A Biology Primer for Computer Scientists

Franco P. Preparata [*]

1 Introduction

The intent of these notes is to outline a minimal background for computer scientists wishing to deal with computer algorithms relevant to problems in molecular biology. In such context, it appears very pretentious to even try to formulate definitions concerning the enormously complex problem of life. With this *caveat*, however, one can identify some of the most characteristic features of living organisms, which distinguish them from other physical systems that no one would label as living.

Perhaps the most distinguishing feature of life is the ability to *replicate*, i.e., the ability of a living entity to reproduce its structure as an entity capable to further the process of replication (*inheritance*). This ability is embodied by a single molecular structure, which carries the "blueprint" (or *code*) for the entity being replicated. This molecular structure is the DNA (deoxyribonucleic acid). The entity being replicated is not necessarily an autonomous organism – it may just be a (nonautonomous) cell of a living organism. As we shall see, the DNA of an organism is a code for both the physical structures and the functions of the organism.

We begin by reviewing the physical constituents of living organisms, in order of increasing complexity, starting at the atomic level.

2 Basic chemical constituents

Only a small subset of the physical elements appear in the chemical composition of living organisms. Within this subset, four elements are preponderant. They are C (carbon), H (hydrogen), N (nitrogen), and O (oxygen). Two more elements appear in significant

[*] Computer Science Department, Brown University, 115 Waterman Street, Providence, RI 02912-1910, USA. E-mail: franco@cs.brown.edu

fractions: P (phosphorus) and S (sulfur). Some other elements appear only in small traces: Cl (clorine), Ca (calcium), Mg (magnesium), Cu (copper), Fe (iron), Mn (manganese), Zn (zinc), and Co (cobalt).

Individual atoms bond to each other to form more complex entities (molecules). There exists a number of binding mechanism, whose detailed study is not necessary in this context. Suffice it to say that the strength of different binding mechanisms varies substantially, and that the energy necessary to break a strong bond may be up to two orders of magnitude as large as that required for a weak bond. The strong binding mechanism encountered in molecules occurring in living organisms (*biomolecules*) is the covalent bond, whereby two atoms share pairs of electrons in their outer shells (this is the type of bond normally occurring among C,H,O,and N).

The structure of a molecule is traditionally displayed in two principal forms, the formula and the diagram. The formula basically describes the composition of the molecule, i.e., the multiplicity of each of its constituent atoms. The diagram is structurally more informative, since it consists of an undirected weighted graph, whose nodes are atoms and whose edges are the bonds (their weight denoting the number of shared electron-pairs). However, even the molecule diagram is only a partial description of the molecular structure, since the atoms of a molecule arrange themselves in a three-dimensional structure. In fact they assume a configuration that is stable, i.e., it minimizes the internal energy over all possible configurations. The backbone of this structure is provided by covalent bonds, but a number of other (much weaker) bonds between pairs of atoms determine a three-dimensional configuration for the molecule, which is not really rigid (since the weaker bonds are not sufficient to enforce rigidity) but has a high degree of rigidity within small neighborhoods of the diagram (we refer to this property as local rigidity). As we shall see, this molecular geometry plays a central role in the processes of life and is completely determined by the chemical structure of the molecule.

In fact, to simplify (perhaps oversimplify) the complexities of the biological processes, one may say that life is expressed as sequences of chemical reactions. A reaction proceeds in a certain direction (from reagents to reaction products) depending upon the concentrations of the molecules involved. Given that the concentrations are appropriate for a certain reaction to occur, the velocity of the latter is essentially determined by the ability of the molecules of the reagents to come in contact with each other. An agent which aids the realization of such proximity is, in general, called a *catalyst*, and a biological catalyst is called an *enzyme* (as we shall see later, enzymes belong, in the near-totality of cases, to the class of macromolecules called proteins).

The basic way an enzyme works is the following. An enzyme is itself a molecule, and has a given three-dimensional configuration, locally rigid. An enzyme for a given reaction has a region of its surface which matches a specific region of a reagent (in the way a plug matches its socket or a key matches its lock. See, for a two-dimensional sketch, Figure 1). This geometric compatibility coupled with weak binding, enables temporary locking of the enzyme to the reagents, which are thereby held in a proximity which favors the reaction. This explains the crucial roles played by geometry and weak binding.

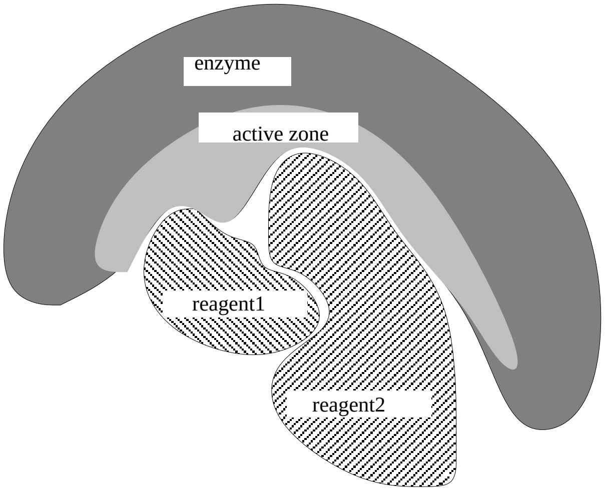


Figure 1: A two-dimensional sketch of enzymatic action. Reagents are usually called substrates.

After briefly outlining these basic processes, we consider the next level of complexity, i.e., the building blocks of living organisms, called *biomolecules*. The basic biomolecules are relatively simple, and are of four basic types: sugars, fatty acids, amino acids, and nucleotides. The amazing complexity of biological structures and processes emerges through the combination of these simple building blocks (as, metaphorically, the 26 letters of the alphabet are adequate to express the most sophisticated literary work).

Sugars are aldehydes with the typical formula CHO-$(CHOH)_n$-CH_2OH, for some $n \geq 1$. For $n > 2$, they tend to form cyclic structures, whereby the carbon of the term CHO- forms a covalent bond with the O of one of the intermediate terms –CHOH–. As we shall see below, of interest to us are two 5-carbon sugars known respectively as *ribose* and *deoxyribose*, which have of course cyclic structures. Fatty acids are of no immediate interest to our considerations. Much more relevant to us are other organic acids known as *amino acids*. Their general formula is NH_2-**R**CH-COOH, and the term **R**, called the *side chain*, characterizes the amino acid; of interest to us are 20 choices for **R**, although additional types of amino acids appear in nature (of these 20 amino acids only one, Proline, does not have the general structure just introduced). Finally, a crucial role is played by *nucleotides*, which we shall describe in some detail below.

Biomolecules of each of the above categories can join together to form long amino acids, and nucleotides are respectively called polysaccharides, lipids, proteins, and nucleic acids (RNA and DNA).

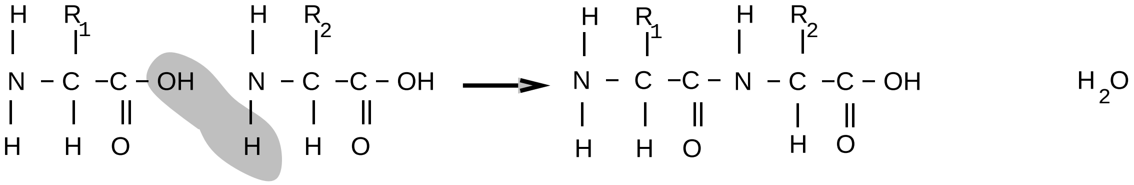


Figure 2: Polymerization of two amino acids.

We shall consider first the structure of proteins. The formation of an amino acid polymer is illustrated in Figure 2, where two amino acids (with respective side chains R_1 and R_2) are polymerized. The process consists of joining the –OH component of the –COOH group (carboxyl end) of the left amino acid with a hydrogen atom of the NH_2– group (amino end) of the right amino acid, creating a CN covalent bond and freeing one molecule of water. Such polymers, consisting of the concatenation of amino acid residues (residues, because of the subtraction of H_2O) are commonly referred to as *polypeptides*.

Next we examine the structure of nucleotides and their polymers.

3 Nucleic acids.

The building blocks of nucleic acids are the nucleotides. A *nucleotide* is a rather simple molecule whose structure is a chain of three components (see Figure 3(a)): (i) a base **B**, (ii) a sugar **S**, and (iii) a phosphoric acid **P**. The phosphoric acid component is common to all nucleotides, which are differentiated by the (not independent) selections of two types of sugars and five types of bases, for a total of eight different nucleotides.

The central constituent of the chain, the sugar, is actually a 5-carbon (therefore cyclic) sugar, known as *ribose*. The carbon atoms are consecutively numbered along the chain as $1', 2', \ldots, 5'$. The two types of ribose are differentiated by the fact that the $2'$-carbon appears in one case in the group –HCOH– (ribose), in the other case in –HCH– (deoxyri-

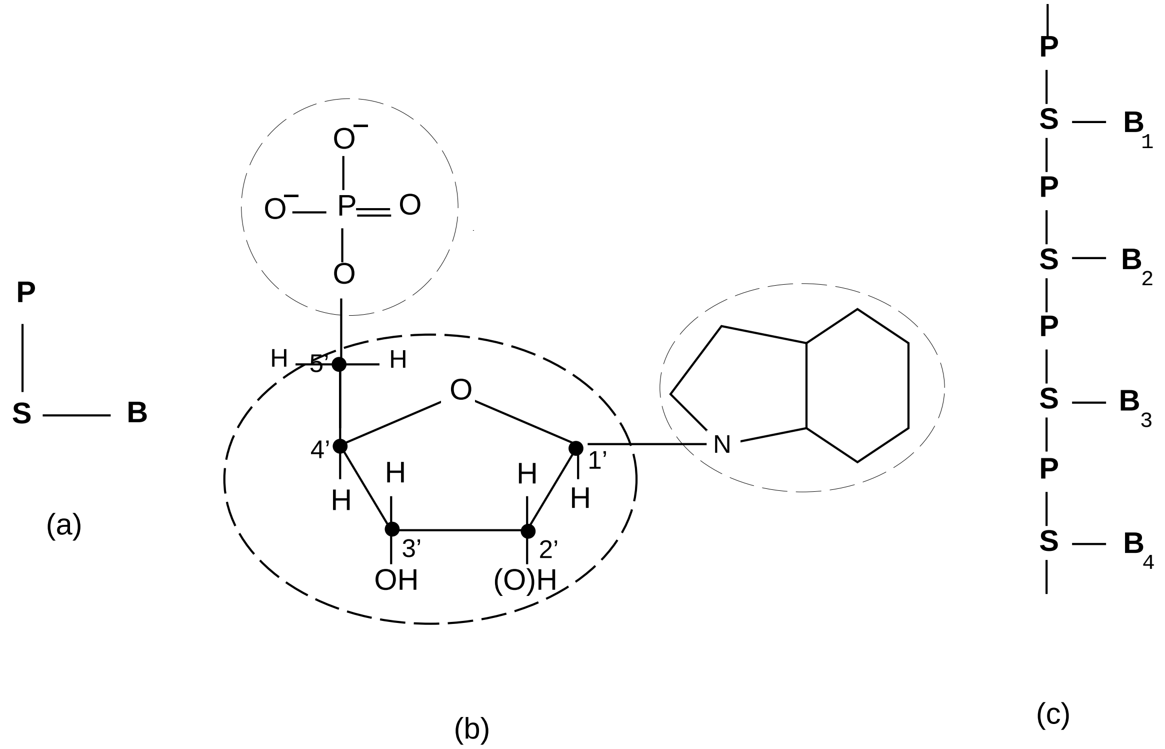


Figure 3: Nucleotides. Basic (a) and detail (b) structure. Single-stranded polymer.

.

bose, where a H replaces the oxygenated OH). The 1′-carbon is linked to the base and the 5′-carbon is linked to the phosphoric acid. The phosphoric acid component is in the ionized form. The bases are compounds consisting of one or two cycles, whose backbones are C and N atoms, respectively denoted pyrimidines and purines. They bind to the 1′-carbon of the ribose realizing a CN covalent bond. There are two purines (A, adenine, and G, guanine) and three pyrimidines (C, cytosine, T, thymine, and U, uracil). A, C, and G bind to both ribose and deoxyribose, and occur in DNA as well as RNA; T binds only to deoxyribose and occurs only in DNA; U binds only to ribose and occurs only in RNA. This shows that there are eight types of nucleotides appearing in nucleic acids. A nucleotide is specified by its base: Therefore when considering a base appearing both in DNA and in RNA (such as A,C, and G) the context will make clear that different types of ribose are involved.

Nucleotides polymerize as nucleic acids. Referring for simplicity to DNA, a DNA strand is a chain (sequence) of nucleotides, so that two consecutive terms are linked as follows (in a conventionally chosen direction): one of the ionized oxygens of the phosphoric acid component binds to the $3'$-carbon of the *preceding* term. This shows that each nucleotide participates in the chain through its $5' - 4' - 3'$, carbons, and a long DNA strand is conventionally oriented from the $5'$ end of its first term to the $3'$ end of its last one. (This orientation agrees with direction of execution of the fundamental biological processes occurring in the cell.) In the polymer each phosphorus group retains one ionized oxygen atom, which makes DNA (as well as RNA) strands negatively charged (a property essential for reading their base sequences, as we shall see). An analogous structure describes RNA.

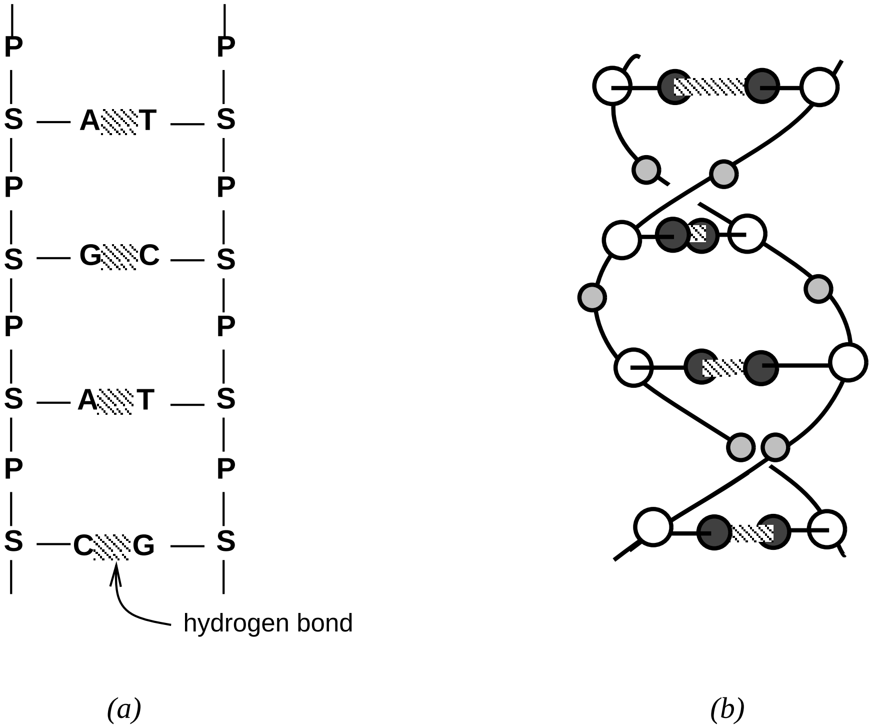


Figure 4: Structure of double-stranded DNA (a). Illustration of the double-helical structure of DNA (b).

A distinguishing feature between DNA and RNA is that the former appears double-stranded whereas the latter is single-stranded (There are other important differences, to be discusses below). The mechanism at the basis of DNA double-strandedness is the fact that each of the four DNA bases has a strong affinity for one of the other three bases, or, as they say, A is *complementary* to T, and G to C. In either case, a purine is complementary to a pyrimidine. This complementarity is due to the establishment of weak bonds, called *hydrogen* bonds, between the atoms of a base pair. Specifically A and T realize two such hydrogen bonds, whereas C and G realize three of them (and the coupling is correspondingly stronger). Therefore a given DNA strand is joined with a complementary strand (as shown in Figure 4(a)), where each base of one strand is mirrored by its complementary base in the other strand (Watson-Crick pairing). This pair of complementary strands (a fragment of which is shown in Figure4(b)) assumes a celebrated three-dimensional minimum-energy configuration: the *double helix*, whose period is nearly 10 base pairs (see Figure 4(b) for a sketch). It must be pointed out that the mutual attraction of complementary bases is not by itself a sufficient condition for double-strandedness of nucleic acids. In fact, complex molecular machinery is needed to effect the pairing, and such machinery exists for DNA but not for RNA (in other words, the existing DNA machinery is inoperative on RNA).

4 Overview of the fundamental cell processes.

As mentioned earlier, life is expressible as collection of amazingly complex chemical reactions. The fundamental chemical processes take place in the cell, which contains the genetic material (nucleic acids) and the machinery necessary to synthesize the required structural constituents and to implement the necessary functions. Cells are of different sizes and complexities, and are structurally different in different classes of organisms (without a nucleus in simpler organisms – *prokaryotes* – and with a nucleus in more complex ones – *eukaryotes*); such details, however, are nearly irrelevant to our considerations.

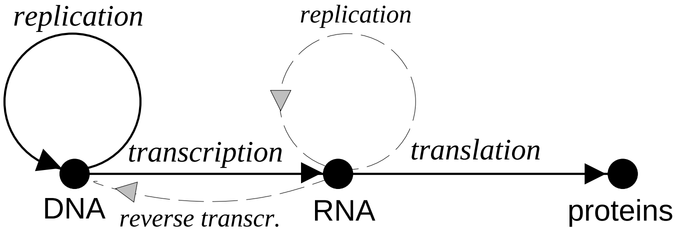


Figure 5: Diagrammatic illustration of the ”central dogma” and of its subsequent amendments (in light broken line).

Three major processes occur in the cell: **DNA replication, DNA-RNA transcription**, and **RNA-protein translation**, referred to briefly as replication, transcription, and translation. These processes are schematically represented by the graph of Figure 5, with vertices {DNA, RNA, proteins}, and edges labelled by the corresponding processes. This graph specifies that, besides DNA replication, RNA is derived exclusively from DNA, and proteins are derived exclusively from RNA. These relationships are referred to by biologists as the *central dogma*, to which, however, in more recent times rare exceptions have been discovered (in the form of RNA direct replication, and RNA-DNA reverse transcription).

In the next section we shall examine in some detail each of these three major processes.

5 DNA replication.

DNA replication is the process by which a double-stranded DNA sequence produces two double-stranded sequences identical (in the absence of errors!) to the original one. The way this happens is that the original complementary strands unwind and for each of them a new complementary strand is synthesized. For the synthesis to occur, a specific site (*origin*) on the original double-stranded sequence is located, beginning at this site the two strands are unfolded, and synthesis of both new complementary strands starts (in more advanced organisms with longer DNA sequences, there may be multiple origins). The process continues as follows: The strands are progressively unfolded and synthesis occurs very near the point of separation, until the entire sequence has been duplicated. We shall discuss shortly the mechanics of this synthesis; suffice it to say now that strand replication always proceeds from the $5'$ end towards the $3'$ end. This implies that it proceeds in opposite directions on the two strands, which are respectively denoted *leading* and *lagging*. Replication of the leading strand is structurally straightforward. Replication of the lagging strand, however, is more problematic, because for this strand unfolding occurs from the $3'$ end towards the $5'$ end. To reconcile unfolding with duplication, the process is intermittent on the lagging strand, i.e., after a relative short segment (about 1000 bases) has been unfolded it is correctly duplicated (from the $5'$ end to the $3'$ end), and two consecutive segments thus created (called *Okazaki fragments*) are subsequently joined together (ligated).

Each of the processes sketched above is assisted by very complex and not yet fully understood enzymatic machinery (*DNA-polymerases*), which recognizes the origins, unfolds the strands and facilitates the synthesis. Recognition of the origin (a string of about 200 bases) is very important, because without such an activating device replication cannot start. This satisfactorily explains duplication of the leading strand. For the lagging strand, a different activating mechanism initiates the replication of each Okazaki fragment. This is in the form of a *primer*, a short RNA string (complementary to DNA substring) which is synthesized by a specific enzyme and forms the initial segment of the fragment, to be later replaced by the corresponding DNA string by another enzyme. [This detailed discussion is preparatory to the explanation of a very important technique in molecular

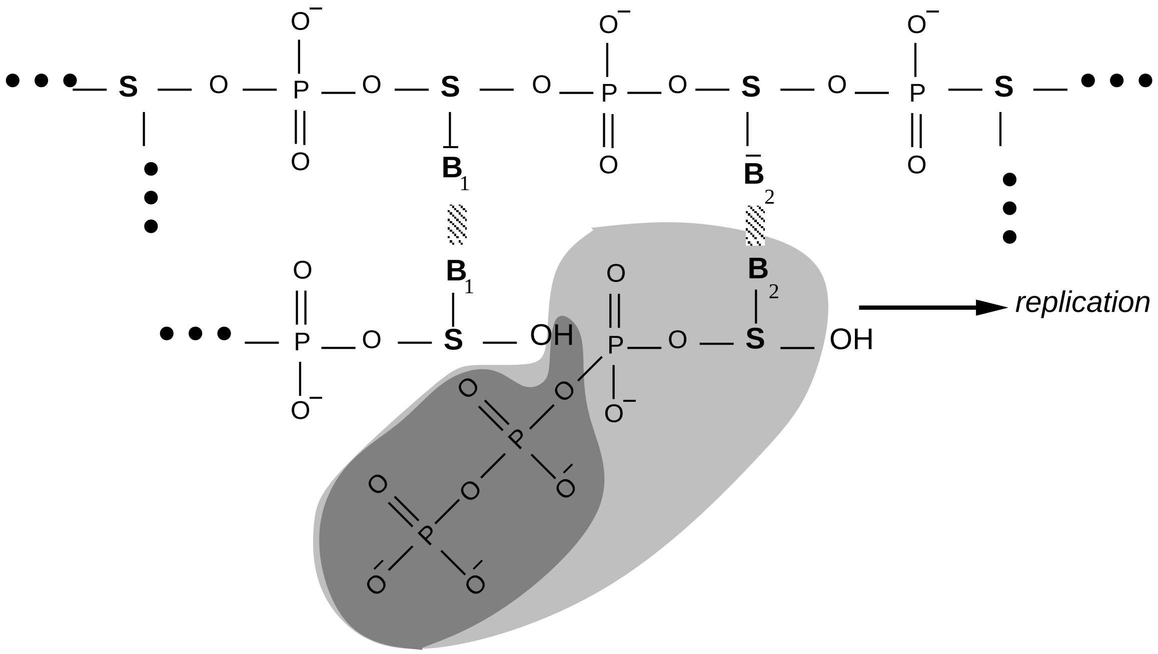


Figure 6: Illustration of the addition of one nucleotide in DNA replication. The shaded region corresponds to the pyrophosphate produced by the process. $\overline{B_i}$ is the complementary base of B_i.

biology.]

The latter occurs in the following manner. Reagents and enzymes are suspended in the cellular "soup" (called *cytosol*). In particular this soup contains the necessary elementary building blocks in the form of triphosphates, denoted here dNTP, where TP is an abbreviation for "triphosphate", $N \in \{A,C,G,T\}$ and dN denotes the deoxy- version of nucleotide N. Denoting by $(\mathrm{dNTP})_n$ an n–basepairs DNA-polymer, the typical reaction is

$$(\mathrm{dNP})_n + \mathrm{dNTP} \leftrightarrow (\mathrm{dNP})_{n+1} + \mathrm{PP}_i$$

where PP_i is an inorganic phosphate, called *pyrophosphate* (see Figure 6). Under enzymatic assistance the pyrophosphate reacts with water (hydrolysis) with production of heat (decrease of free energy). In the above reaction we have purposely used the bidirectional arrow to emphasize that it could proceed in either direction depending upon the relative concentrations of the reagents. Indeed, if we artificially produced a high concentration of pyrophosphate, the reaction would proceed in the opposite direction.

One may observe that DNA-replication is a process where a random mixture of triphosphates gets organized as the precise copy of a given DNA sequence. Locally we have the transition from disorder to order, a process that apparently contradicts the second principle of thermodynamics. The emergence of order from disorder had suggested in the past the presence of some "vital" principle governing phenomena of life. In reality, everything is explainable by means of the standard physical principles governing the natural world. If one considers the extended process, of which the actual copy process is simply an intermediate step, one recognizes an overall transformation of "high-quality" energy into "low-quality" energy. Quite simplistically, we may recall that forms of high-quality energy are transformable into one another and into low-quality energy (heat), and that the latter transformation is irreversible. Moreover, the quality of the heat decreases with its temperature. Qualitatively, in a steam engine high-temperature heat is partly transformed into mechanical work (a high-quality form of energy) but the rest of the energy is delivered to the environment as lower-temperature heat; or, in an air-conditioner heat extracted from an interior is brought to a higher temperature – to be delivered to the exterior – at the expense of mechanical work.

In the biological context, there is a large number of analogous mechanisms. For example, in plants ATP and other triphosphates are synthesized with the use of light energy (a high-quality energy). Subsequently triphosphates intervene in DNA replication and generate pyrophosphates, which are promptly hydrolyzed with the production of heat (low-quality energy). Heat is then absorbed by the environment. The overall cycle involves degradation of light energy to heat, with the intermediate product of a highly ordered structure, in total agreement with the second law of thermodynamics.

6 DNA-RNA transcription.

The transcription of DNA into RNA is a process that shares several features with DNA replication but it also differs from it in some fundamental ways.

The common feature is that transcription also rests on base complementarity. Specifically DNA-bases A,C,G,T are respectively paired with RNA-bases U,G,C,A.

An essential difference, however, is that only one specific strand (the so-called "*genomic*" strand) is used in transcription, which, as in replication, proceeds from the 5' end to the 3' end. The machinery involved in the process is an enzymatic complex (RNA-polymerase) analogous to DNA-polymerase. This enzyme separates the two DNA strands along a short stretch and trascription occurs on the exposed short segment of the genomic strand. To ensure error-free operation the RNA strand being transcribed forms a tight double-helix with the exposed DNA segment, while its tail floats freely in the cellular environment. The two separated DNA strands rejoin as the length of the free-floating RNA strand increases.

Other essential differences are that, whereas DNA is replicated in its entirety, RNA transcription is selective both in space and in time. Space-selectivity means that only selected substrings of the DNA string are replicated, and such selectivity is effected through the recognition by RNA-polymerase of unique starting and ending substrings (so that only the substring between such special signals if effectively transcribed). In addition, only substrings of the transcribed string are actually used by the cell, through the following mechanism. An RNA string σ consists of an alternating sequence of concatenated substrings $\epsilon_1\iota_1\epsilon_2\iota_2\ldots\epsilon_r$, for some integer r, where the ϵ-strings are called *exons* and the ι-strings are called *introns*. Introns are excised from the transcribed string and adjacent exons are spliced together, so that the string $\epsilon_1\ldots\epsilon_r$ is ultimately produced. The excision of introns is carried out by short RNA sequences (called *sn*RNA) that binds to the two extremities of the intron causing its removal (see Figure 7).

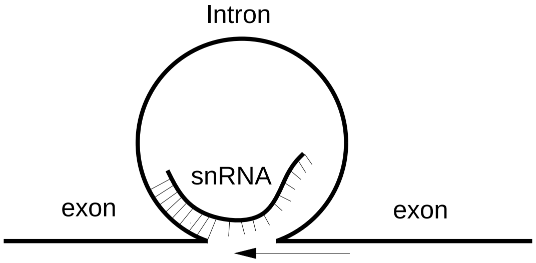


Figure 7: Simplified illustration of the excision of introns.

Time-selectivity refers to the mechanism whereby a substring is transcribed depending upon the environment, i.e., its instantaneous conditions. Specifically, let σ be the RNA string in question. String σ acts (through a complex process irrelevant to the present considerations) on a certain molecule μ, and therefore must be produced (transcribed from DNA) only when μ is present in appreciable concentration in the cell. Such regulated transcription requires a detector that activates the process. In simplified terms, the following happens: In the DNA the string $\overline{\sigma}$, complementary of σ is preceded by a starter string ω, called its operator. RNA-polymerase binds to ω to initiate transcription. Therefore, if ω is blocked, binding and transcription cannot occur. The blocking of ω is performed by a fourth agent, an enzyme γ that binds to ω when μ is absent (thereby preventing transcription) and to μ when present (thereby exposing ω and allowing transcription). The collection of DNA substrings involved in this complex regulating process is called *operon*.

Finally we note that the (relatively short) transcribed RNA strings are classified according to their different functions as follows:

1. mRNA,*messenger* RNA, strings involved in the process of RNA-protein translation (see next section). Each mRNA string is translated into a protein and is called a *gene.*
2. rRNA, *ribosomal* RNA, strings that participate untranslated in the the structure of the ribosome, the complex cellular machinery effecting the translation process.
3. tRNA, *transfer* RNA, untranslated strings assuming a sufficiently rigid three-dimensional configuration and act as linkages (see next section) between mRNA and protein chains (polypeptides).
4. snRNA, *small nuclear* RNA, acting in the excision of introns and splicing of exons, as described above.

7 RNA-protein translation (Protein synthesis)

In the preceding section we noted that strings of mRNA are translated into strings of amino acids, also called polypeptide chains or proteins. The protein is actually the end product of the process, i.e., the three-dimensional structure into which the polypeptide chain folds after its synthesis: the protein is really the functional agent, both structural (cell membrane, etc.) and biochemical (enzymes), while a polypeptide chain is to be viewed exclusively as a string over the amino acid alphabet. In fact the translation process is a string to string transduction.

A crucial feature of this transduction is that it is length-preserving, i.e., the ratio of the lengths of corresponding mRNA and polypeptide strings is constant. Recall that the RNA-alphabet has 4 symbols and the amino-acid alphabet 20 symbols. Let s and n be the respective lengths of a polypeptide chain and of the RNA strings encoding for it. Since there are 4^n RNA strings of length n and 20^s polypeptide chains, and there must be at least one RNA-string for each polypeptide chain, we have the obvious inequality

$$4^n \geq 20^s$$

i.e., $n \geq s\frac{\log 20}{\log 4} = s2.1609\ldots$. Such ratio, however, could be attained only by block-translation of blocks as long as the entire mRNA string. Such a method would involve a machinery so complicated that, not surprisingly, evolution has settled for an informationally less efficient but extremely reliable mechanism. The simplest case of transduction is memoryless (namely, the present output depends only upon the present input and not on past inputs) and has an output block of size 1, that is, a single amino acid. This implies that a string of at least $\lceil \frac{\log 20}{\log 4} \rceil = 3$ bases must be used to encode for a single amino acid, or that $4^3 = 64$ input configurations map to 20 output configurations. Indeed, this turns out to be the nature of the transduction, which is characterized by considerable redundancy (and, presumably, by increased reliability). The actual map from base triplets to amino acid is known as the *genetic code*, which was completely deciphered by the end of the sixties.

Before analyzing the structure of the genetic code, it is instructive to consider an incorrect hypothesis put forward before the code was elucidated. It was postulated that the code had to be *comma-free*, since there are no special markers (commas) to separate consecutive triplets thereby providing synchronization (synchronization is the device by which the translation mechanism locks on the first base of the triplet and not on any of the subsequent two). Therefore it was thought that only a subset of the triplets were recognized as valid (legal) inputs; in other words, a legal triplet would exclude two other triplets, because a periodic sequence ABCABCABCA... of repeating legal triplet ABC should be uniquely translated. This implies exclusion of triplets BCA and CAB from the legal set. Since each of the four triplets of the type XXX ($X \in \{A,C,G,U\}$) excludes itself, there appears to be a set of (64-4)/3=20 legal triplets, coinciding with the number of amino acids! This persuasive circumstancial evidence turned out to be wrong.

Indeed, the genetic code is not self-synchronizing. The DNA sequence is segmented into triplets of nucleotides (each triplet called a *codon*), and each codon is individually translated into an amino acid. There are 64 codons and only 20 amino acidis, i.e, an average of 3.2 codons per amino acid. As it turns out, 3 amino acids are coded by 6 codons, 5 by 4, 1 by 3, 8 by 2, and 2 by 1 (this accounts for a total of 61 codons); the three remaining codons are STOP signals, terminating the translation process. The actual detail of the code is not very relevant to our considerations; it suffices to note that of the $\log_2 21 = 4.32\ldots$ bits corresponding to the selection of a peptide (including the STOP condition), 3.85 are carried by the first two bases and only 0.54 by the third one, i.e., the discrimination provided by the latter is very weak.

More interesting is the synchronizing mechanism. In the absence of synchronization, there would be three reading frames, of which only one would yield functional proteins. It is therefore essential that translation be rigorously synchronized. This function is provided by the ribosome. The initial segment of an RNA gene contains a unique pattern of a few bases which binds to a complementary sequence of rRNA contained within the ribosome; this binding correctly positions the portion of mRNA to be translated.

Translation occurs in the ribosome, which is a complex and not yet completely elucidated three-dimensional structure, consisting of proteins and rRNA. Intuitively, the ribosome can be likened to a tape reader (of the mRNA sequence), which also produces an output tape (of the corresponding protein). Since there is no spatial fit between the codon and

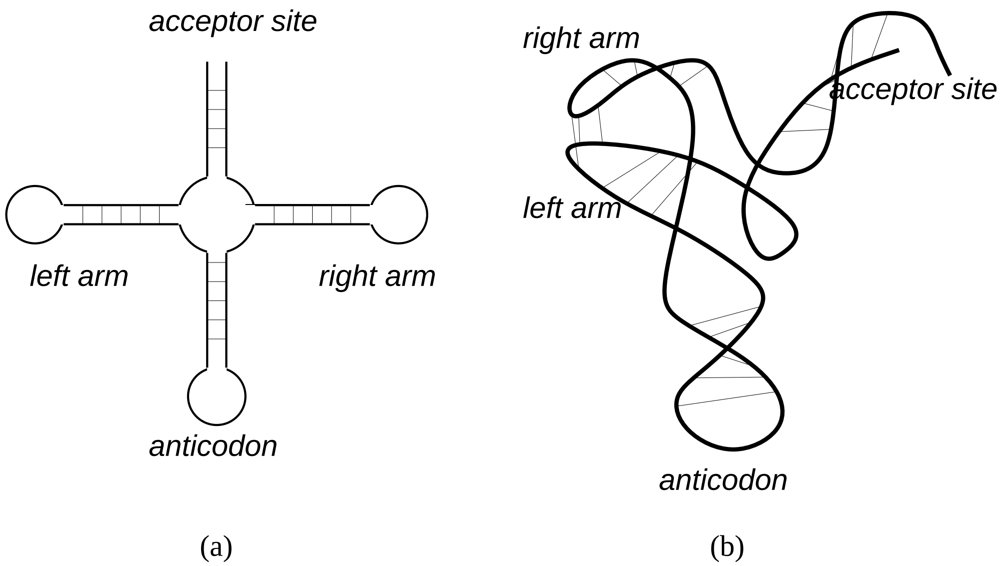


Figure 8: Simplified illustration of secondary and tertiary structures of tRNA.

the binding site of the corresponding amino acid, an adaptor structure is necessary. The latter is the tRNA, which is – to a first approximation – specific for a given (codon, amino acid) pair. A tRNA must be sufficiently rigid to maintain the special relationships required by the process. Rigidity is provided by the fact that, since RNA is single-stranded, complementary-base pairing may occur among its constituent bases. This phenomenon creates bindings between different sections of the RNA strings, causing it to bundle (technically, to fold). Defining its sequence of the RNA its *primary* structure, its *secondary* structure represents the pairing just described. The secondary structure of a typical tRNA molecule is schematically shown in Figure 8 (a). The contiguous base pairs arrange themselves as double helices (which are locally rigid); the unpaired bases of the loops at the end of the two arms pairs to each others (*tertiary* structure), giving rise to a three-dimensional configuration (in the shape of an L), of which an equally simplified illustration is given in Figure 8(b).

To a rough first approximation the ribosome consists of two major coupled three-dimensional domains (see Figure 9), of which RNA is only a relatively small fraction. However, RNA plays a crucial role in protein synthesis, both because it provides the synchronizing mechanism as mentioned above (rRNA), and because it provides the codon-amino acid adaptors (tRNA). Molecules of tRNA, attached to their specific amino acid at their acceptor site, float abundantly in the cytosol. In the mRNA sequence, close to and downstream from the

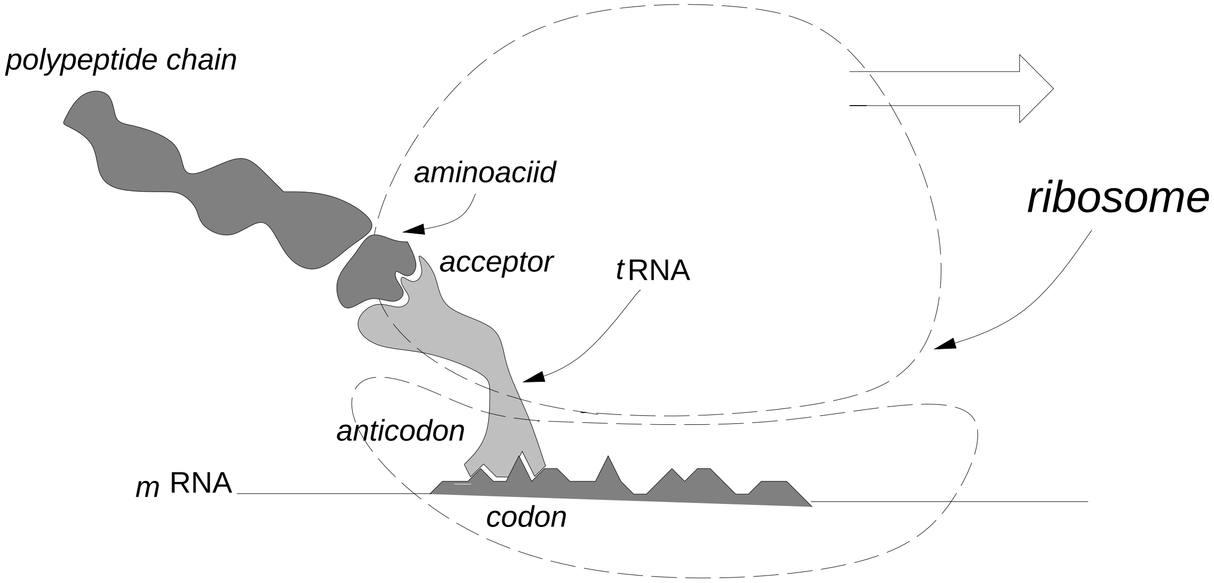


Figure 9: Illustration of the codon-amino acid adaptation within the ribosome.

synchronizing pattern, the portion to be translated begins with a codon which is unique within broad classes of organisms (for example, in eukaryotes the codon for amino acid methionine). Its specific tRNA binds to it to initiate the polypeptide synthesis (enzymes and distinct tRNA's differentiate between the occurrences of the methionine codon as initial or as internal). Once the polypeptide chain has been initiated the tRNA matching the next codon is captured from the cytosol and positioned behind the methionine-tRNA. The ribosome moves along the mRNA bringing this second tRNA in position for the attachment of the next amino acid. The process (polypeptide *elongation*) continues according to this general pattern until a STOP codon is detected. At this point, the completed polypeptide chain is released within the cell and its folding is completed.

8 Protein structure

As we discussed earlier, proteins are polymers (polypeptides) of 20 distinct amino acids resulting from the RNA-protein translation process. The amino acid (residue) sequence fully specifies the protein, but it is its spatial arrangement that determines its function. Normally a polypeptide has a unique spatial arrangement, but the elucidation of this mapping is only partially understood and remains one of the most daunting problems in

Computational Biology (protein folding).

—

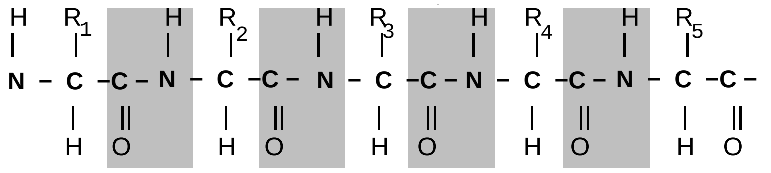


Figure 10: A polypeptide chain, with highlighted backbone

Let us reconsider the polypeptide chain, repeated for convenience in Figure 10. It may be viewed as an alternating sequence of a fixed "spacer" components, –CO–NH–, called peptide groups, and of variable components –HCR$_j$–, called "substitute methylene groups", where R$_j$ is a side chain (the carbon atom is called an α-carbon to distinguish it from the carbon in the peptide group). Essentially, the polypeptide sequence has a backbone of the type $(C_\alpha CN)^n$ and the residue sequence R$_1$R$_2$R$_3$... completely describes it. This description is called the *primary* structure.

To describe the spatial configuration of the molecule, we begin by noting that the C–N bond, referred to as the *peptide bond*, is quite rigid, so that the four atoms of the peptide group –CO–NH– are coplanar and rigidly connected. This rigid configuration is connected on both sides to α-carbons and can *rotate* around the axes of these connections: these two rotation angles, although not susceptible to assume all values in $[0, 2\pi)$, are degrees of freedom of the spatial configuration of the protein. Indeed the configuration (folding) of an n-amino acid protein is fully specified by $2n$ such angles.

In the spatial structure of a protein, special substructures are identifiable. Specifically:

- *α-helices.* In an α-helix the peptide chain winds around an axis (a straight line), with the planes containing the peptide groups parallel to this axis. In one period there are 11/3 peptide groups and consecutive turns are tied to each other by hydrogen bonds between hydrogen and oxygen atoms of peptide groups adjacent on contiguous turns (see Figure 11).

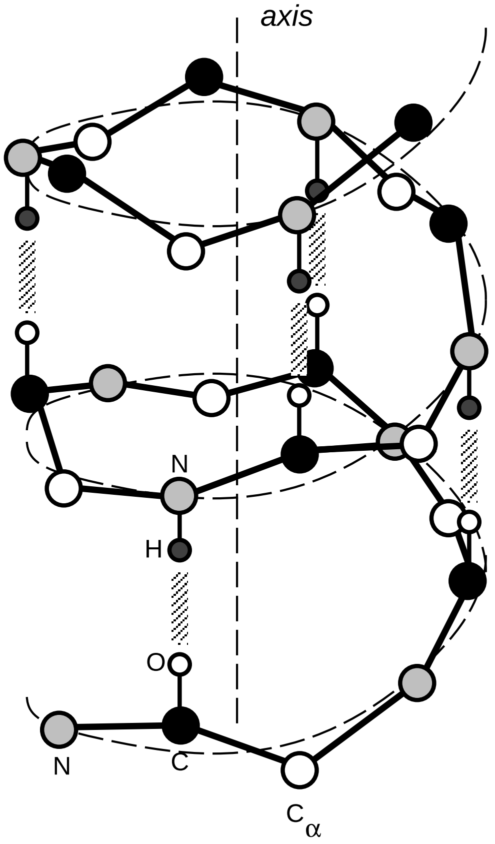
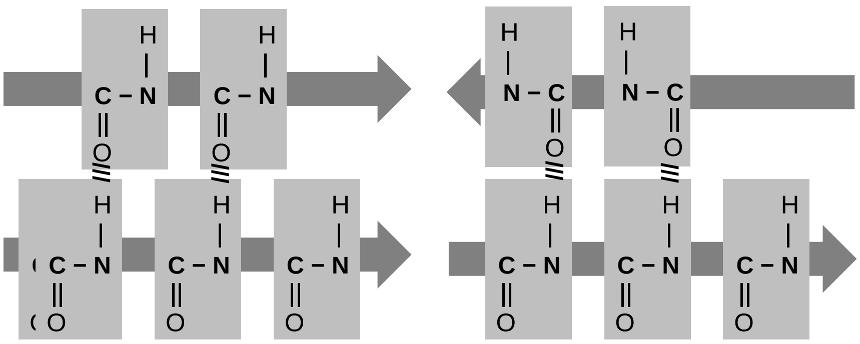


Figure 11: Illustration of a fragment of an α-helix. Not shown are the lateral chains

- β-strands and β-sheets. In a β-strand the peptide backbone is laid out basically along a straight-line segment. Two β-strands can run side-by-side and bind via the H–O hydrogen bonds. The structure is denoted parallel or antiparallel depending upon whether the corresponding backbones run in the same or in opposite directions. Two or more β-strands, connected in such fashion, tend to form a planar structure known as a β-sheet (see Figure 12).

- Loops. these are the section of the sequence which connect substructures of the two previosly described types.

As a result, we may view the peptide chain as a sequence loop($u$loop)^s, where the term u



(a) parallel

(b) antiparallel

Figure 12: Illustration of the two alternative forms of a β-sheet. (a) parallel,(b) antiparallel

is either an α-helix or a β-strand and the term "loop" is possibly empty. This structure is referred to as the *secondary* structure.

As defined, the secondary structure should be viewed as possessing some mechanical flexibility. In fact, if the bonds in the α and β components impart some local rigidity, the loop components are eminently flexible. This enables the various terms of the secondary structure to come in mutual proximity and to realize, if appropriate, additional bonds. The resulting structure is called *tertiary*, and to a great extent may be considered a rigid body.

Finally, two a more tertiary structure constituents (each originating from a distinct polypeptide chain) may form a more complex aggregate, called the *quaternary* structure.