1
Introduction

Soon after the discovery that deoxyribonucleic acid (DNA) is the carrier of the genetic information in all kingdoms of life, it became apparent that proteins are not directly produced from genes located on the DNA. Instead, the genetic information of protein coding genes is transcribed into ribonucleic acid (RNA) in a highly sophisticated process termed transcription. Such RNA molecules are termed messenger RNAs (mRNAs). In eukaryotic cells (cells that contain a nucleus), the genetic information on DNA is stored in the nucleus. The sites of protein production, however, are in the cytoplasm. MRNAs carry the genetic information from the nucleus across the nuclear membrane into the cytoplasm, where it is decoded and proteins are synthesized. This process is known as translation and the cellular factories that synthesize proteins are referred to as ribosomes.

For a long time, the fact that genetic information flows from DNA via mRNA to protein molecules was considered as a central dogma in molecular biology. However, still in the early days of molecular biology it turned out that not all genes that are found in the genome give rise to proteins. Such genes produce non-coding RNAs, which have highly specialized functions in the cell. The main classes that were initially found were ribosomal RNAs (rRNAs) and so-called transfer RNAs (tRNAs). Both classes are present in all organisms and are essential for protein production by the ribosome.

Initial genome sequencing projects revealed that only a minor portion of complex genomes such as the human genome consist of protein coding genes. The vast majority of such genomes are not coding for proteins. With the advent of new technologies such as large-scale sequencing (deep sequencing) or high-resolution microarray analysis, it became possible to analyze the transcription activity of whole genomes and in particular of genomic regions where no protein coding genes are located in molecular detail. The surprising outcome of such studies was that almost the entire human genome is constantly transcribed and therefore vast amounts of non-coding RNAs are produced. Today, we know that many classes of non-coding RNAs exist and we are only beginning to understand how such RNA molecules act and what their impacts on cell function are.

This book is divided into two parts. The first part focuses on *mRNA biology* and describes the journey of the genetic information stored on the mRNA from the

RNA Biology: An Introduction, First Edition. Gunter Meister.

nuclear sites of transcription to the cellular places where it is destructed and removed from the cell. Classical and well characterized parts of the mRNA life cycle such as transcription or translation are kept shorter, because these cellular processes are well covered by almost all biochemistry or molecular biology textbook. In contrast, more space is reserved for processes that are usually not comprehensively discussed by other textbooks. The second part of the book is dedicated to *non-coding RNA biology*. Non-coding RNAs ranging from the classical rRNAs and tRNAs to classes that have been identified more recently (such as microRNAs or short interfering RNAs) are discussed in detail.

1.1
RNA Building Blocks

Both RNA and DNA molecules are assembled from individual building blocks termed nucleotides. A nucleotide consists of a sugar molecule – a pentose for DNA and RNA – that is linked to a so-called base. The pentose that is the basis for RNAs is ribose and therefore such nucleotides are also referred to as ribonucleotides (Figure 1.1). The pentose found in DNA molecules is deoxyribose (Figure 1.1) and DNA nucleotides are also known as deoxyribonucleotides. Deoxyribose is a ribose that lacks the hydroxyl group at the 2′ position and this seemingly minor difference has a tremendous impact on the chemical nature of the molecule. The 2′ hydroxyl group is typically highly reactive and many enzymes that degrade RNAs make use of this hydroxyl group. Therefore, RNA molecules are usually much less stable than DNAs.

In nucleotide molecules, the hydroxyl group at the 1′ position is linked to the base. Bases are divided into two classes, namely purines and pyrimidines. This classification is based on whether a purine or a pyrimidine is the structural basis of the molecule (Figure 1.2). In DNA nucleotides, the two purine bases adenine (A) and guanine (G) as well as the two pyrimidines thymine (T) and cytosine (C) are found. RNA molecules are composed of the same bases, except for thymine, which is replaced by uracil (Figure 1.2). A ribose linked to one of the bases is termed a nucleoside and they are termed adenosine, guanosine, thymidine, cytidine and

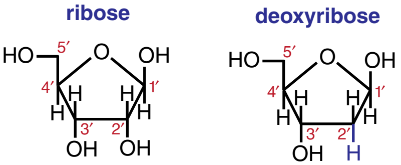

Figure 1.1 Nucleic acids are composed of a sugar backbone and individual bases (see Figure 1.2). A pentose (a sugar molecule that contains five carbon atoms. The carbons are numbered 1′, 2′, 3′, 4′ and 5′; highlighted in red) forms the backbone of DNA as well as RNA. In RNA, the pentose ribose is present. In DNA, deoxyribose is found. Deoxyribose is characterized by the lack of a hydroxyl group at the 2′ carbon position (highlighted in blue).

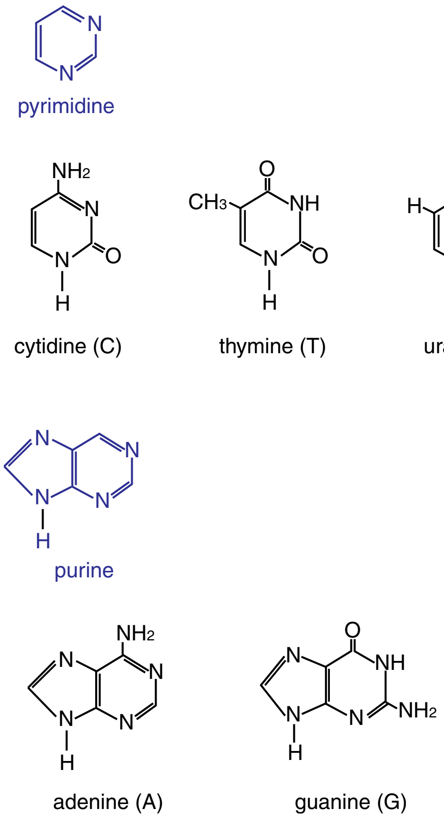

Figure 1.2 Bases present in both ribonucleotides and deoxyribonucleotides. Bases are classified as purines or pyrimidines depending on their structural basis (purine or pyrimidine, indicated in blue). The pyrimidine bases cytidine (C) and thymine (T) are found in DNA. In RNA molecules, thymine is replaced by uracil (U). The pyrimidines adenine (A) and guanine (G) are equally present in DNA and RNA molecules.

uridine. To form a nucleotide, the 5′ carbon of the ribose or deoxyribose reacts with a phosphate group. Therefore, a nucleotide is a nucleoside-5′-phosphate.

Ribonucleotides are not only used as the building blocks for RNA molecules but are also important cellular components with distinct functions. Nucleotides can carry one (e.g., *a*denosine *m*ono*p*hosphate, AMP; a ribose carrying an adenine base on the 1′ carbon and a phosphate group at the 5′ carbon; Figure 1.3), two (e.g., *a*denosine *d*i*p*hosphate, ADP) or three (e.g., *a*denosine *t*ri*p*hosphate, ATP) phosphate groups at their 5′ carbon. When ATP is hydrolyzed to ADP and a phosphate group is liberated (referred to as orthophosphate, p_i) energy is released that is used by many enzymes in almost every cellular pathway. In addition, GTP is also used as structural component for many proteins and enzymes, which are referred to as small GTPases because of their typical sizes and their catalytic activities. In summary, ribonucleotides have various cellular functions besides their role as structural components of DNA and RNA in all kingdoms of life.

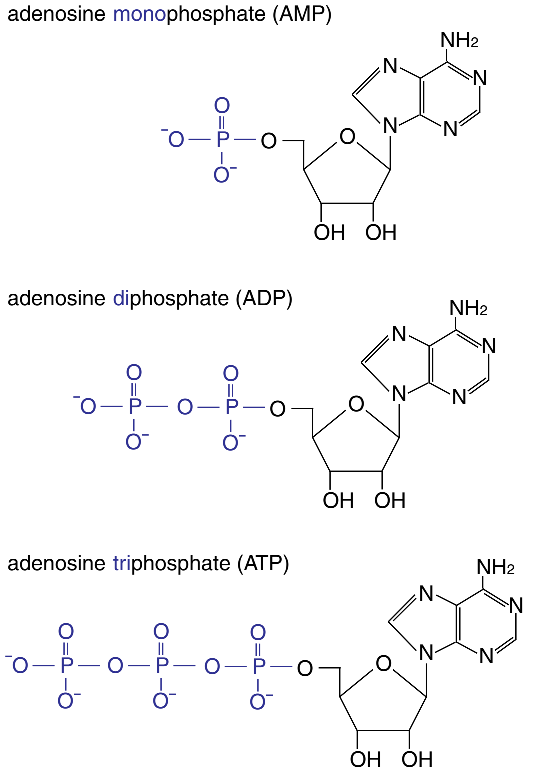
adenosine monophosphate (AMP)
adenosine diphosphate (ADP)
adenosine triphosphate (ATP)
NH2
N
O
P
OH

Figure 1.3 Nucleosides covalently bind to phosphate groups with their 5′ hydroxyl group to form nucleotides. Depending on the number of phosphate groups that are present on the molecules, they are referred to as nucleotide monophosphates (AMP, GMP, UMP, CMP), nucleotide diphosphates (ADP, GDP, UDP, CDP) or nucleotide triphosphates (ATP, GTP, UTP, CTP). Particularly the tri- and diphosphorylated forms are used as energy sources in living cells. The ester bonds between the phosphorus atoms are energy-rich and energy is liberated upon hydrolysis.

1.2 RNA Folding

Ribonucleotides as well as deoxyribonucleotides can use their 5′ phosphate group to form so-called phosphodiester bonds leading to long RNA or DNA molecules (Figure 1.4). Both nucleic acids contain free 5′ ends (the terminal 5′ carbon is not linked via a phosphodiester bond with another nucleotide) as well as free 3′ ends.

DNA molecules are typically double stranded, that is, two single stranded DNA polymers interact with each other to form a double stranded DNA molecule. To form such double stranded DNA, bases form hydrogen bonds with each other. Adenine pairs with thymine and guanine with cytosine. While guanine and cytosine pairing is based on the formation of three hydrogen bonds, adenine and thymine form only two

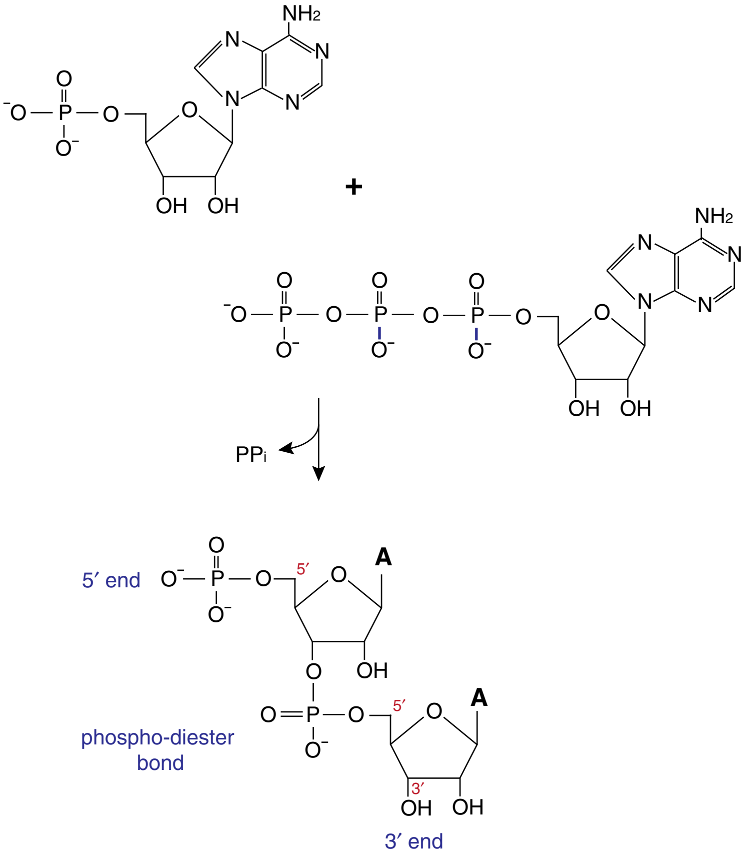

Figure 1.4 Formation of a phospho-diester bond. Two nucleotides can react with each other and form a phospho-diester bond. Adenosine triphosphate reacts with the 3′ carbon of a adenosine monophosphate, leading to the formation of a phopho-diester bond. The resulting dinucleotide possesses a free 5′ end and a 3′ hydroxyl group. Other triphosphonucleotides can react with the 3′ hydroxyl group and extend the RNA chain at the 3′ end.

hydrogen bonds and are therefore thermodynamically less stable (Figure 1.5). James Watson and Francis Crick solved the structure of DNA in the 1950s and established the famous model of the DNA double helix. In the Watson–Crick model, two DNA molecules (also referred to as DNA chains) form a double helix around a common axis. The two DNA chains are antiparallel to each other and the sugar molecules (known as the sugar backbone) define the outer surface of the helix, whereas the bases face each other and form the above-mentioned hydrogen bonds. The DNA double helix is also characterized by a minor and a major groove, which are frequently contact points for proteins interacting with the DNA double helix (Figure 1.5). Depending on the sequence, salt conditions and various other parameters, DNA double helices can adopt different conformations. These forms are called A form (A-DNA), B form (B-DNA) and Z form (Z-DNA) and differ in physical parameters

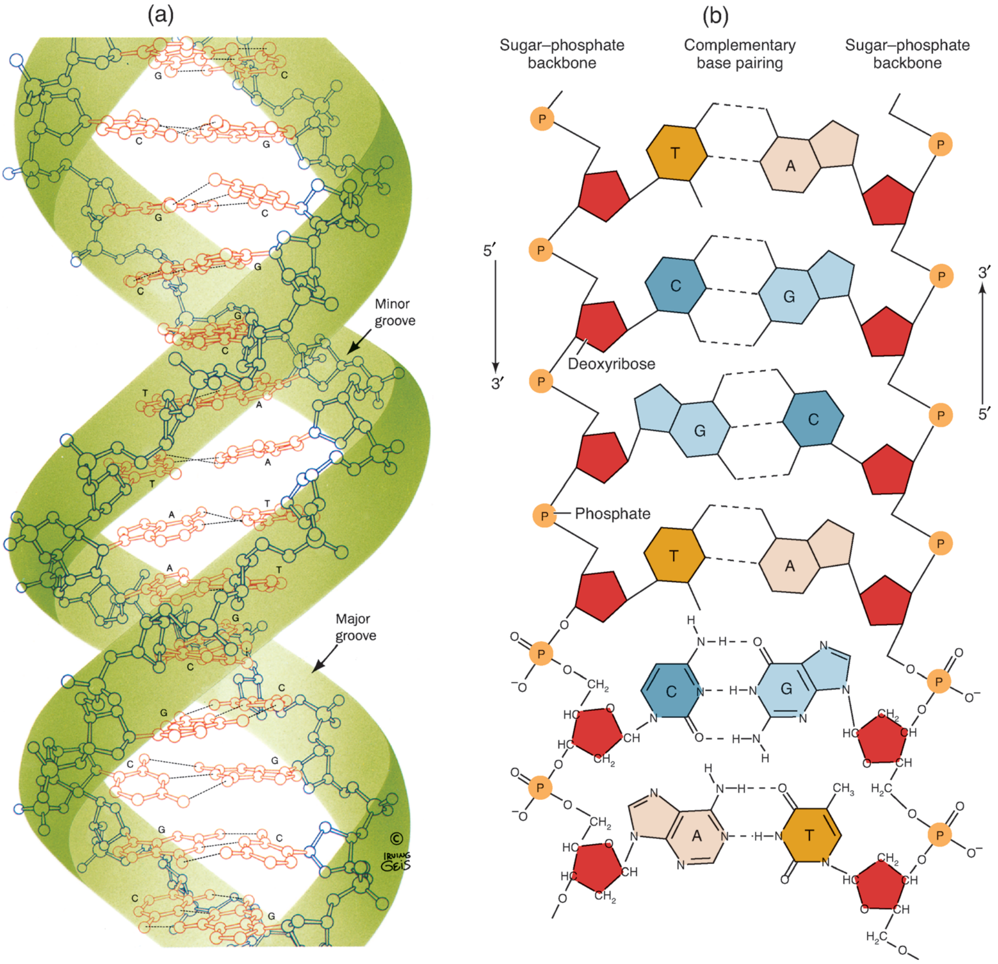

Figure 1.5 Two DNA strands associate to a double stranded DNA helix. (a) The sugar backbone of the DNA helix defines the outer surface of the helix (shown in green). The individual bases face each other and form hydrogen bonds with each other. The two strands are antiparallel, i.e. the ends of the DNA helix contain one 5′ and one 3′ end. The helix is also characterized by major and minor grooves, which are very often interaction platforms for DNA-binding proteins. (b) The base G pairs with C and forms three hydrogen bonds (blue and light blue). A pairs with T and forms two hydrogen bonds (orange and pink). One strand is shown in 5′ to 3′ direction and the other one in 3′ to 5′ direction indicating the antiparallel nature of the two strands.

such as diameter, size of major or minor grooves or the nucleotides that are found in one turn of the helix.

For RNA synthesis, DNA is used as the template that carries the genetic information. Generally, RNA molecules are synthesized by cellular enzymes (so-called RNA polymerases) as single stranded RNA molecules. However, RNAs very often form secondary structures as well. Although it is rather rare in higher eukaryotes, two

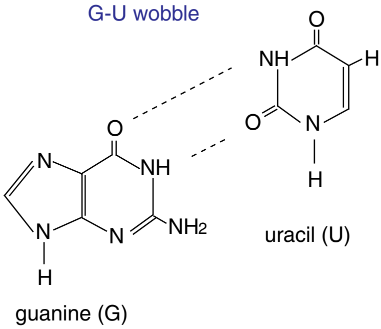

Figure 1.6 RNA helices are often characterized by unusual base pairings, which are referred to as wobble pairings. The most common wobble interactions are G–U wobbles. The guanine base interacts via two hydrogen bonds with uracil.

independent RNA molecules can form double stranded RNAs. Such double stranded RNAs or RNA secondary structures also form a double helix, but because of the different chemical nature of RNAs the RNA double helix differs significantly from the DNA double helix. RNA helices are mainly found in conformations that are similar to the A form of DNA double helices and therefore such double stranded RNA structures are called A-RNA.

In contrast to DNA, where G almost exclusively pairs with C and A with T, so-called wobble base pairs are frequently found in double stranded RNAs. The most common non-canonical interaction is the G–U wobble in which G forms two hydrogen bonds with U (Figure 1.6).

Similar to proteins, which are characterized by primary, secondary, tertiary and quaternary structures, RNAs can also form specific spatial organizations, which are classified as follows [1, 2]:

1) The order of the individual nucleotides can be viewed as the primary RNA structure.
2) Secondary RNA structure: although RNAs are single stranded molecules, they form intramolecular secondary structure. Partially complementary sequences that are present on one molecule can fold back onto each other and form partial double stranded RNA stretches. Single stranded regions are looped out and so-called hairpins (also referred to as stem-loop structures) are formed (Figure 1.7). The double stranded portion is termed the stem and the single stranded part the loop of the hairpin. Such hairpins can vary extensively. Very short hairpins exist in parallel to extensive secondary structures on long RNA molecules [3]. Longer stems of hairpins often contain mismatched regions, which are termed bulges.
3) Tertiary RNA structure: tertiary interactions between more distal domains of one RNA molecule (secondary structures are rather local) are responsible for the formation of a compact three-dimensional structure (Figure 1.7). Such compact folding units are required for the function of catalytic RNAs, for example.
4) Quaternary RNA structure: in some larger RNA protein complexes (RNPs) several RNAs exist. One example for such a large assembly is the ribosome,

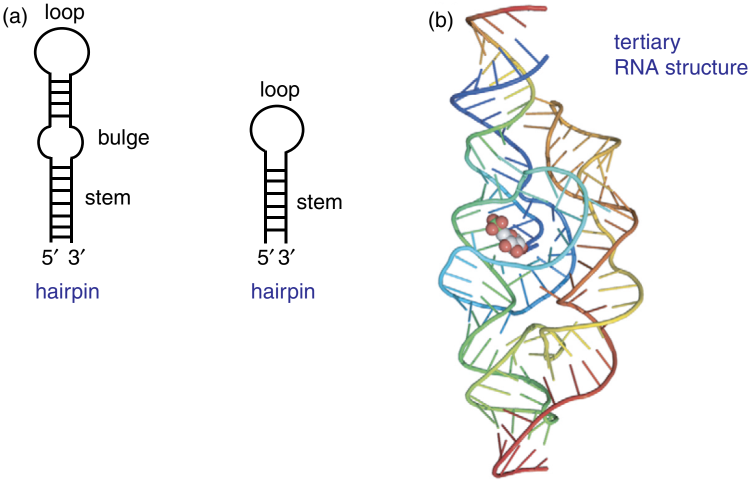

Figure 1.7 Structural organization of single stranded RNA molecules. (a) The most common secondary structure of single stranded RNAs are so-called hairpins or stem–loop structures. Hairpins consist of antiparallel double stranded RNA stems and single stranded loops. Large hairpins very often contain mismatches in their stems leading to bulges of unpaired RNAs. (b) Example of a tertiary RNA structure. In tertiary RNA structures, distant regions of the RNA molecules contact each other and form higher order structures. Secondary structures are typically restricted to more local contacts. The structure represents the bifunctional glmS ribozyme – or riboswitch (for further details, see Chapter 21).

which synthesizes protein products from mRNA templates. The different ribosomal RNAs (rRNAs; for further details, see Chapters 7 and 12) interact with each other and form higher order RNA structures, which are referred to as quaternary RNA structures.

Especially non-coding RNAs as well as the untranslated regions (UTRs) of mRNAs are very often characterized by pronounced secondary folding structures. Viral genomes, which possess limited space to store the genetic information, are characterized by the production of RNAs that contain extensive secondary structures. RNA folding constitutes in most cases a regulatory means and influences gene expression.

1.3
The RNA World Hypothesis

Genetic information is stored in the genome that is composed of DNA. The information is copied into RNA, which serves as template for the ribosome that decodes the genetic information and produces proteins. DNA maintenance, mRNA synthesis and protein production require the action of proteins. An intriguing question in the evolution of life is: what was first? DNA, RNA or proteins? – a classical hen and egg scenario. One concept that tries to explain this problem is the RNA world hypothesis that was put forward by Walter Gilbert in 1986. This model describes a world in which only RNA molecules existed long before DNA and proteins have evolved. The RNA world model implies that DNA molecules evolved from RNA to serve as a storage medium for the precious genetic information. Finally, proteins evolved to function as highly specialized and potent catalytic enzymes. The RNA world hypothesis is based on a number of observations that are essential for life:

1) RNA molecules can have catalytic functions similar to proteins. Catalytic proteins are termed enzymes and catalytic RNAs are referred to as ribozymes (see Chapter 21). Highly specialized RNA molecules can catalyze various chemical reactions.
2) The nucleotide composition of an RNA molecule contains genetic information. Retroviruses, for example, still use RNA as a storage medium for their genetic information.
3) RNAs can also replicate without the help of other factors. Ribozymes can catalyze RNA ligation (two RNAs are fused together) or the addition of single nucleotides to a growing RNA molecule. Even today, replication of DNA requires the help of RNAs, which might be reminiscent of an ancient RNA world.
4) Today, many cellular processes require the action of non-coding RNAs, although the RNAs no longer function as ribozymes.

Taken together, the above-mentioned points strongly suggest that an ancient DNA and protein-free RNA world might have existed during the evolution from simple molecules to complex organisms [4, 5].

1.4 Functions of RNA

In living organisms that are found today, RNA molecules can have various functions.

1) RNAs function as carriers of the genetic information. Such RNAs are termed messenger RNAs (mRNAs).
2) Since RNAs are single stranded, they can use their bases and find complementary nucleic acids (DNAs or RNAs) within a cell. Therefore, they can function as guides that bring proteins to specific target RNAs for modification or destruction.
3) RNAs can also function as storage for genetic information. Viruses such as retroviruses contain a RNA genome that is converted into DNA before it can be integrated into a host genome.
4) RNAs can serve as scaffolds for the assembly of larger RNA–protein complexes.
5) RNAs can function as sensors that measure physical parameters such as temperature, ion or metabolite concentrations (see Chapter 22).

1.5 Protein Classes that are Required for RNA Function

Nucleic acids are associated with proteins within a cell. These proteins can have various functions and many of these functions will be discussed throughout this book. DNA is incorporated into higher order chromatin structures in which the DNA is wrapped around proteins termed histones. Similarly, RNAs are also incorporated into RNA–protein complexes. Such higher order assemblies are generally referred to as *r*ibo*n*ucleoprotein *p*articles or *r*ibo*n*ucleo*p*roteins (RNPs). For example, mRNAs are incorporated into mRNPs at any given time during their life cycles.

1.5.1 RNA Binding Proteins

At least one protein component of a given RNP interacts with the RNA. Therefore, highly specialized RNA binding proteins are important for RNP formation and subsequent RNA function. Depending on the RNA structure that they bind, RNA binding proteins can be divided into single stranded RNA binding proteins and double stranded RNA binding proteins. These two classes of proteins as well as their distinct RNA binding motifs will be introduced in this paragraph.

1.5.1.1 Proteins that Interact with Single Stranded RNAs

In many organisms, proteins with RNA binding activity are very abundant. The most common structural family or protein motif that binds single stranded RNA is the *R*NA *r*ecognition *m*otif (RRM). In human, for example, this motif is present in about 0.5–1.0% of all proteins. RRM motifs are also known as *R*NA *b*inding *d*omains (RBDs). RRMs are characterized by two distinct folding units termed *r*ibo*n*ucleo*p*rotein domain *1* (RNP1) and RNP2, which are located on two antiparallel beta sheets of the RRM (Figure 1.8). Both RNP motifs are required for RNA interaction. Some RRMs bind to single stranded RNA molecules sequence-specifically. Others, however, can have a broader range of target RNAs and the interaction is facilitated independently of the RNA sequence. RRM-containing proteins play crucial roles in almost all pathways is which RNA is involved. It is very common that RRM-containing proteins are composed of arrays of multiple RRMs, which can influence each others RNA binding activity. In addition to RNA binding, RRMs can also serve as protein interaction platforms. The orientation of the helix α1 (Figure 1.8) towards then RNP-containing β-sheets determines, whether a RRM domain interacts with RNA or proteins [6–8].

Another common structural motif that interacts with single stranded RNAs is the so-called KH domain. KH domains were first identified in the protein named *h*eterogeneous *n*uclear *r*ibo*n*ucleo*p*rotein *K* (hnRNP K) and therefore the domain was termed hnRNP *K*-*h*omology (KH) domain. KH domains are ubiquitous and are found in bacteria, archaea and eukaryotes. The domain is composed of about 70 amino acids and a specific sequence element establishes the contact to the

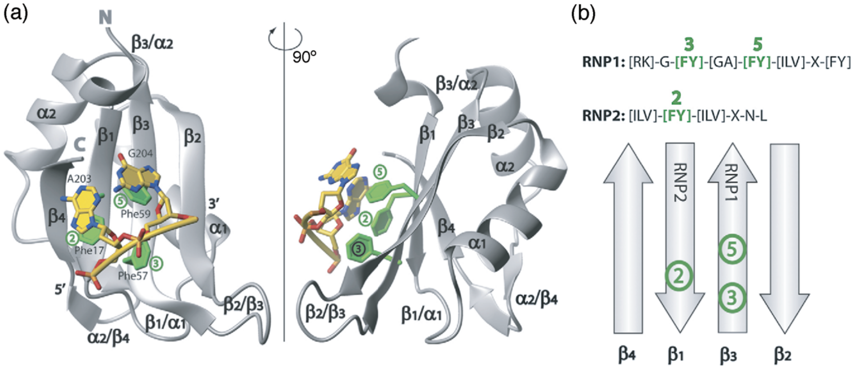

Figure 1.8 Crystal structure of a RRM domain in complex with a short RNA molecule. (a) Structure of one of the hnRNP A1 RRM domains. As apparent from the structure, the contacts between the nucleic acid and the RRM domain are mainly facilitated by β-sheets. (b) Conserved amino acids and structural organization of the RRM domain . The RRM domain contains the RNP1 and the RNP2 motif, which directly contact the nucleic acid. Contacts 2, 3 and 5 shown in (a) are indicated. The figure was reproduced from [7].

nucleic acid. The motif is composed of α-helices and β-sheets and the interaction with the RNA is established mainly by hydrogen bonds (Figure 1.9). KH domains interact with RNA substrates independently of the RNA sequence. KH domains interact both with single stranded DNA and single stranded RNA. Similarly to

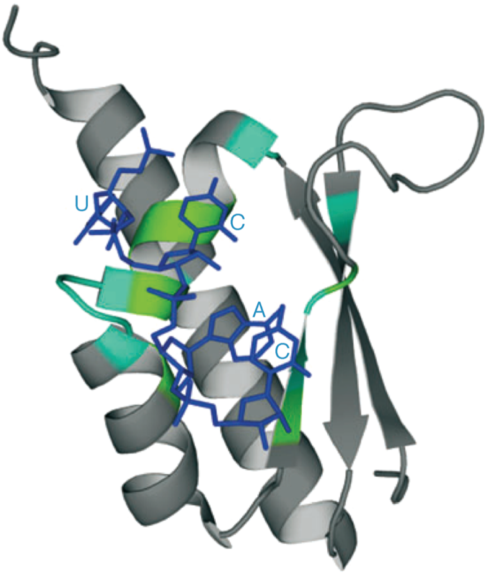

Figure 1.9 Crystal structure of a KH domain bound to RNA. The structure shows one of the KH domains found in a eukaryotic protein termed Nova-2. It is associated with the tetranucleotide sequence 5′-UCAC. The figure was reproduced from [9].

RRMs, KH domains are frequently found in multiple copies on one RNA binding protein [8, 9].

1.5.1.2 Proteins that Interact with Double Stranded RNAs

As mentioned above, RNAs can form secondary structures and such folding states often contain stretches of double stranded RNA (Figure 1.7). Furthermore, complementary RNAs are also capable of forming double stranded RNA. Consequently, proteins have evolved that recognize and specifically bind double stranded RNAs. Such proteins are particularly important for the diverse functions of non-coding RNAs (see Part Two of this book). Most double stranded RNA binding proteins are characterized by *d*ouble stranded *R*NA *b*inding *d*omains (DRBDs). Proteins that contain DRBDs are found in bacteria, archaea and eukaryotes. Even various viruses encode double stranded RNA binding proteins. The DRBD consists of about 65–68 amino acids and some proteins contain more than one DSRBD. Viral proteins, however, usually contain only one [10].

1.5.2
RNA Helicases

RNAs are capable of forming extensive secondary and tertiary structures. However, for specific RNA functions, for example, translation of mRNAs into proteins, such secondary structures have to be removed. Moreover, whole RNPs are constantly remodeled and proteins associate with or dissociate from RNPs. For this purpose, a specific protein class, so-called RNA helicases, has evolved. RNA helicases are defined based on their ability to utilize the energy of ATP hydrolysis to unwind RNA duplexes. RNA helicases are therefore occasionally termed unwindases.

Helicases (RNA or DNA helicases) are characterized by a number of distinct domains and motifs and can be classified into five different superfamilies, designated SF1–SF5. Most RNA helicases, however, belong to SF2 and only a few to SF1. RNA helicases contain seven to nine conserved motifs that constitute the helicases core. Most RNA helicases include either a DEAD box, a DEAH box or a DExD/H box. Many RNA helicases are therefore commonly referred to as DEAD, DEAH or DExD/H box helicases.

The conserved sequence motifs are located in two different domains: motifs I, Ia, Ib, II and III form domain 1. Motifs IV, V and VI give rise to domain 2 (Figure 1.10). Motifs I and II are also termed Walker A and Walker B motifs after the discoverer John E. Walker, who received the Nobel prize in chemistry in 1997. These motifs are not specific to RNA helicases and are also found in other ATP hydrolyzing proteins. Both motifs are required for binding and hydrolyzing ATP. Motif III couples ATP hydrolysis with the helicases activity, motif VI has a role in RNA binding and ATP hydrolysis and motifs Ia, Ib, IV and V have a role in RNA substrate binding. Recently, another motif has been identified, which was termed Q motif because of an invariant glutamine (Q). The Q motif is specific to DEAD box helicases and regulates ATP binding and hydrolysis.

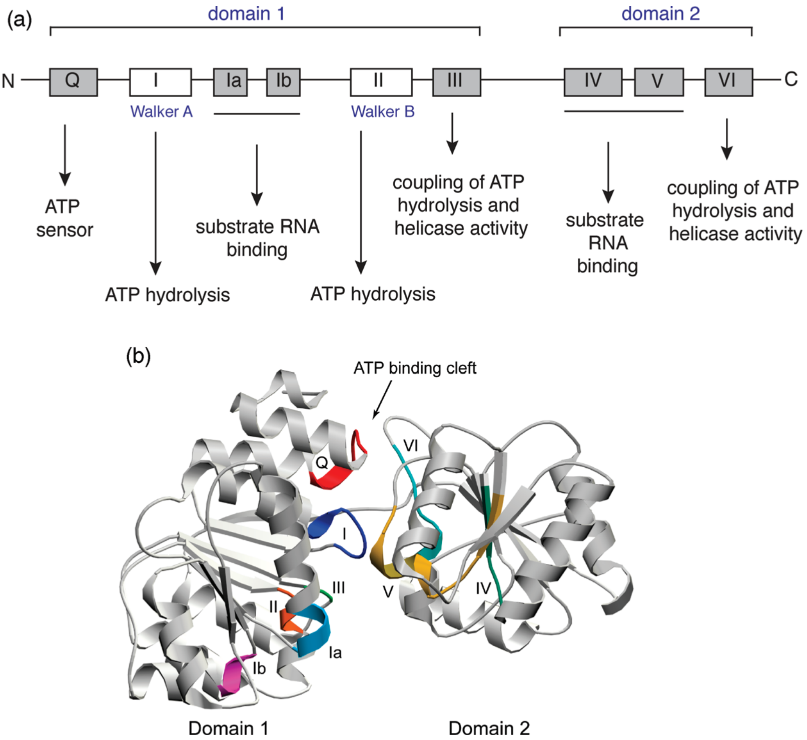

Figure 1.10 Domain architecture and X-ray structure of the DEAD box motif of RNA helicases. (a) Schematic representation of the motifs found in RNA helicases. The motif II (Walker B) contains the DEAD (named after the amino acids Asp-Glu-Ala-Asp) box. The helicase domain is separated into domain1 and domain 2. The figure was modified from [11]. (b) Crystal structure of the *Methanococcus jannashii* DEAD box protein. The conserved motifs shown in (a) are highlighted with different colors. The figure was created with molscript (http://www.avatar.se/molscript/) and POVRAY (http://www.povray.org/) using the coordinates deposited in the Protein Data Bank (1HV8).

RNA helicases remove secondary structures from single stranded RNA molecules. RNA helicases generally catalyze the dissociation of two complementary RNA strands from each other. However, RNA helicases have also an important RNP remodeling function. It has been demonstrated that RNA helicases can actively remove RNA binding proteins from RNA molecules. Therefore, taken together, RNA helicases have dual roles in RNA metabolism: (i) such enzymes can unwind double stranded RNAs, (ii) RNA helicases remove proteins from RNA molecules [11, 12].

References

1 Holbrook, S.R. (2008) Structural principles from large RNAs. *Annu. Rev. Biophys.*, **37**, 445–464.

2 Li, P.T., Vieregg, J., and TinocoJr., I. (2008) How RNA unfolds and refolds. *Annu. Rev. Biochem*, **77**, 77–100.

3 Svoboda, P. and Di Cara, A. (2006) Hairpin RNA: a secondary structure of primary importance. *Cell Mol. Life Sci.*, **63**, 901–908.

4 Spirin, A.S. (2002) Omnipotent RNA. *FEBS Lett.*, **530**, 4–8.

5 Altman, S. (2007) An overview of the RNA world: for now. *Biol. Chem.*, **388**, 663–664.

6 Hall, K.B. (2002) RNA-protein interactions. *Curr. Opin. Struct. Biol.*, **12**, 283–288.

7 Clery, A., Blatter, M., and Allain, F.H. (2008) RNA recognition motifs: boring? Not quite. *Curr. Opin. Struct. Biol.*, **18**, 290–298.

8 Lunde, B.M., Moore, C., and Varani, G. (2007) RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.*, **8**, 479–490.

9 Valverde, R., Edwards, L., and Regan, L. (2008) Structure and function of KH domains. *FEBS J.*, **275**, 2712–2726.

10 Saunders, L.R. and Barber, G.N. (2003) The dsRNA binding protein family: critical roles, diverse cellular functions. *FASEB J.*, **17**, 961–983.

11 Bleichert, F. and Baserga, S.J. (2007) The long unwinding road of RNA helicases. *Mol. Cell*, **27**, 339–352.

12 Rocak, S. and Linder, P. (2004) DEAD-box proteins: the driving forces behind RNA metabolism. *Nat. Rev. Mol. Cell Biol.*, **5**, 232–241.