcomes when tossed, heads and tails. So to store and transmit more than two possible messages, the results of tossing a coin must be grouped. If coins are tossed and read three at a time, then eight messages can be assigned to 3 coins.

The code is important because it assigns the messages. The code performs the function of the Spanish to English dictionary in figure 3.1. Consider the following code, and how it can be used to create a message (figure 3.2).

Coin 1	Coin2	Coin 3	Message
T	T	T	The
T	T	H	Cat
T	H	T	Dog
T	H	H –coin code---->	Is
H	T	T	Big
H	T	H	Small
H	H	T	Tall
H	H	H	Short

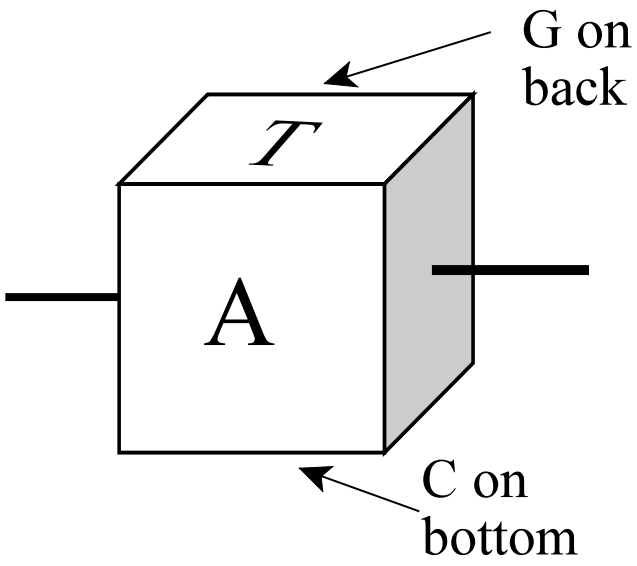
Figure 3.2: How Coins Can be Transcribed and Translated



A Four Sided Coin

Life uses the equivalent of a four sided coin. To imagine this coin, visualize a playing die. Instead of numbers on each side, each side has a single letter. The letters are A,T,C and G. Since dice have 6 sides, this die has to be modified. This is easy to do. Imagine a long pin inserted into one side and out the other. This pin only allows the die to land on the 4 sides with letters. See figure 3.3.

Figure 3.3: Four Sided Die (Coin) with the Letters A, T, G, and C



If this altered die is used as money, then it is just as good as any other 4 sided coin. But now instead of each coin having two possible outcomes, head or tails, each coin has 4 possible outcomes, A, T, G and C. Thus, each coin now contains 2 bits of information because 2^2 =4 (equation 1, p24).

If these special coins are grouped three at a time, how many possible outcomes are there? Each coin contains 2 bits of information. Thus, 3 coins must contain 6 bits (2+2+2=6). There are now 64 possible outcomes because 2^6 =64. This means that life could potentially have at its disposal 64 words, but life only needs 20 words. So some outcomes are assigned the same word.

The Four Sided Coin Code

The four sided coins are always tossed and read 3 at a time. The following table assigns a word to each possible outcome. Since only 20 words are used, some outcomes are assigned to the same word. The period in the parenthesis is not a word, but indicates the end of a sentence. All sentences must start with the word, *the*, and this is why it is capitalized. The table is like a Spanish to English dictionary. It assigns a message to an outcome.

Table 3.1: The Coin Code

Outcome	Message
CCA, CCG, CCT, CCC	dog
CGA, CGG, CGT, CGC	cat
CAA, CAG, CAT, CAC	bird
GAA, GAG, GAT, AAC, GAC, AAT	animal
TAA, TAG, TAT	is
GGA, GGG, GGT, GGC	runs
ACC	sleeps
AAA, AAG	under
TAC	The (Capitalize)
AGA, AGG, AGT,AGC, TCA, TCG	lazy
TGA, TGG, TGT, TGC	quick
GAT, GTG	big
GTT, GTC	tall
CTT, CTC	small
CTA, CTG	tiny
TTA, TTG	tree
TTT, TTC	buried
ACA, ACG	home
ATA, ATG	white
GCA, GCG, GCT, GCC, TCT, TCC	black
ATC, ACT, ATT	(Period)

Using this table, it is possible to construct many sentences. Consider the following outcomes and the associated messages. These messages are constructed by tossing the special coin 15 times and then grouping the results into triplets. Each triplet is called a codon.

TAC-CCA-TAA-AGA-ATC => The dog is lazy.
TAC-CGA-TAA-GAT-ATC => The cat is big.
TAC-GAA-TAA-CTG-ATC => The animal is tiny.

Many more nonsense messages are possible.

TAC-ATA-GCA-CTA-ATC => The white black tiny.
TAC-ATA-TAA-CTA-ATC => The white is tiny.

Also notice that the reading frame is very important. Consider the following message:

TAC-CCA-TAA-AGA-ATC => The dog is lazy.

If the first *T* is deleted, then the message becomes ACC-CAT-AAA-GAA-TC => sleeps bird under animal. This sentence is nonsense. To get the correct message the letters must be read three at a time, and the reading frame must be correct. Shifting the frame changes how the letters are grouped. Thus, it changes the words, and very simple change like deleting a single letter can have far reaching consequences. Because such mutations change the reading frame, they are called frame shift mutations. In contrast, point mutations only change a single base and thus preserve the reading frame.

The Information in DNA is Similar to the Four Sided Coin

Life uses a system very similar to the four sided coin, but instead of coins and dictionaries, life uses chemicals. DNA looks like a twisted ladder. The steps of the ladder are composed of four chemicals, adenine, thymine, cytosine and guanine. These four chemicals correspond to the four letters on the four sides of the four sided coin. That is adenine is A, thymine is T, cytosine is C and guanine is G. The one letter abbreviations for these chemicals will be used from now on. The side of the ladder is composed of a chemical called deoxyribose.

Notice that the steps are A-T, T-A, G-C or C-G, where the dash represents a chemical bond between the chemicals. Adenine always forms a chemical bond with thymine, and guanine always forms a bond with cytosine. A-T is called a base pair as is G-C. There are four possibilities for each step, so each step can hold the same amount of information as the four sided coin or 2 bits.

Figure 3.4: Untwisted DNA Ladder

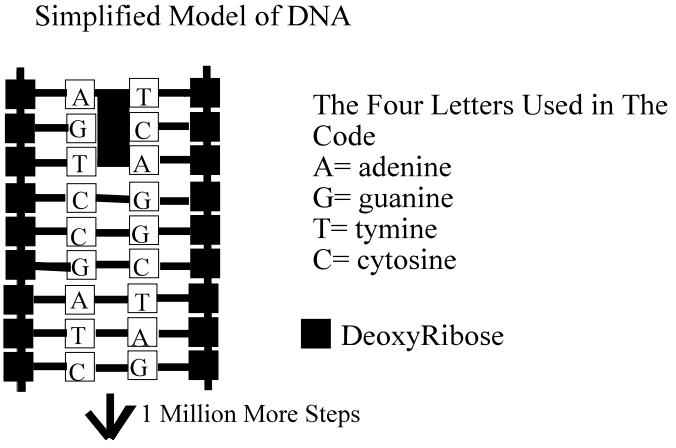
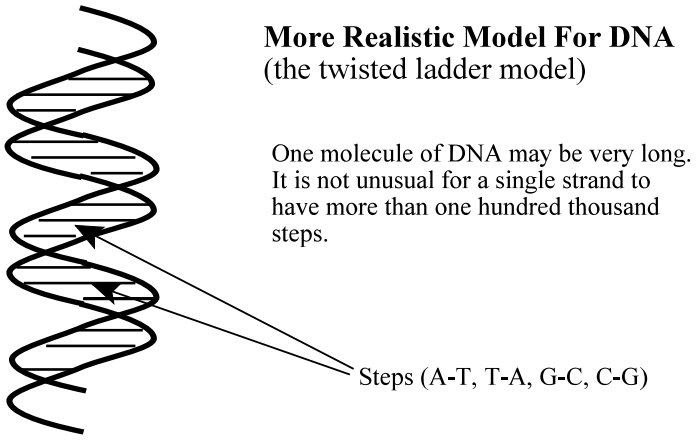


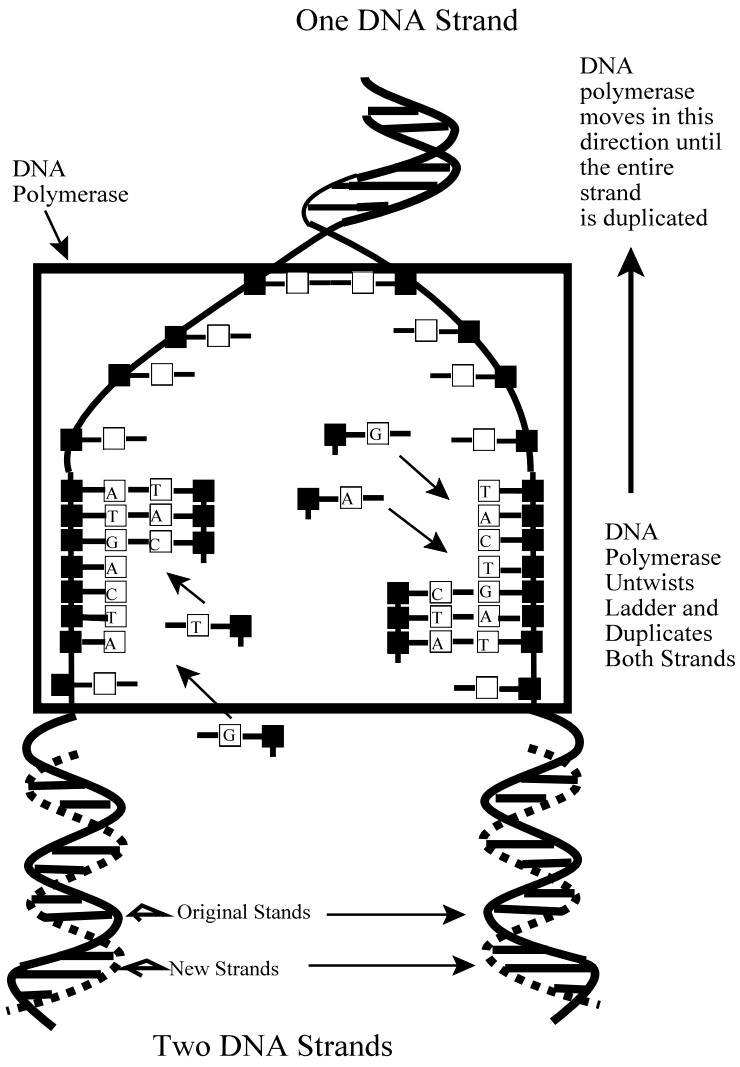
Figure 3.5: Twisted DNA Ladder



DNA Replication

In order to replicate, the DNA molecule must be untwisted, and the chemical bonds between the base pairs must be broken. This process is controlled by many different proteins. In the first step, a protein binds to the DNA targeting the site to be replicated, another protein untwists the DNA breaking the chemical bonds between the base pairs, other proteins keep the base pair bonds from reforming, and a protein called DNA polymerase replicates the untwisted sections. The two original strands serve as templates (solid lines in figure 3.6) for the growing strands (dashed lines). Figure 3.6 shows how DNA polymerase replicates each strand, creating two DNA molecules from one. Figure 3.6 is simplified in that most of the proteins involved in DNA replication are not shown.

Figure 3.6: DNA Replication



During DNA replication, sometimes the bases pair incorrectly. That is sometimes adenine pairs with guanine, and thymine pairs with cytosine. Proofreading corrects most of these errors. Thus, DNA replication is very accurate and mistakes are rare, but mistakes happen. Most mistakes just create variability, but some can create information.

The Genetic Code

The coin code in table 3.1 applies to the four sided coin. The genetic code works in the same way, except instead of words the code specifies amino acids. The coin code groups the results of the coin tosses into groups of three. The genetic code does the same, except the letters are now chemicals. A group of three bases is called a codon. Each codon specifies an amino acid. For example, with the coin code:

TAC-CCA-TAA-AGA-ATC should be translated as follows: The dog is lazy.

but with the genetic code (see table 3.2)

TAC-CCA-TAA-AGA-ATC should be translated as follows: methionine-glycine- isoleucine-serine.

So the coin code specified words, and the genetic code specifies amino acids. Table 3.2 lists the genetic code, and figure 3.7 shows the process of transcription in which DNA is used to create RNA. Notice that in figure 3.7 the coding DNA strand creates a complementary RNA strand (G is replaced with C, C is replaced with G, T is replaced with A, and A is replaced with a new chemical unique to RNA - uracil or U for short).

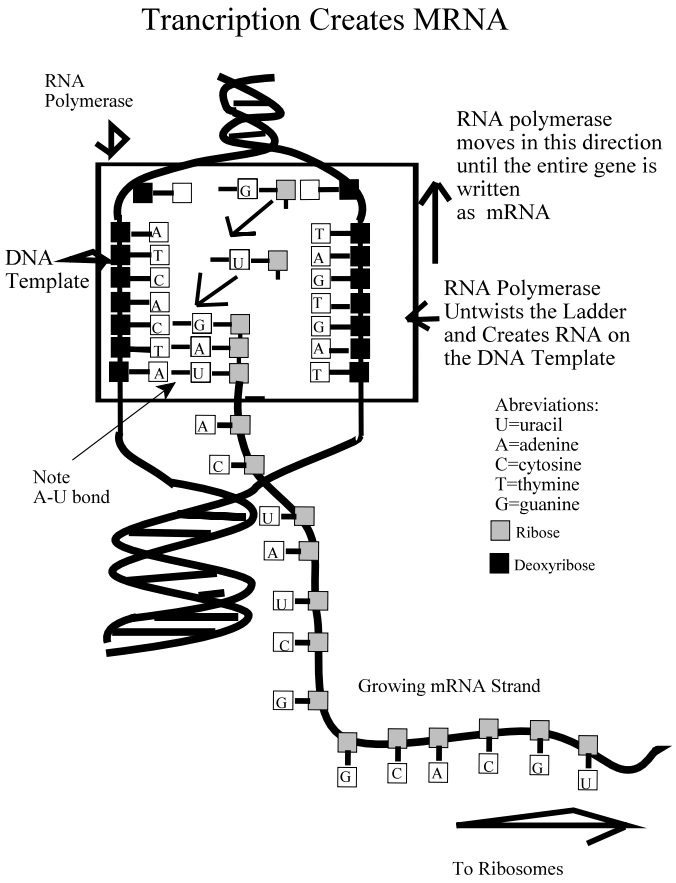
Table 3.2: The Genetic Code

DNA Codon (Coding Strand)	Corresponding RNA Codons (Transcription)	Amino Acid (Translation)
CCA, CCG, CCT, CCC	GGU, GGC, GGA, GGG	glycine
CGA, CGG, CGT, CGC	GCU, GCC, GCA, GCG	alanine
CAA, CAG, CAT, CAC	GUU, GUC, GUA, GUG	valine
GAA, GAG, GAT, AAC, GAC, AAT	CUU, CUC, CUA, UUG, CUG, UUA	leucine
TAA, TAG, TAT	AUU, AUC, AUA	isoleucine
GGA, GGG, GGT, GGC	CCU, CCC, CCA, CCG	proline
ACC	UGG	tryptophan
AAA, AAG	UUU, UUC	phenylalanine
TAC	AUG	methionine
AGA, AGG, AGT,AGC, TCA, TCG	UCU, UCC, UCA, UCG, AGU, AGC	serine
TGA, TGG, TGT, TGC	ACU, ACC, ACA, ACG	threonine
GAT, GTG	CUA, CAC	histidine
GTT, GTC	CAA, CAG	glutamine
CTT, CTC	GAA, GAG	glutamate
CTA, CTG	GAU, GAC	aspartate
TTA, TTG	AAU, AAC	asparagine
TTT, TTC	AAA, AAG	lysine
ACA, ACG	UGU, UGC	cysteine
ATA, ATG	UAU, UAC	tyrosine
GCA, GCG, GCT, GCC, TCT, TCC	CGU, CGC, CGA, CGG, AGA, AGG	arginine
ATC, ACT, ATT	UAG, UGA, UAA	Stop

Transcription

Transcription copies the information in DNA into RNA as shown in figure 3.7. Note that RNA is a single strand, whereas DNA is double stranded. Also note the RNA does not contain the chemical thymine. Another chemical called uracil replaces it. Thus, an A-U bond (adenine-uracil) is formed between the DNA and RNA strand (as opposed to adenine-thymine). The enzyme RNA polymerase directs the synthesis of RNA.

Figure 3.7: Transcription

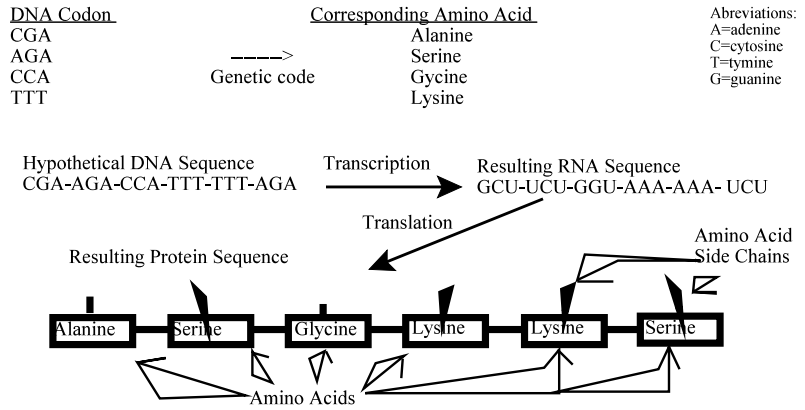


Each Codon Specifies an Amino Acid

Each codon in a gene specifies a particular amino acid. Therefore, a gene determines the chemical properties of a specific protein by specifying its amino acid sequence. A simplified example is shown in figure 3.8. Four DNA codons and the amino acids that they specify are depicted in this picture. During transcription of the hypothetical DNA sequence, the DNA is written into a corresponding messenger RNA (mRNA). During translation, mRNA is translated using the genetic code to create a new protein. The code assigns an amino acid to each codon.

The properties of amino acids are determined by their side chains. In figure 3.8, these side chains branch from the main protein chain. The side chains of amino acids give them unique chemical properties. 1) Some side chains are chemically reactive. These side chains are used by proteins to interact with other chemicals. 2) Other side chains do not like to dissolve in water. Such side chains are called hydrophobic. Hydrophobic side chains cause proteins to fold up into complex three-dimensional shapes.

Figure 3.8: Translation



To verify understanding, all readers should make sure that they can use table 3.2 to follow how the hypothetical DNA sequence in figure 3.8 is transcribed into the resulting RNA sequence, and how this messenger RNA sequence is translated into the final amino acid chain.

Translation

Messenger RNA is translated at ribosomes. Figures 3.9-3.12 illustrate this process. A special type of RNA known at tRNA recognizes amino acids and brings them to the ribosome. The tRNA matches up its own sequence of 3 bases to the codon in the messenger RNA.

Figure 3.9: Transfer RNA Brings and Aligns Amino Acids

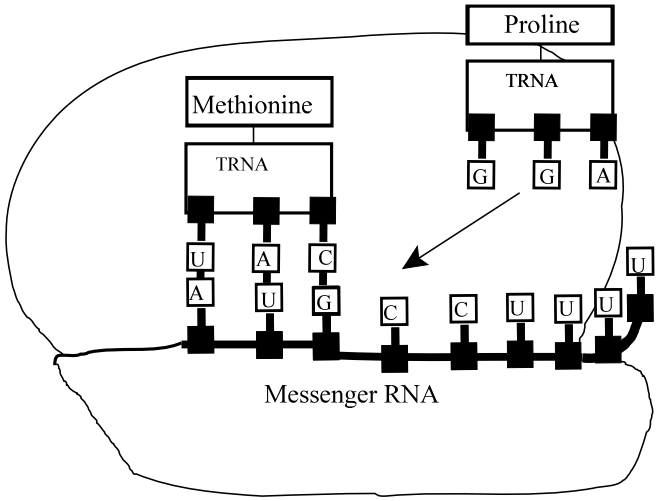


Figure 3.10: A Peptide Bond Forms between the Amino Acids

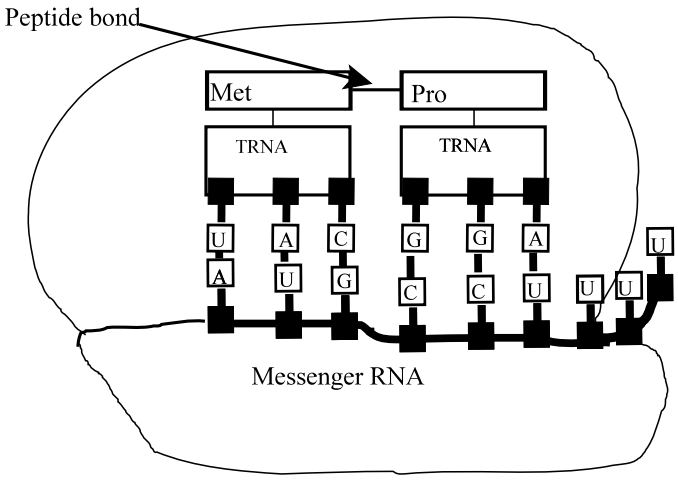


Figure 3.11: Ribosome Shifts Along Messenger RNA

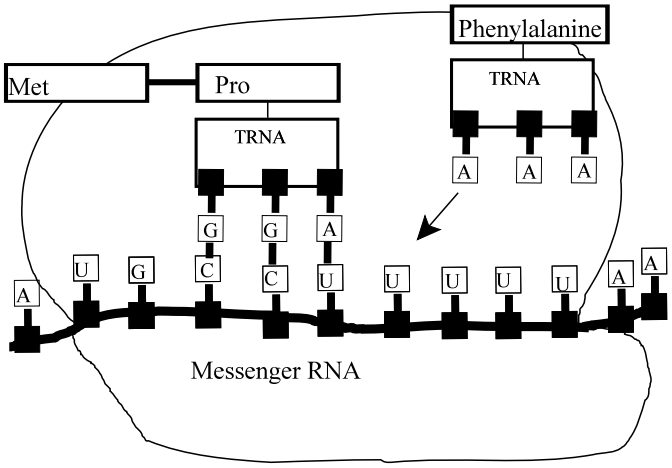
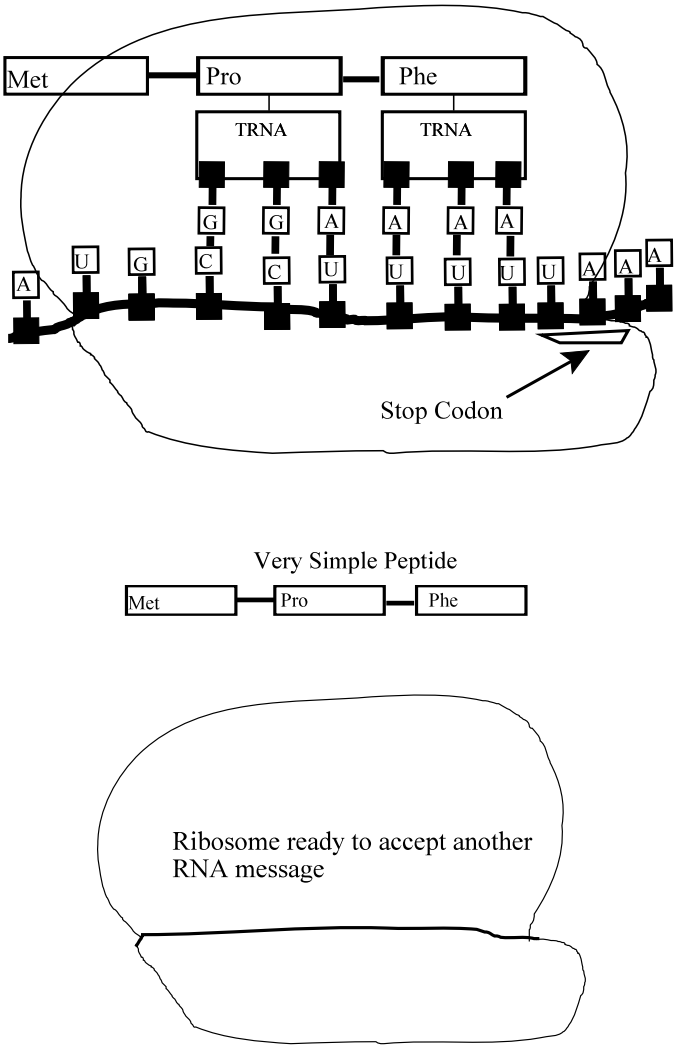


Figure 3.12: Second Peptide Bond Forms and Cycle Starts Over



Proteins

Proteins are the most versatile chemicals found in life. Proteins are so important because they are involved in everything that life does. They implement the knowledge contained in DNA. Some of the primary functions of proteins are listed below.

A special kind of protein called an enzyme regulates chemical reactions. Chemical reactions that would take years without enzymes proceed in fractions of a second with enzymes. Enzymes make life possible. Teams of enzymes working together enable cells to synthesize all sorts of complex chemicals.

Proteins have additional functions as well. Some proteins regulate genes. Others control which chemicals can pass though the cell membrane, and still others are responsible for movement (muscles are composed of two proteins, actin and myosin). The list does not stop here. Some proteins can transport other chemicals. Hemoglobin is the blood protein that transports oxygen. Proteins also serve as signals. For example, insulin signals cells to take up sugar from the blood stream. Diabetes results when this process does not function properly.

Proteins are very versatile molecules, and it is this versatility that allows proteins to implement the knowledge stored in DNA.

Protein Folding

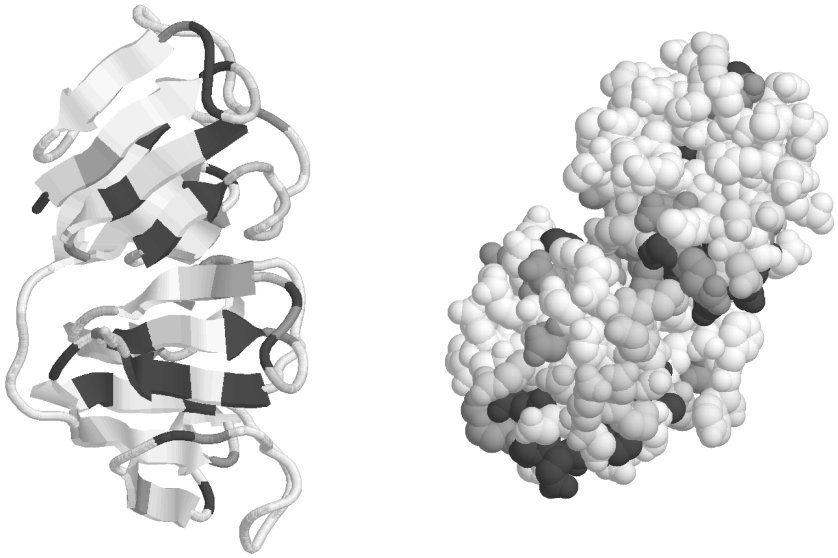
Proteins fold into complex 3 dimensional patterns. The amino acid sequence of a protein determines this pattern. The amino acids that do not like water tend to cluster in the protein's core where water is excluded. Amino acids that like water are typically found on the surface of the protein where they can interact with water. A random amino acid sequence will rarely fold properly into a compact 3-D shape. So the amino acid sequence is very important to ensure that a protein folds properly.

Some segments of a protein may form helixes. Others form sheets. Because of the complexity, the structure of proteins is best illustrated with cartoons. Rather than drawing the amino acids that form the helix, the cartoon model just draws the helix. The sheets are drawn as straight lines with arrows on the end. The protein bacterial rhodopsin is shown in figure 3.13. Notice the helixes. Also note that black and gray are used to represent amino acids that contain information.

Figure 3.13: Bacterial Rhodopsin

The protein that composes the eye lens in mammals is called crystallin. It tends to form sheets instead of helixes. Its cartoon representation is shown in figure 3.14. Again, black and gray represent amino acids that carry information. Figure 3.14 also shows crystallin using the space filling model for the atoms (right side). That is all of the atoms in the protein are represented by spheres. It is very hard to visualize the structure of a protein using this model. This is why the cartoons are so useful. When viewing these structures, keep in mind that they are formed by hundreds and sometimes thousands of amino acids.

Figure 3.14: Carton and Space Fill Model for Crystallin

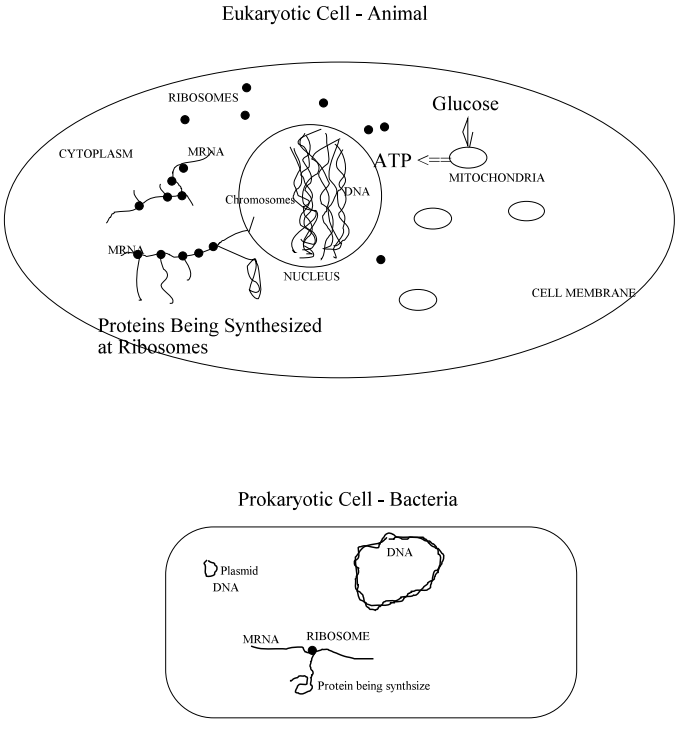


The top right picture on the back cover shows bacterial rhodopsin in color. Purple and shades of purple are regions rich in information. This image was created by applying a script generated by the CONSURF program to the protein data bank file and then viewing the results in the RASMOL program (see page 289).

Eukaryotes and Prokaryotes

Figure 3.15 depicts a eukaryotic and a prokaryotic cell. The DNA in eukaryotes is bundled together with proteins to form chromosomes. This DNA resides inside the nucleus which is separated form the rest of the cell by the nuclear membrane. In prokaryotes, there is no nucleus. The DNA is usually one large circular ring, and the cell contains much less DNA. Prokaryotes are very simple organisms such as bacteria. The total information contained in their DNA is an order of magnitude less than the information found in eukaryotes. The mitochondria in eukaryotes converts sugar into ATP - the fuel for the cell. In bacteria, proteins in the cell membrane create ATP.

Figure 3.15: Prokaryotes and Eukaryotes

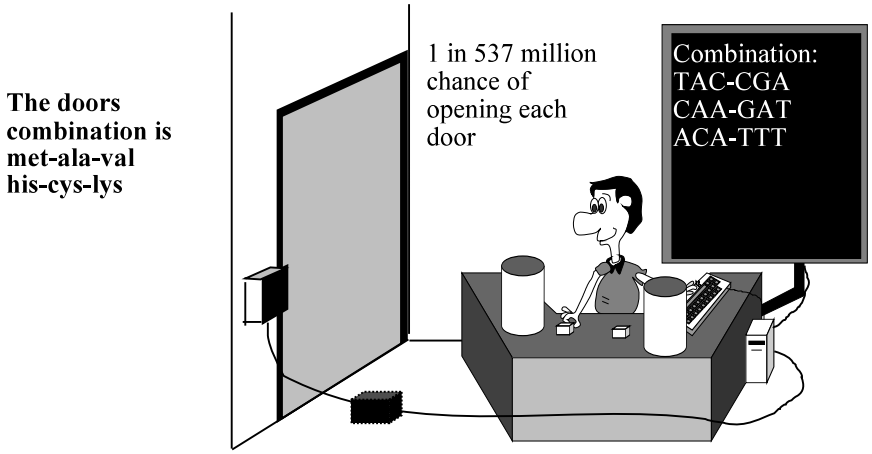


The Accurate Trapped Scientist

The previous trapped scientist models are oversimplified because they model changes to amino acid sequences without considering DNA. Since mutations alter DNA, the trapped scientist really should change DNA.

When considering DNA, another factor will surface. Six codons specify the amino acid, arginine. Only one codon specifies methionine. Three codons specify isoleucine**.** From table 3.2, each amino acid may have 1,2,3,4 or 6 corresponding codons. If the bases A, T, G and C are changed at random, the probability of creating a codon that specifies arginine is much higher than the probability of creating a codon that calls for methionine. Information theory must take this effect into account when computing the information content of a protein. For simplicity, only point mutations that change A, T, C or G will be considered (no insertions or deletions allowed), and all mutations will be considered random.

Figure 3.16: Trapped Scientist Using Genetic Code



The scientist now requires two baskets. In one he has four blocks labeled A, G, C and T. In the other, he has 18 blocks numbered 1 through 18. He is told to pull a lettered block and enter it into the computer. He is to put the block back and repeat this procedure. After he enters 18 letters, he is to press enter to see if the door opens. If the door does not open, he is to draw one lettered block and one numbered block. He is to use the number to find a corresponding letter on the computer screen (the letters on the computer screen are numbered sequentially 1 through 18). After he finds the letter corresponding to the number, he is to replace this letter with the new letter. For example, in figure 3.16, the last letter on the computer screen is *T*. If the scientist draws the number 18 and the letter A, then he should change the letter *T* to an *A*. He should repeat this procedure until he opens the door.

The probability that he will open the door is now a little more complex to calculate. The door will open if the scientist enters any sequence of letters that specifies the combination of the door, methionine-alanine- valine-histidine-cysteine -lysine. Using table 3.2 for reference, there are 64 possible codons that can be typed into the computer. One of these codons specifies methionine. Four of these codons specify alanine and valine. Two codons specify histidine, cysteine and lysine.

Methionine has a 1 in 64 chance of arising by chance. The knowledge it contributes is $2^{(\text{information})} = 64$. Since $2^6 = 64$, methionine contributes 6 bits. Alanine and valine each have a 1 in 16 chance of arising by chance. The knowledge they contribute is thus $2^{(\text{information})} = 16$. Because $2^4 = 16$, each contributes 4 bits. Histidine, cysteine, and lysine each have a 1 in 32 chance of arising by chance. Because $2^5 = 32$, each contributes 5 bits. Thus, the total knowledge required to open the door is $6 + 4 + 4 + 5 + 5 + 5 = 29$ bits. The odds that the scientist will open the door on the first try are 1 in 2^{29} or 1 in 537 million.