1

INTRODUCTION

This is a brief introduction to the molecular biology and biochemistry necessary to understand what follows in the book. Those readers who are already familiar with the basics of these subjects should feel free to skip this chapter. For those of you, who need some introductory material, the next fifteen pages is an effort to get you started. Obviously, this material cannot substitute serious courses in biochemistry and molecular biology. Hopefully, as you get familiar with the material of the book, you will be find the questions we try address in the book interesting enough. This interest, in turn, would encourage you to study these subjects further. In fact, this chapter contains a bit more than the absolute minimum of information necessary to understand the rest of the book. That is by choice. The idea is to enable you to engage in discussions in biology and perhaps find a research problem of your own. Biology offers an inexhaustible number of such problems for those interested in quantitative modeling, as we will see.

1.1 Life and evolution

Life on Earth started about 4 billion years ago within a billion year of the Solar system forming and the planets coming into existence. Life has been around for about a third of the period the Universe has been around. The world early life coped with was very different from the world we see now. As physical environment on Earth changed, life changed accordingly. All of us, bacteria, fungi, plants, worms, insects, all the way to vertebrate animals, come from common ancestors. These early organisms possibly resembled eubacteria and archea. It is not surprising that these many of these bacteria are highly evolved and are extremely adaptive: ready for many different contingencies. The early single cell organisms have a relatively simple structure. The cell does not have separate envelope holding the genetic material, DNA (more on it shortly). More or less, everything happens in the bag that is inside the cell wall. These organisms are called prokaryotes. Some prokaryotes, found a way of using sunlight for energy, and created oxygen as a byproduct two and a half billion years ago. It changed the whole dynamics of evolution.

About two billion years ago, the eukaryotes, the ones with well defined nucleus, came into being. This is the branch we, humans, come from. Eukaryotic cells are far more complex and compartmentalized. We will discuss this in greater detail in a later section. Life was still unicellular. However, one could make a case that the molecular machinery needed for more complex multicellular life was already being developed, albeit, for a different purpose.

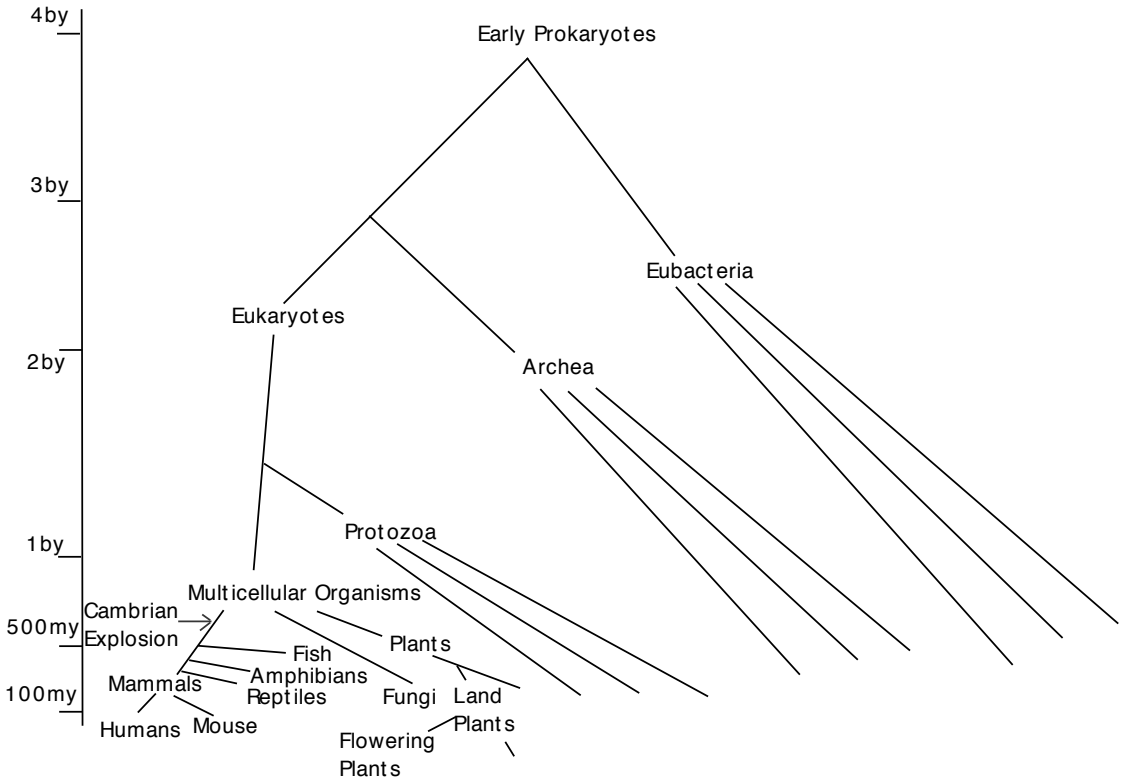

FIG. 1.1. The Tree of Life.

The multicellular organisms came into being about seven hundred million years ago. For hundreds of millions of years from that point, there were soft bodied sponge like animals. There are very few distinct kinds of fossils from that period. Whether that meant that there were not too many species at that time or that, for some reason, the fossils were not found, is open to debate. What is known is that from the Cambrian period, which starts about 570 million years back, one finds a plethora of fossils showing an amazing diversity of body forms. These animals often had skeletal body parts, which might have lead to more stable fossil records. Anyway, some researchers believe that evolution of body design speeded up at this point. Till that point, major evolutionary steps often happened by invention of new proteins or of new cellular morphology. One could possibly argue that the later stages of evolution often had to do with how various molecular components interact with each other and thereby regulate each others functions, than with evolution of new components.

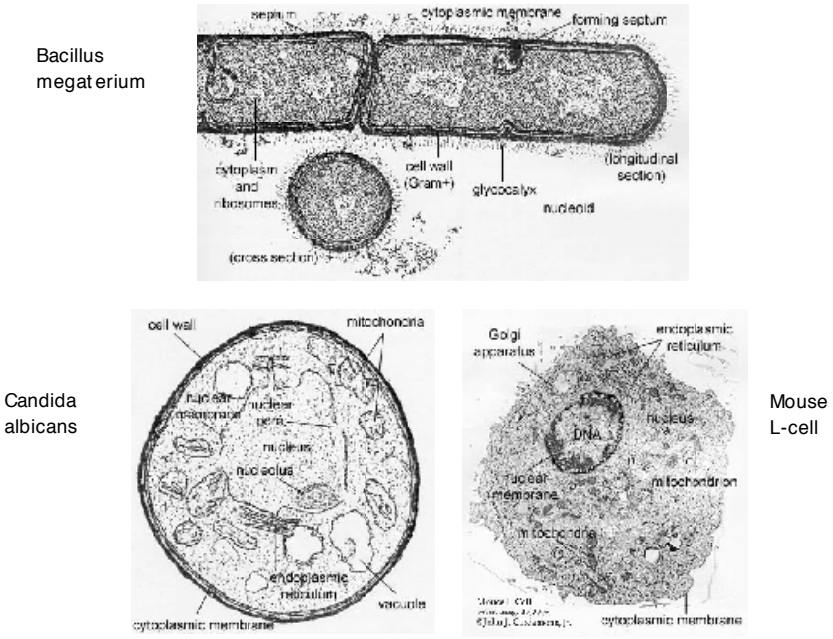

FIG. 1.2. Prokaryotic vs Eukaryotic Cell (courtesy Dr. Kaiser and Dr. Carbonne).

At the end of the Cambrian explosion, as that enormous diversification is called, almost all the different kinds of body plans that we see today have already been invented. Some have called this process biology's big bang. In the next four hundred million years saw the arrival of fish, reptiles, mammals and birds. It also saw the evolution of land plants which later gave rise to flowering plants. We would be able to relate to the world a hundred million years ago. Hominids (human like apes) are recent phenomenon: 20-25 million years old. We humans have been around for only one and half million years. If life on Earth was an hour long movie, human existence would occupy a blink of an eye.

Much of the story described so far is from fossil record combined with geological inferences. However we could also get a record of the history from the life forms present today, by looking at their relations at a microscopic and ultimately molecular level. For that we need to understand what is in our cells.

Type	Prokaryote	Eukaryote
Size	$0.1 - 10\mu m$	$10 - 100\mu m$
Genome	Single circular DNA	Multiple chromosomes
Organelles	None	Mitochondria and others
Metabolism	Many strategies	Mostly oxidative
Internal membranes	None	Complex folded ER
Motility	Flagella	Complex undulipodium

Table 1.1 *Some differences between prokaryotic cells and eukaryotic cells, other than existence of a nuclear envelope.*

1.2 Structure of a Cell

The basic unit of all living beings are cells [1]. Cells are enclosed within cell walls. The figure shows basic components of prokaryotic and eukaryotic cells and contrast their structures. In prokaryotes, there is no compartmentalization of the basic processes making proteins from DNA. Eukaryotes distinguish themselves by having a nuclear membrane which encloses the genetic material, DNA. RNA, a molecule that copies information about genes, from DNA, carries the instructions out of nucleus into a rather convoluted surface called the endoplasmic reticulum (ER), where proteins are made. Eukaryotes also have many additional organelles, some of which (like mitochondria and chloroplasts, for plants) come with their own genetic material. It is believed that such components may have come from free living prokaryotes at one point of time.

1.3 Biopolymers: DNA, RNA and Proteins

The nucleic acids, DNA and RNA, and the polypeptide chains making proteins are biopolymers of fundamental importance. Without some basic familiarity with them, and understanding of the relationships between these polymers, it is impossible to appreciate why modern molecular biology is so powerful.

1.3.1 *The Nucleic Acids*

The central piece of the nucleic acids is the sugar ribose (see Fig. 1.3). Note the numbering of carbons from 1' to 5'. Nucleotides are made by attaching phosphate groups to the 5' carbon and a base to the 1' carbon. We need to be concerned with only five of the bases in the present context: Adenine (A), Thyamine (T), Guanine (G), Cytosine (C)and Uracil (U), the structures of which could be seen in Fig. **??**. Long polymers could be made by joining the phosphate group to the 3' carbon of the next nucleotide and so on. The link is called the phospodiester bond. A chain of linked nucleotide thus have a sugar phosphate backbone and bases sticking out from the sugars.

The major difference between between **r**ibo**n**ucleic **a**cid (RNA) and the **d**eoxyribo**n**ucleic **a**cid (DNA) is the missing oxygen at the 3' carbon. The "ribo-" is from ribose

[1] Viruses are an exception. But viruses do not live on their own. They need other cells to thrive.

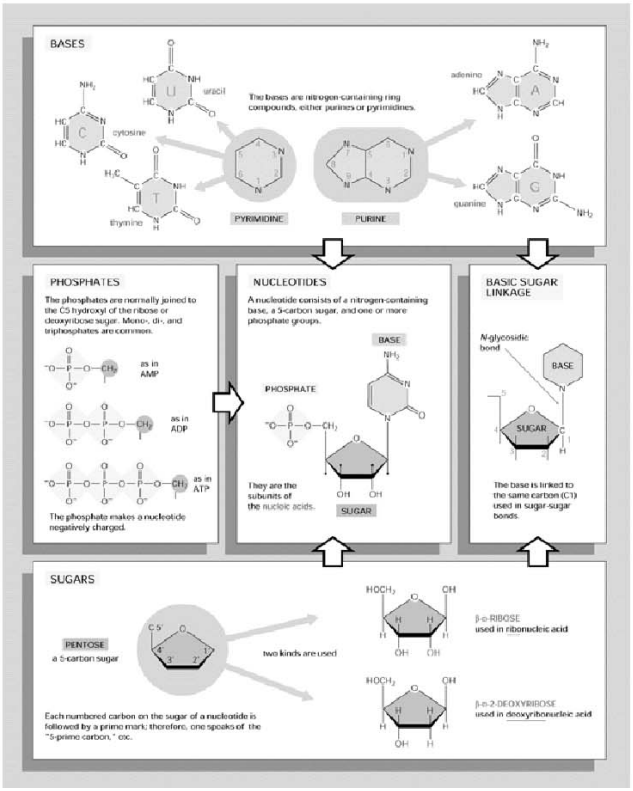

FIG. 1.3. Building blocks of nucleic acids.

and "nucleic" has got to do with being from the nucleus of cells. These are weak acids, the phosphate group being negatively charged having lost a proton. DNA utilizes the bases A,T, G and C.

Now that you have picked up enough chemistry, you can understand what is so magical about nucleic acids, especially DNA. It has to do with base pairing. Look how A -T and G-C form base pairs (Fig. 1.4 and Fig. 1.5) between two strands of DNA with complimentary sequences. The interaction between bases are due to hydrogen bonds.

This is a good place for an aside on hydrogen bonds, which play an extremely important role in molecular biology. A hydrogen bond is a kind of attractive intermolecular force between *p*artial electric charges on a hydrogen atom and a strongly electronegative atom (like oxygen, nitrogen or fluorine). Hydrogen bonds are quite strong but are much weaker than covalent bonds or full strength ionic bonds. They are strong enough to provide stability to a molecular structure but weak enough to be broken when a biological function demands structural changes.

As these base pairs are formed, sugar phosphate backbones twirl around each other making this pretty staircase. DNA is, universally, the repository of genetic information across biological species[2]. One of the important things about the double helix structure is that it immediately lead to a hypothesis about how specificity of base pairing could be utilized for copying genetic information. We will discuss that when we come to replication in the next section.

RNA has, another base, uracil (U) instead of thyamine (T), which base pairs with A. RNA is more often found in the single stranded form. As a result, it could form base pairs with itself, making complex three dimensional structures as in Fig. 1.6. Such structure is important for RNA molecules, like t-RNA, which have special functions.

RNA plays the role of the messenger in the flow of genetic information (see the coming sections on transcription and translation) as well as form molecules with enzymatic activity. Enzymes are biomolecules (in most cases proteins) that speed up certain biochemical reactions. In other words, they play the role of catalysts. Some hypothesize that in the earliest form of life, RNA was the genetic material as well as the constituent of the enzymes encoded in the genes. Whether this hypothesis, known as the "RNA" world, is true or not, recently there has been surge of interest in additional role RNA plays in the genetic networks. The regulatory role played by small RNA pieces is beginning to be explored and has already been used as a very powerful molecular tool for perturbing gene expression patterns.

1.3.2 *Proteins*

Much as we have come to appreciate the different roles of RNA, past and present, the fact remains that the proteins are the workhorses of present day cells. Proteins are made of amino acids, which get linked by peptide bond formation, shown in fig. 1.7. Individual polymers, made this way by stringing of amino acids, are known as polypeptide chains. A single protein molecule might consist of one or more of these chains.

There are 20 amino acids (out of about 80, found in nature) which make up almost all proteins. these amino acids differ from each other in their side chains. The biologically relevant amino acids are often classified based on the properties of the side chains into basic, acidic, polar etc. See Table 1.2.

[2]Some viruses use RNA as genetic material. But then, as we mentioned before viruses are special, somewhere between the living and the dead.

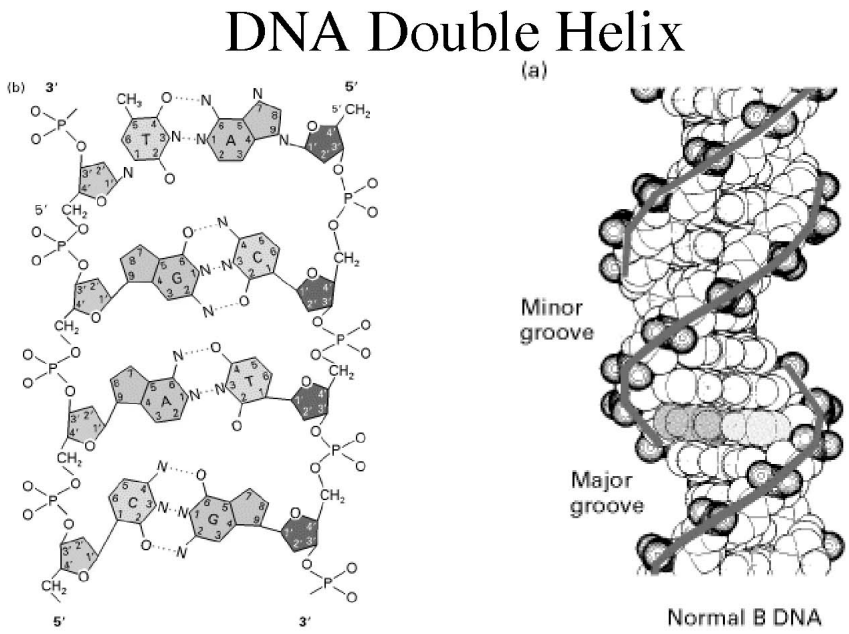

FIG. 1.4. Base pairing and the double helix.

Properties of side chains	Common amino acids
Basic (positively charged)	Lysine, Arginine, Histidine
Acidic (negatively charged)	Aspartic acid, Glutamic acid
Uncharged but polar (has a dipole moment)	Asparagine, Glutamine, Serine, Threonine, Tyrosine
Nonpolar	Alanine, Valine, Leucine, Isoleucine, Proline, Phenylalanine, Methionine, Tryptophan, Glycine,Cysteine

Table 1.2 *Classification of side chains of common amino acids*

Another Look

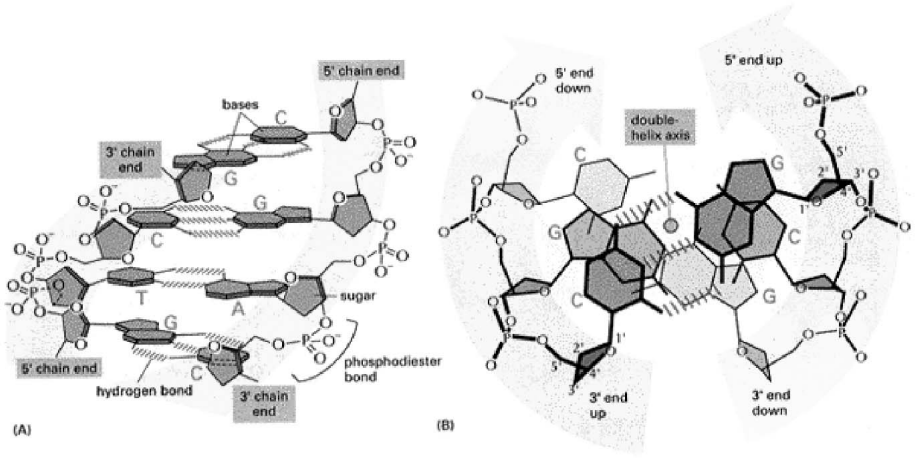

FIG. 1.5. 3-d structure of the double helix.

Proper functioning of the protein depends upon its three dimensional structure, and therefore, folding of the polypeptide into the right structure is extremely important. For most proteins, the ultimate structure is determined by the precise sequence of amino acids of individual polypeptide chains. This sequence, that is determined genetically, as we will see, is often referred to as the primary structure.

The next level of structure refers to certain repeating themes found in many proteins. Secondary structure, like alpha helix or beta sheet, is formed when the positively charged hydrogen from the amine group is brought close to the negatively charged oxygen on the carboxyl group. This is one more example of hydrogen bonds playing an important role in biomolecular structure. Further higher level organization, namely, tertiary structures has to do with how these helices, sheets and unstructured loops interact with each other to form a stable structure for a polypeptide chain. How multiple polypeptide chains weave together in a protein often goes by the name of quaternary structure. These higher

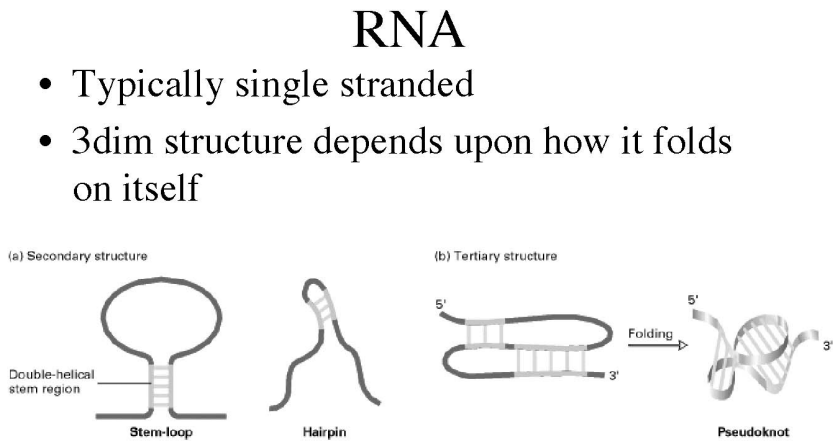

FIG. 1.6. Base pairing and 3-d structure of RNA.

order structures are affected by the nature of side chains. How the interaction of side chains affect the protein's shape is a complex story and is at the heart of the challenges faced by *ab initio* protein structure prediction. One might say that, prior to the genomic era, with high-throughput sequencing and genome-wide expression data, demanding computational processing, the problem of protein folding drew the largest number of investigators trained in quantitative methods into biology. It is still one of the major open problems, which is being attacked by a variety of methods, ranging from biophysical simulation to pattern recognition algorithms.

1.4 Central Processes: Replication, Transcription and Translation

The essential processes underlying life are remarkably similar among all the living beings. Fig. 1.9 refers to the the central dogma of molecular biology regarding how the sequence of a strand of DNA is transcribed into an RNA molecule which is then translated into amino acid sequence in polypeptide chains in proteins. DNA, the carrier of blueprints for proteins (and also some RNA machines), gets replicated, allowing one cell to divide many cells carrying the same genetic in-

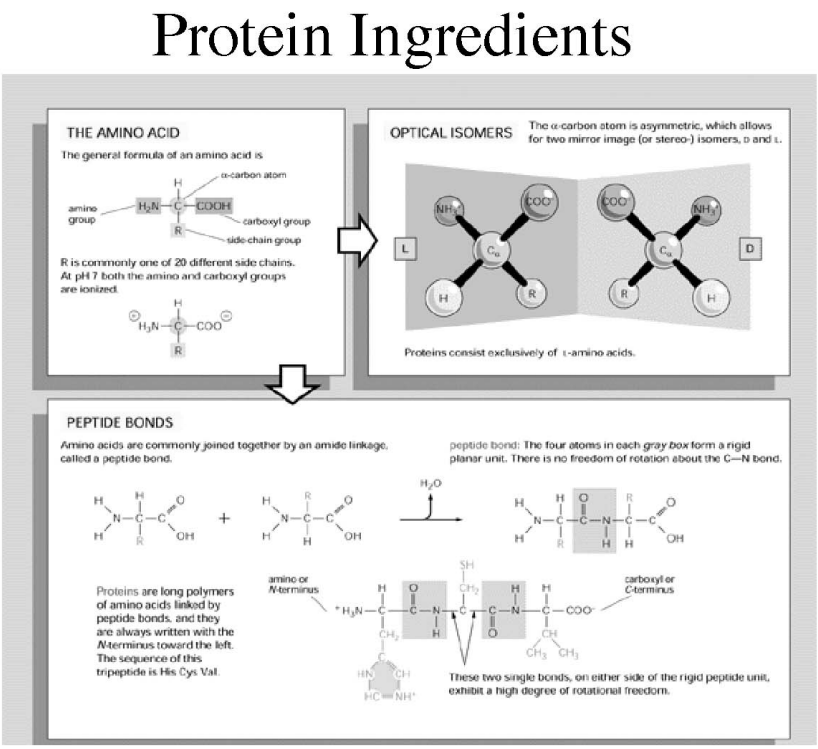

FIG. 1.7. Building proteins from amino acids.

formation. We have already discussed the biomolecules involved. Now we discuss these basic processes, replication, transcription and translation, and the machinery involved in carrying them out.

1.4.1 *Replication*

One of the the essential features of life is cell division and each cell receiving a reliable copy of the genetic information coded in DNA (more about how it is coded in the subsection to come). The celebrated understatement, "It has not escaped our attention that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." , made in the paper that elucidated the molecular structure of DNA (J.D. Watson and F.H.C. Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature, 171:737–738, 1953.), relied on the following observation. The two

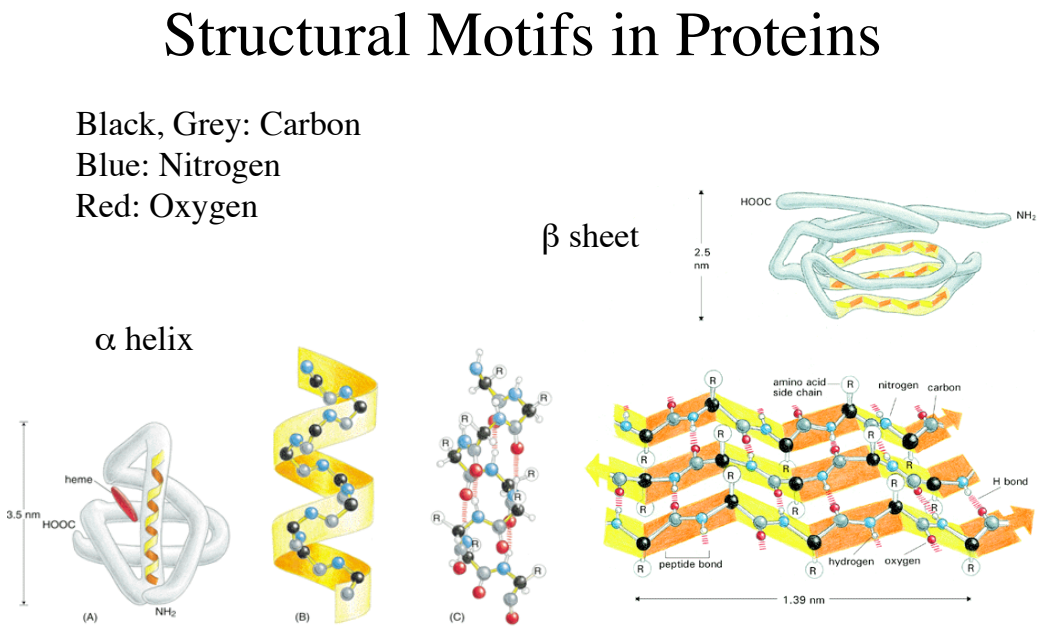

FIG. 1.8. The alpha helix and the beta sheet.

strands of DNA were held together by hydrogen bonds of the base pairing. Since hydrogen bonds are much weaker than covalent bonds, the two strands could be easily made to come apart, allowing each to be used as a template for further synthesis of two double stranded DNA molecules. This would be a possible mechanism of DNA replication, the process by which double stranded DNA generates a copy of itself. This picture turned out to be essentially correct, with some interesting twists.

The simplest model would be that DNA opens up and replication happens on both strands continouously (Fig. 1.10). Cosidering the model carefully, one notices that it requires one strand to be synthesized in the 5' to 3' direction whereas the other to be in the 3' to 5' direction. This would require two kinds of polymerase (polymerases are proteins that catalyzes the formation of polymers, like single stranded nucleic acids, from monomers, like nucleotides) actions. After

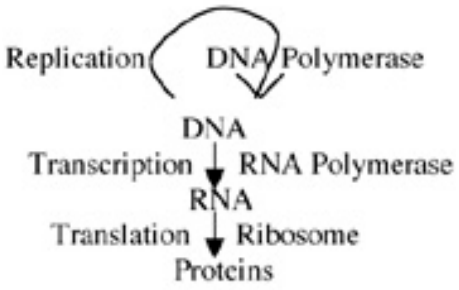

FIG. 1.9. Central biological processes.

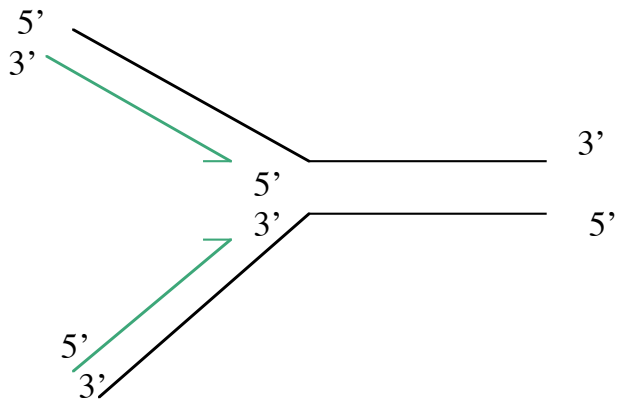

FIG. 1.10. Continuous replication model.

the discovery of the DNA polymerase, it became clear that the enzyme only synthesizes DNA in the 5' to 3' direction. No one has found a DNA polymerase that makes DNA in the 3' to 5' orientation. As resolution of the puzzle, Reiji Okazaki suggested that the DNA replication might happen in a discontinuous fashion. Today we know that the process is semi-discontinuous (Fig. 1.11). In

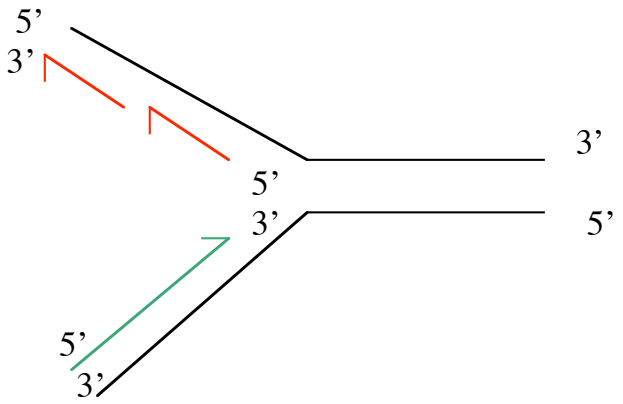

FIG. 1.11. Semi-discontinuous replication model and Okazaki fragments.

fact, it is quite complicated. Enzymes called primases make short RNA fragments hybridized to DNA to act as primers (starting points) for the DNA polymerase to come along and make the so-called Okazaki fragments. Finally the RNA parts are excised and replaced by DNA and then another enzyme called DNA ligase joins (“ligates” , in technical terms) all these discontinuous DNA fragments. Quite remarkably, all this happens very fast: at a rate of about 1000 base pairs/second. This speed is good enough to replicate prokaryotic genomes with few million bases in less than an hour. Prokaryotes therefore usually have only one location on their DNA from which an open replication bubble spreads. This location called the origin of replication. Eukaryotic genomes are much bigger and hence need multiple origins of replication. Another difference between prokaryotes and eukaryotes is that the fidelity of replication. Prokaryotes have an error rate of $10^{-6} - 10^{-7}$ per base per replication. In eukaryotes, it is about 10^{-9} per base per replication, thanks to additional proof-reading mechanisms.

1.4.2 *Transcription*

Transcription is the process by which RNA is made from DNA. The key player in this process is another polymerase, one that makes RNA instead of DNA. RNA polymerase copies a the information on a stretch of DNA, usually a few thousand bases, into a piece of RNA. This process is called transcription. The transcribed stretch of DNA, along with regions around it important for controlling transcription rates, l is the basic unit of hereditary material, often referred to as a gene. To start the process of transcription, the RNA polymerase has to

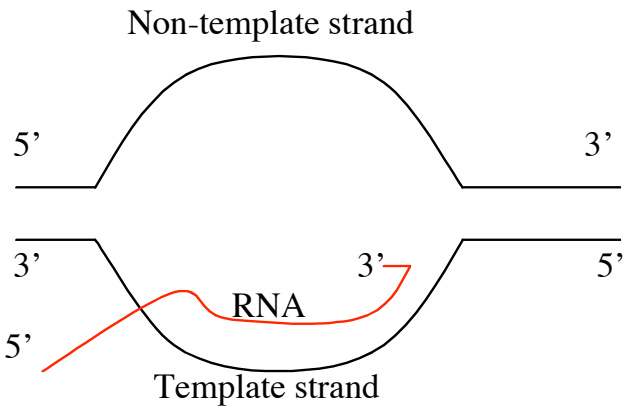

FIG. 1.12. Transcription.

bind upstream (to the 5’ end) of the transcribed sequence (regarded as running from 5’ to 3’ end of the non-template strand, see Fig. 1.12) in DNA. Much of genetic regulation happens through controlling this step. Once the polymerase binds, it forms the so-called open complex, by opening a DNA bubble. It then goes ahead making RNA, using one of the two strands as a template, as shown in the same figure (Fig. 1.12). In this process, called elongation, the RNA is made in the 5’ to 3’ direction.

In prokaroytes like *E. coli*, the initial recognition happens because the core enzyme (the essential parts of RNA polymerase) come in complex with a sigma factor, which is a DNA binding protein. It recognizes certain sequences upstream of the gene. This process can be regulated by additional DNA binding proteins bound to neighboring sites. Some proteins interact with the polymerase and enhance its binding. Other regulatory proteins, often known as repressors, get in the way of the initial RNA polymerase binding or affects the process of elongation negatively.

Transcription is far more complicated for eukaryotes. In eukaryotes, the DNA is packed around nucleosomes, which are complexes of proteins called histones. Nucleosomes are then packed into higher order structures, forming chromatin. DNA, tightly packed this way, does not bind proteins necessary for transcription. In many cases, there are regulatory DNA sequences, typically upstream of the coding sequence, known as upstream activator sequences (UAS). Such a sequence binds an activator, a regulatory protein, which in turn can recruit

other proteins that could modify chromatin structure making it more permissive for transcription. This modification exposes sequences (like the so-called TATA box, a sequence similar to TATAAAA) to which Transcription factor IID or TFIID complex binds, beginning the recruitment of many other components finally leading to transcription.

The speed with which RNA polymerase synthesizes RNA is an order of magnitude lower than the rate for DNA synthesis by DNA polymerase. In prokaryotes it is of the order of 30-85 bases/sec.

1.4.3 *Translation*

The information contained in RNA is converted to a polypeptide by the ribosome, an enzyme made of RNA itself and some proteins. The steps to protein synthesis, subsequent to making of the primary RNA transcript are different in prokaryptes and eukaryotes. In prokaryotes, the ribosomes come attach themselves to specific ribosome binding sequences on the nascent RNA (Fig. 1.13(a)). Prokaryotic RNA often encodes, in tandem, for multiple proteins, with the ribosome binding sites, also called the Shine Dalgarno sequences, between the regions coding for different proteins. Eukaryotic primary transcript goes through far more processing (Fig. 1.13(b)). The RNA gets capped on the 5' end by a methylated G. This cap is finally what the eukaryotic ribosomes recognize. The regions in RNA which does not code for proteins, the introns, are spliced out. A poly-A tail is added to the 3' end. The resulting messenger RNA (mRNA) is exported out of nucleus. The mRNA finally arrives at the endoplasmic retiuculum (ER) where it meets ribosomes.

The process of making a polypeptide, once the ribosome is bound to the mRNA, is very similar across all living beings. The translation apparatus moves from the 5' to the 3' end. It starts from an AUG, coding for Methionine. RNA is read in non overlapping triplets coding for different amino acids. It goes on till it recognizes any of the three triplets that code for a stop (UAA, UAG and UGA). The recognition happens by transfer RNAs (t-RNA) for each amino acid having a triplet complimentary to the code. The t-RNA comes covalently bonded with the amino acid, binds to mRNA at the right place and offers the amino acid for joining to the polypeptide chain. The process is indicated in Fig. 1.14.

The rule book of triplets to amino acids (Fig. 1.15) is nearly universal, with a few exceptions, like for mitochondrial genes. Universality of the genetic code, once more, indicate the common origin of life. Whether there is any deep reason why the code is this way or whether this is just "a frozen accident" (as Francis Crick puts it) we don't know.

We have studied how information on DNA is converted into to a protein molecule. Scientifically, it was a remarkable achievement to have solved the essential aspects of this process within two decades, the 1950s and the 1960s. It turns out that the whole process is heavily regulated. Nature has used any knob it could find to tweak the rate of protein synthesis adaptively. We will come back to this aspect when we study genetic networks.

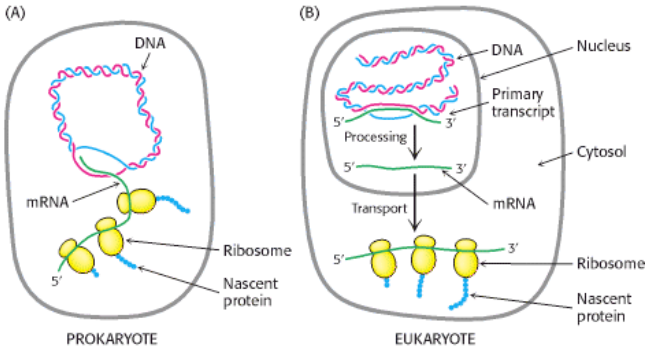

FIG. 1.13. From transcription to translation.

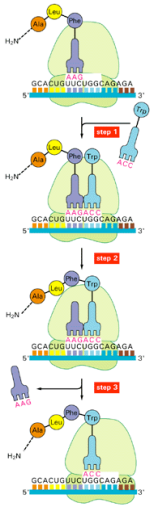

FIG. 1.14. Translation.

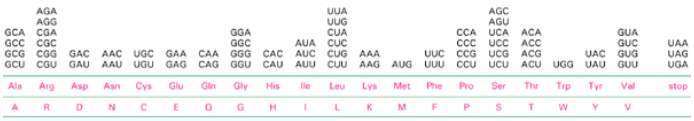

FIG. 1.15. The genetic code.