Bioinformatics

Bearbeitet von
Andrzej Polanski, Marek Kimmel

1. Auflage 2007. Buch. xvii, 376 S. Hardcover
ISBN 978 3 540 24166 9
Format (B x L): 15,5 x 23,5 cm
Gewicht: 754 g

10

RNA

Ribonucleic acid (RNA), is a polymer of repeating units, namely ribonucleotides, with a structure analogous to single-stranded DNA. It has a backbone composed of sugars (riboses) and phosphate groups, with organic nitrogen bases bonded along it. The differences between RNA and DNA (see Fig. 8.7) concerning their components are (i) sugar deoxyribose which appears in DNA is replaced in RNA by another sugar, ribose, and (ii) the organic base thymine (T) that appears in DNA is replaced in RNA by another organic base, uracil (U). Compared with DNA, RNA molecules are less stable and exhibit more variability in their three dimensional-structure which underline their different function.

In living organisms, RNA arises in the process of transcription which involves creating single-stranded RNA, based on a DNA template according to the complementarity pairing rules A–U, C–G, $G - C$ and $U - A$. Similarly to DNA, the RNA chain has a direction, two ends of RNA are labeled $5'$ and $3'$. For example, an RNA strand copied from the left strand of DNA piece from Fig. 8.4 will have the sequence of bases $3'$–$UGACUG$–$5'$. By the mechanism of transcription, many copies of RNA corresponding to one piece of DNA can be created. The process of transcription is catalyzed by an enzyme RNA polymerase, which performs two functions: it unwinds the DNA (separates the two DNA strands), and slides along the DNA strand, forming RNA according to the complementarity rules. In eukaryotic organisms, there are various types of RNA polymerase, specializing in producing different types of RNA [118, 216, 287].

In the central dogma of molecular biology illustrated in Fig. 8.6, RNA serves mainly as a carrier of information from the DNA to the ribosomes, where it is utilized for protein construction. However, recent developments in molecular biology suggest that the importance of the various kinds of RNA molecule for the development and functioning of living organisms has not yet been sufficiently appreciated. In the forthcoming sections, we will mention these arguments and reference some related recent publications.

Experimental research on the structure and functions of RNA, proceeds along many of the directions that we have were already sketched for DNA and proteins. These include electrophoresis techniques for estimating the length of molecules, sequencing, blots, and X-ray diffraction and NMR methods for analysis of the three-dimensional structure. These methods must also be supported by appropriate developments in bioinformatics, to order, organize, and make accessible large amounts of data. Information on the sequences of ribonucleotides in RNA, functions of RNA molecules and their spatial structures are available in appropriate databases, coding RNA sequences are in gene banks [326], some recently established databases for non coding sequences of RNA can be found in [106, 181, 318, 334], and databases of spatial structures of RNA can be found in [321, 329]. The mathematical and computational aspects of research on RNA are also in most respects parallel to those for DNA and/or proteins, concerning inference on function by aligning sequences (as in the case of DNA and proteins) and studying and predicting the three-dimensional shapes of RNA molecules associated with efforts to relate the function of RNA to its spatial conformation.

10.1 The RNA World Hypothesis

The Processes involving replication, transcription, and translation are catalyzed by proteins. On the other hand proteins are constructed on the basis of the information written in DNA and carried by mRNA. This poses an evolutionary puzzle about how this functional organization evolved. Early theories gave prominence to amino acids and short peptides, as the earliest molecules in evolution. However, explaining the evolutionary scenario by the protein-first hypothesis suffers from serious problems related mainly to the lack of molecular mechanisms for the self replication of proteins. The present hypothesis, called the RNA world hypothesis, is that in the process of evolution, RNA molecules preceded DNA strands and proteins [98]. The scenario of self catalytic replication of RNA in the early stages of evolution is being reproduced in several laboratory experiments, see, for example, [277] and references therein.

10.2 The Functions of RNA

On the basis of on their functions, RNA molecules can be divided into two groups, coding and noncoding. The coding RNA is messenger RNA (mRNA) assembled by the machinery of (eukaryotic) cell during the processes of transcription, splicing, and polyadenylation. The order of the codons formed by the ribonucleotides in mRNA corresponds to the order of the amino acids in the linear content of proteins.

Noncoding RNA is transcribed from noncoding regions of DNA. Classically, there are two types of noncoding RNA, transfer RNA (tRNA) and ribosome RNA (rRNA). The transfer RNAs are short-chain RNA molecules (74–93 ribonucleotides) involved in transporting amino acid molecules to ribosomal sites, where the process of growing polypeptide chains occurs. Ribosome RNA (rRNA) is a component of ribosomes.

Recently there has been a great interest in RNA transcribed from noncoding regions of DNA, leading to the development of a much broader taxonomy of noncoding RNA. This direction of development is partly related to recent theories of the "dark matter of noncoding DNA and RNA". According to these theories, noncoding DNA is not "evolutionary junk" in the interpretation based on the classical molecular-evolutionary viewpoint, but rather has some important, not yet discovered, function in organisms [186, 187, 198]. One of the chief arguments supporting these theories is based on comparison of genomes of simple and complex organisms. The scale of complexity, seemingly, is not related to the number of genes nor to the number of chromosomes, but correlates better with the size of noncoding DNA. The conclusion is that non coding DNA and the non coding RNA transcribed from it are responsible for important functions of organisms, which are behind their diversification. Examples of the functions of noncoding RNAs are [186, 187, 198] regulatory functions in the process of gene expression, maintaining the telomeres, gene splicing, and chemical modification of ribosomal RNA. Organizing information about RNA families and their functions in a systematic way involves creating electronic databases that allow the relevant data to be deposited and downloaded. The Rfam database, [106, 334] contains curated lists of noncoding RNA families, classified by aligning their sequences.

10.3 Reverse Transcription, Sequencing RNA Chains

The standard direction of information flow is copying from the DNA template to RNA. However, the opposite direction, called reverse transcription, is also possible. This process is performed by the enzyme reverse transcriptase. In nature, reverse transcription is often met with in retroviruses, which appear to consist of two or more RNA molecules and attack the host cells by a self-replication strategy based on reverse transcription of their sequence to the host's genome. A very well known example is the HIV virus.

In the area of scientific research, reverse transcription is most often used for sequencing RNA by use of polymerase chain reaction (PCR) as in the case of DNA. The PCR can only be applied to DNA. By using reverse transcription, RNA can be first copied to DNA and then amplified (replicated) by means of the PCR.

10.4 The Northern Blot

The Northern blot is a technique analogous to the Southern blot described in Chap. 8. RNA from a specimen is separated by electrophoresis and fixed on a supporting plate. In the next step, single-stranded DNA fragments complementary to the specific mRNA being sought are labeled with radioactive atoms and then hybridized to the immobilized mRNA. If the specific mRNA is present, a radioactive band is detected. The name "Northern blot" was invented by altering the term "Southern blot" used for DNA assays.

10.5 RNA Primary Structure

The primary structure of RNA is the linear sequence of ribonucleotides. Comparative sequence analysis of ribonucleotides is a basic premise for the classification of RNA and prediction of its function and structure. RNA databases [334] contain a large number RNA sequences, enabling newly discovered sequences to be compared with existing RNA families.

10.6 RNA Secondary Structure

The shape of RNA is to a substantial extent formed by a series of hydrogen bonds that occur between the nitrogen bases in its backbone. These bonds result in the formation of characteristic motifs, shared by many RNA molecules, which can be represented graphically by using two-dimensional plots. Graphical representation of RNA including these motifs is called the secondary structure of RNA. The motifs encountered in RNA strands are, hairpin loops, internal loops, multibranched loops, bulges, and stems. An example of a secondary structure of an RNA strand is presented in Fig. 10.1. Hydrogen bonds occur between the complementary (Watson Crick) pairs of bases in RNA A–U and C–G. Additionally, a bond between G and U is also energetically favorable and is often encountered in RNA molecules.

10.7 RNA Tertiary Structure

The formation of the spatial (tertiary) structure of RNA molecules is related to the occurrence of hydrogen bonds between bases, other than those responsible for the formation of the secondary structure. These additional bonds contribute to the final 3D shape of RNA molecules. The tertiary structure of RNA can be established experimentally by X-ray diffraction of crystallized molecules or by NMR techniques. After the spatial shape has been computed, the hydrogen bonds that stabilize the spatial form of the molecule can be

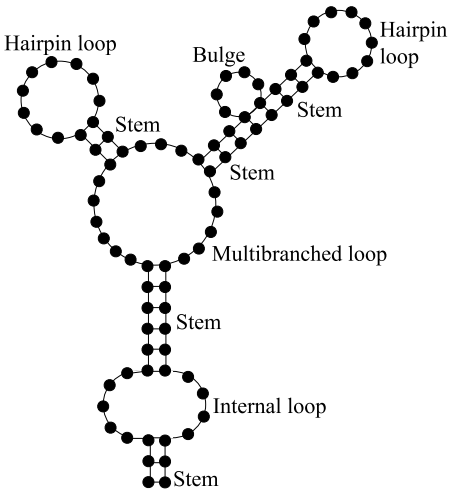

Fig. 10.1. Motifs in RNA secondary structure: stem, hairpin loop, bulge, multibranched loop, and internal loop. Nitrogen bases are represented by small circles and hydrogen bonds between bases are depicted by thin, ladder-like line segments

found by analysis of the distances between the bases and their orientations [286, 19].

The motifs encountered in RNA tertiary structure and their occurrence in RNA species are stored in the SCOR database [153, 338]. Among the motifs in RNA spatial structure probably most important are coaxial bonds, which contribute to the formation of helical structures in RNA molecules. Other motifs include pseudoknots, kissing hairpin loops, and ribose zippers.

10.8 Computational Prediction of RNA Secondary Structure

The most reliable approach for prediction secondary structure, especially for long RNA sequences, is comparative sequence analysis discussed later in this chapter. This approach involves alignment of multiple RNA sequences and uses a covariance-type analysis aiming at identification of conserved base-pairing interactions in the RNA. Most of the secondary structures of long RNA sequences, accepted by experts and available via internet were obtained by using the method of comparative sequence analysis.

In cases where multiple RNA sequences either are not available or do not have enough diversity for comparative analysis, an alternative method of prediction of RNA folding is energy minimization. Although it is not as accurate as comparative analysis, it leads to useful predictions and can be applied, for

example, when one wants to estimate quickly the second-order structure of a single RNA strand. Numerical minimization of the folding energy is performed by using the principle of dynamic programming. Below, we describe two basic algorithms in detail and also mention some other variants. The first approach, often called the Nussinov algorithm [210] simplifies the energy minimization problem by using the hypothesis that the more pairings there are between bases in an RNA strand, the lower the energy of the molecule. The second approach, [241, 281, 298, 299] assigns thermodynamic energies to motifs in the secondary structure. These energies depend on the size of the motifs, and the overall energy of an RNA strand is the sum of the energies of the motifs. The thermodynamic energies of the motifs have been tabulated [253] and they allow modeling the true ratios of the energy components. This enables, for example, the influence of the temperature on the shape of an RNA molecule to be modeled and studied.

10.8.1 Nested Structure

An important property of RNA secondary structures is that there are no crossings between bonds; in other words the structures are nested. The nested character of the secondary shape of RNA is presented in Fig. 10.2. In the upper part of this figure, a fragment of RNA of "hairpin" shape is presented and then in the lower part chain of bases is straightened out. Bonded bases are connected by half circles. The nested structure requires that the half circles cannot make crossings.

Observe that if there is at least one bond in an RNA chain then the secondary structure of this RNA string contains at least one hairpin loop. Contemplating the lower part of Fig. 10.2, one can notice an analogy between defining a nested structure of bonds between bases in RNA and arranging left and right parentheses in the correct order [104]. There is also a correspondence between nested secondary RNA structures and unlabeled trees, whose terminal leaves correspond to unbonded bases and whose topology reflects the structure of the bonds [281].

10.8.2 Maximizing the Number of Pairings Between Bases

As already said, the secondary structure of RNA is formed by series of bonds between bases. We assume that the number of bases in an RNA strand is N and we denote the sequence of the bases by $b_1, b_2, \ldots, b_N$. In this subsection we describe an algorithm for maximizing the number of pairings in the sequence $b_1, b_2, \ldots, b_N$ [210, 281]. It does not change the structure of the algorithm if we assume, more generally, that the score for a pair b_k–b_l is $s(b_k, b_l)$ and we maximize the sum of scores over all possible nested second-order foldings. This formulation becomes equivalent to maximizing the number of pairings when we take $s(A, U) = s(C, G) = s(G, U) = 1$ for energetically favorable pairs, and $s(b_k, b_l) = -\infty$ for all pairs b_k–b_l other than A–U, C–G, and G–U.

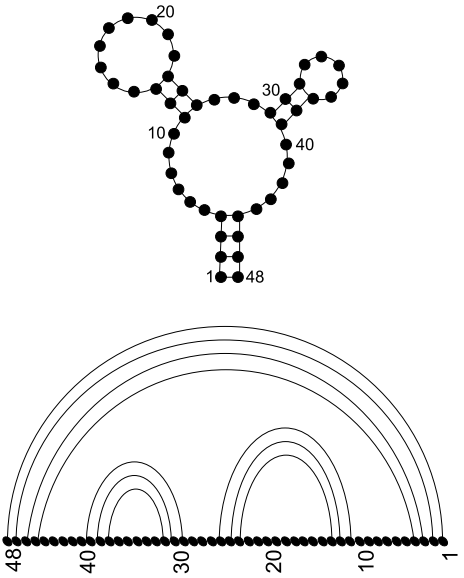

Fig. 10.2. Illustration of the nested property of the bonds between bases in the secondary structure of RNA. The *upper part* presents an example of the fragment of an RNA chain. In the *lower part*, the RNA chain is straightened out and bonded bases are connected by half circles

We introduce two triangular N-dimensional matrices $V(i,j)$ and $W(i,j)$, $i \leq j$, with the following meaning: $V(i,j)$ is the score of the best folding of the RNA subsequence $b_i, b_{i+1}, \ldots, b_j$, given that bases b_i and b_j form a bond, and $W(i,j)$ is the score of the best folding of the RNA subsequence $b_i, b_{i+1}, \ldots, b_j$ (no matter whether b_i and b_j are paired or not).

We shall state and explain recursions for $V(i,j)$, and $W(i,j)$. We start from the case where it is given that b_i and b_j form a bond. We then have

$$V(i,j) = s(b_i, b_j) + W(i+1, j-1). \tag{10.1}$$

In order to formulate a recursive relation for $W(i,j)$ one has to consider the following possibilities: (1) b_i and b_j form a bond, in which case $W(i,j) = V(i,j)$, and (2) b_i and b_j do not form a bond. In case (2), the nested property of bondings described in Sect. 10.8.1 guarantees that the strand $b_i, b_{i+1}, \ldots, b_j$ can be split into two strands $b_i, b_{i+1}, \ldots, b_k$ and $b_{k+1}, b_{k+2}, \ldots, b_j$ such that there are no bonds between them, and consequently their total score is $W(i,k) + W(k+1,j)$. Summing up (1) and (2) we obtain

$$W(i,j) = \max\{V(i,j), \max_{i \leq k < j} [W(i,k) + W(k+1,j)]\}. \tag{10.2}$$

The recursions (10.1) and (10.2) can easily be shrunk into one,

$$W(i,j) = \max\{s(b_i, b_j) + W(i+1, j-1), \max_{i \le k < j}[W(i,k) + W(k+1,j)]\}. \quad (10.3)$$

Starting from $W(i,i) = 0, i = 1, 2, \ldots, N$, and $W(i, i+1) = 0, i = 1, 2, \ldots, N-1$ and then using (10.3) we can fill in all entries of $W(i,j)$ $i = 1, 2, \ldots N$, $i \le j$.

The algorithm (10.3) has a complexity of order $O(N^3)$ since filling in each entry of $N \times N$ matrix requires running the index k over a range $O(N)$.

10.8.3 Minimizing the Energy of RNA Secondary Structure

Predicting the secondary structure of RNA only by maximizing the number of pairings between bases is an oversimplification. Results closer to the shapes found in experiments are obtained by assigning energies to the motifs of RNA secondary structure, i.e., stems, hairpin loops, bulges, internal loops, and multibranched loops, and searching for the structure with the lowest energy. This more accurate model assumes that RNA folding stems from an interplay between the stabilizing role of base pairings and the destabilizing effects of unpaired segments of hairpin loops, bulges, internal loops, and multibranched loops.

Hairpin RNA Structure

Let us start by describing an algorithm for minimizing energy of an RNA strand under some restrictions. Namely, we aim at minimizing the energy of RNA sequence of bases $b_1, b_2, \ldots, b_N$, assuming additionally that (i) b_1 and b_N are paired, and (ii) the possible secondary motifs are hairpin loops, stems, bulges, and internal loops. We call such a structure a hairpin RNA structure. Since multibranched loops are excluded, hairpin structures do not branch. An example of a hairpin RNA structure is shown in Fig. 10.3. As seen in this figure, a hairpin RNA structure can be thought of as a sequence of motifs, such as those shown, $stem1$, $iloop1$, $stem2$, $bulge1$, $stem3$, and $hloop1$. Each of the possible motifs is characterized by its size:

- $stem(k)$–a stem of k consecutive pairs of bases;
- $bulge(1,k)$, $bulge(k,1)$–right and left bulge of k bases,
- $iloop(k_1, k_2)$–an internal loop of k_1 bases on the left and k_2 bases on the right;
- $hloop(k)$–a hairpin loop of k bases.

The energies of motifs have been experimentally measured and their values are available in the literature and on the Internet [253, 344]. The energies of motifs depend not only on their size but also on the composition of bases, and it is necessary to specify the location of a motif relative to the sequence of RNA bases $b_1, b_2, \ldots, b_N$. It is convenient to introduce the concept of motifs of type 1 and type 2 (based on the presentation in [241]). A motif of type

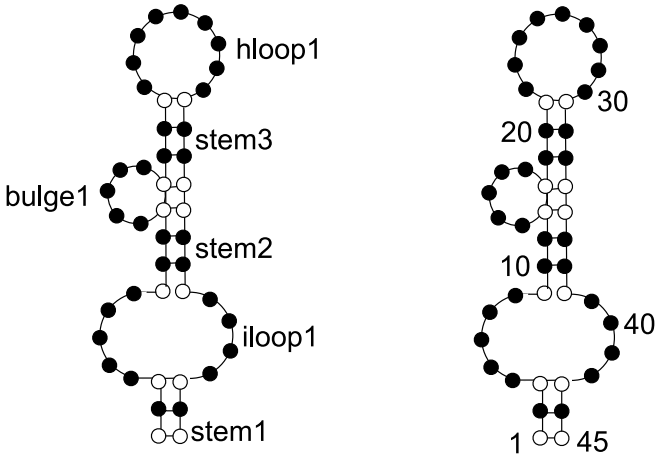

Fig. 10.3. A hairpin structure in RNA secondary folding. On the *left-hand* side, the sequence of motifs of the hairpin loop is marked *stem*1, *iloop*1, *stem*2, *bulge*1, *stem*3, *hloop*1. On the *right-hand* side, the numbers of ribonucleic bases are depicted, to be used in the text

1 is a hairpin loop. It is fully specified by a pair of indices i_s, j_s of bases, $i_s, j_s \in 1, 2, \ldots, N$, and is denoted by

$$M1(i_s, j_s). \tag{10.4}$$

The motifs of type 2 are stems, bulges, and internal loops. These motifs are defined by specifying indices of paired bases, $i_s, j_s, i_t, j_t \in 1, 2, \ldots, N$, where pairings are i_s–j_s and i_t–j_t, and is denoted by

$$M2(i_s, j_s, i_t, j_t). \tag{10.5}$$

In (10.4) and (10.5), the subscript s stands for "start" and the subscript t stands for "termination". In Fig. 10.3 the ribonucleotides corresponding to the indices i_s, j_s, i_t, j_t are marked in white. The hairpin loop *hloop*1 is a motif of type 1, and $hloop1 = M1(21, 31)$; the stem *stem*1 is a motif of type 2, and $stem1 = M2(1, 45, 3, 43)$, and the *bulge*1 is a motif of type 2, and $bulge1 = M2(12, 35, 18, 34)$. We denote energies of motifs (10.4) and (10.5) by

$$EM1(i_s, j_s) \tag{10.6}$$

and

$$EM2(i_s, j_s, i_t, j_t), \tag{10.7}$$

respectively Since the motifs in a hairpin RNA structure appear sequentially, arranging a dynamic programming recursion for minimization of its energy is particularly easy. We denote by $V(i, j)$ the lowest folding energy of the RNA

hairpin structure, over the sequence $b_i, b_{i+1}, \ldots, b_j$, with b_i and b_j paired. We can then state the following recursion for $V(i,j)$:

$$V(i,j) = \min \begin{cases} EM1(i,j), \\ \min_{i_t,j_t}[EM2(i,j,i_t,j_t) + V(i_t,j_t)]. \end{cases} \cdot \tag{10.8}$$

In (10.8), the indices run over the ranges $1 \leq i +$ minimal size of $M1 < j \leq N$ and $i < i_t +$ minimal size of $M1 < j_t \leq j$, and the values of $V(i,j)$ on the diagonal and neighboring positions are initialized at $+\infty$, $V(i,i) = V(i,i+1) = \ldots = V(i, i +$ minimal size of $M1) = +\infty$.

The order of complexity of the algorithm (10.8) is $O(N^4)$, because filling in each of the entries of $V(i,j)$ requires $O(N^2)$ operations, following from minimization over two indices i_t, j_t.

RNA with Multibranched Loops

In our notation multibranched loops are motifs of types 3, 4, and so forth. Assume that in the RNA sequence $b_1, b_2, \ldots, b_N$ ending bases are paired and that possible motifs are type 1 (10.4), type 2 (10.5), and type 3, denoted analogously to (10.4) and (10.5) by

$$M3(i_s, j_s, i_t, j_t, i_q, j_q) \tag{10.9}$$

and having an energy

$$EM3(i_s, j_s, i_t, j_t, i_q, j_q). \tag{10.10}$$

We denote by $V(i,j)$ the minimal energy of the RNA strand $b_i, b_{i+1}, \ldots, b_j$ with b_i and b_j paired. The recursion analogous to (10.8) for $V(i,j)$ is

$$V(i,j) = \min \begin{cases} EM1(i,j), \\ \min_{i_t,j_t}[EM2(i,j,i_t,j_t) + V(i_t,j_t)], \\ \min_{i_t,j_t,i_q,j_q}[EM3(i,j,i_t,j_t,i_q,j_q) + V(i_t,j_t) + V(i_q,j_q)]. \end{cases} \tag{10.11}$$

In (10.11), updating $V(i,j)$ requires minimization over four indices, so the overall complexity of the recursion (10.11) is $O(N^6)$.

Comparing (10.11) and (10.8), one can see that adding loops with more branchings, given by motifs of type 4 and higher, will lead to recursions of successively higher complexity. In practical calculations related to the minimization of RNA folding energies, expressions for energies such as (10.10) are not, however, used because there are not enough experimental data describing the exact energies of multibranched loops. Instead, the energies of multibranched loops are approximated by sums of components related to generating a multibranched loop (M), closing base pairs (P), and unpaired bases inside a loop (Q). So e.g., the energy of multibranched loop $M3$ from (10.10) will be approximated by

$$EM3(i_s, j_s, i_t, j_t, i_q, j_q) = M + 3P + Q(i_t - i_s + i_q - j_t + j_s - j_q), \tag{10.12}$$

where M, P, and Q are appropriate coefficients, and $Q(.)$. Analogous formulas hold for motifs of type 4 and higher.

Using the approximation (10.12), we can simplify the recursion for minimizing the energies of RNA foldings with multibranched loops. Assume that in the RNA sequence $b_1, b_2, \ldots, b_N$ ending bases are paired and we allow motifs of all types. Denote by $V(i,j)$ the minimal energy of the RNA strand $b_i, b_{i+1}, \ldots, b_j$ with b_i and b_j paired, and by $W(i,j)$ the minimal energy of the strand $b_i, b_{i+1}, \ldots, b_j$ inside a multibranched loop. Then, for $V(i,j)$, we have a recursion

$$V(i,j) = \min \begin{cases} EM1(i,j), \\ \min_{i_t,j_t}[EM2(i,j,i_t,j_t) + V(i_t,j_t)], \\ M + P + \min_k[W(i+1,k) + W(k+1,j-1)]. \end{cases} \tag{10.13}$$

In the above, the first and second row are the same as in (10.8). The third row stems from inserting a multibranched loop into the RNA secondary structure. The components M and P are related to the energy of creation of the multibranched loop and to the energy of the closing pairing i–j. The term in the third row is also called a bifurcation, because the secondary structures related to $W(i+1,k)$ and $W(k+1,j-1)$ will fold independently one of another. The recursions for $W(i,j)$ are as follows:

$$W(i,j) = \min \begin{cases} P + V(i,j), \\ Q + W(i+1,j), \\ Q + W(i,j-1), \\ \min_k[W(i,k) + W(k+1,j)]. \end{cases} \tag{10.14}$$

In the above the first row is related to closing the multibranched loop by a pairing i–j, the second and third rows represent leaving the ith and jth base, respectively, inside the loop unpaired, and the fourth row is related to adding a new bifurcation. The computational complexity of the algorithm (10.13) and (10.14) is $O(N^4)$. Before starting the recursions, one must initialize $V(i,j)$ and $W(i,j)$ so that the diagonal and neighboring positions are initialized to $+\infty$, i.e., $V(i,i) = V(i,i+1) = \ldots = V(i, i + \text{minimal size of } M1) = +\infty$, $W(i,i) = W(i,i+1) = \ldots = W(i, i + \text{minimal size of } M1) = +\infty$.

External Bases

Up to now, when discussing minimizing the folding energy of RNA, we assumed that the two terminating bases of the strand were paired, which is not the most general situation. The ending bases $b_1, \ldots, b_p$ and $b_{N-p}, \ldots, b_N$ of the RNA strand $b_1, b_2, \ldots, b_N$ may not form pairings. We shall call the unpaired bases $b_1, \ldots, b_p, b_{N-p}, \ldots, b_N$ of the RNA strand $b_1, b_2, \ldots, b_N$ external bases. Generalizing the algorithm (10.13) and (10.14) to the case of external bases is possible and involves adding one more score matrix $W^E(i,j)$, with the same recursion scheme as in (10.14)

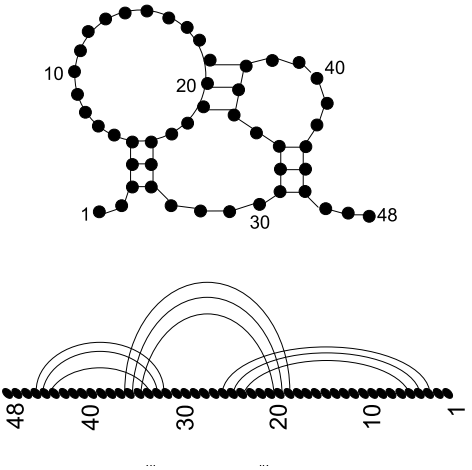

Fig. 10.4. Illustration of the secondary structure in RNA called a pseudoknot. The *upper part* shows the shape of a pseudoknot. The *lower part* illustrates the nonnested character of a pseudoknot

$$W^E(i,j) = \min \begin{cases} P^E + V(i,j), \\ Q^E + W^E(i+1,j), \\ Q^E + W^E(i,j-1), \\ \min_k [W^E(i,k) + W^E(k+1,j)]. \end{cases} \tag{10.15}$$

The difference between (10.14) and (10.15) is in the values of the constants P, Q and P^E, Q^E. For external bases, reasonable values for the parameters are $P^E = Q^E = 0$.

10.8.4 Pseudoknots

In Fig. 10.4 we present a motif of RNA structure, called a pseudoknot, which has not yet been discussed. Pseudoknots are not found in short RNA chains but they can form in longer RNA molecules. The RNA folding is stabilized here by additional pairings, which do not form a nested structure. Deriving dynamic programming algorithms for the analysis and prediction of RNA secondary structures with pseudoknots is possible [241], however, the time complexities of these algorithms increases substantially. Although it is possible to present them in the planar layout, pseudoknots are instead classified as tertiary motifs of RNA, since their occurrence often contributes to forming a nonplanar spatial structure of RNA.

10.9 Prediction of RNA Structure by Comparative Sequence Analysis

Comparative sequence analysis involves aligning a target RNA sequence with a block of RNA sequences of known structure. Then, using the correspondences obtained we can infer the secondary and/or tertiary structure of the target RNA sequence. The idea, analogous to that used in comparative modeling of proteins, is that similar sequences of ribonucleotides lead to similar secondary and tertiary structures of molecules. This idea is related to the paradigm that homologous RNA species result from an evolutionary relationship and that functionally homologous regions will adopt similar structures.

In comparative analysis of RNA species, sequences are searched for compensatory base pair changes. If, in the course of evolution, a base pair has changed, then a compensatory mutation should have occurred on the complementary string, allowing the molecule to maintain its spatial structure. The existing software for alignment-based structure prediction of RNA [55, 206], enables a group of sequences to be aligned with a new target sequence, and using regions of high sequence conservation of the group as predictors of secondary-structure motifs in the target sequence. The growing number of sequences in RNA databases will result in the possibility of quickly adding new RNA sequences to structured databases of homologous RNA molecules.

10.10 Exercises

1. Assume that RNA chain has length N and the bases are numbered $1, 2, \ldots, N$. There are K bonds between bases, depicted as follows:

$$\begin{array}{ccc} i_1 & - & j_1 \\ i_2 & - & j_2 \\ \vdots & \vdots & \vdots \\ i_K & - & j_K. \end{array} \tag{10.16}$$

   How can one determine whether these bonds have a nested structure? Write a computer program for solving this problem.
2. Develop a computer program with graphics for drawing secondary structure of RNA of length N, given a list of nested bonds between bases, as in (10.16).
3. Develop a computer program for maximizing the number of pairings, on the basis of the algorithm described in Sect. 10.8.2.
4. Download a short tRNA sequence from the GtRNA database [181, 318]. Use the program from Exercise 3 to predict its secondary structure. Compare the structure obtained to that available in the GtRNA database.
5. Present the above RNA sequence to one of the RNA secondary-structure prediction servers [346].

6. Develop a computer program for minimizing the free energy of an RNA sequence using one of the algorithms described in Sect. 10.8.3. Use it to data from Exercise 4.