CHAPTER 5

Prediction of RNA Secondary Structure

THE PREVIOUS TWO CHAPTERS DISCUSS the alignment of protein and nucleic acid sequences. The methods used either align entire sequences or search for common patterns in the sequences. In either case, the objective is to locate a set of sequence characters in the same order in the sequences. Nucleic acid sequences that specify RNA molecules have to be compared differently. Sequence variations in RNA sequences maintain base-pairing patterns that give rise to double-stranded regions (secondary structure) in the molecule. Thus, alignments of two sequences that specify the same RNA molecules will show covariation at interacting base-pair positions, as illustrated in Figure 5.1. In addition to these covariable positions, sequences of RNA-specifying genes may also have rows of similar sequence characters that reflect the common ancestry of the genes.

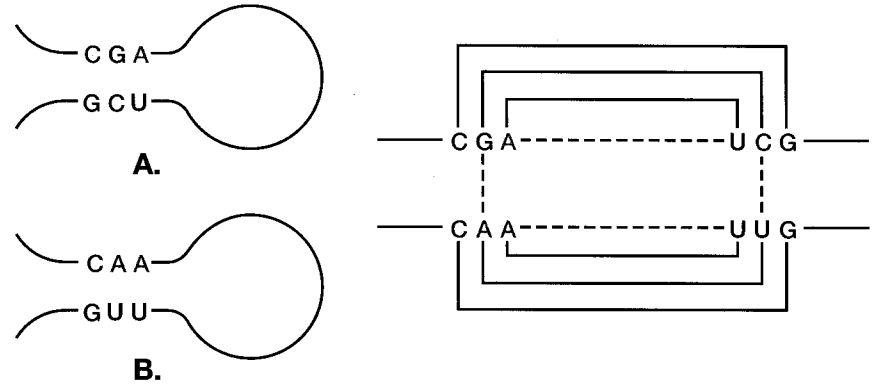


Figure 5.1. Complementary sequences in RNA molecules maintain RNA secondary structure. Shown is a simple stem-and-loop structure formed by the RNA strand folding back on itself. Molecule *A* depends on the presence of two complementary sequences CGA and UCG that are base-paired in the structure. In *B*, two sequence changes, G → A and C → U, which maintain the same structure, are present. Aligning RNA sequences required locating such regions of sequence covariation that are capable of maintaining base-pairing in the corresponding structure.

INTRODUCTION

As genomic sequences of organisms become available, it is important to be able to identify the various classes of genes, including the major class of genes that encodes RNA molecules. There are a large number of Web sites listed in Table 5.1 that provide programs

Table 5.1. *RNA databases and RNA analysis Web sites*

Site or resource	Web address	Reference
5S Ribosomal RNA data bank	http://rose.man.poznan.pl/5SData/ and mirrored at http://userpage.chemie.fu-berlin.de/fb_chemie/ibc/agerdmann/5S_rRNA.html	Szymanski et al. (1999)
5S rRNA database	http://www.bchs.uh.edu/~nzhou/temp/5snew.html	Shumyatsky and Reddy (1993)
Comparative RNA Web site	http://www.rna.icmb.utexas. edu/	see Web site
GenLang linguistic sequence analyzer	http://www.cbil.upenn.edu/	Dong and Searls (1994)
Gobase for mitochondrial sequences	http://alice.bch.umontreal.ca/genera/gobase/gobase.html	Korab-Laskowska et al. (1998)

Site or resource	Web address	Reference
Intron analysis—*Saccharomyces cerevisiae*	http://www.cse.ucsc.edu/research/compbio/yeast_introns.html	Spingola et al. (1999)
tRNA genes, higher plant mitochondria	ftp://ftp.ebi.ac.uk/pub/databases/plmitrna/	Ceci et al. (1999)
MFOLD minimum energy RNA configuration	http://bioinfo.math.rpi.edu/~zukerm/rna/	Zuker et al. (1991)
Nucleic acid database and structure resource	http://ndbserver.rutgers.edu/	Berman et al. (1998)
Pseudobase–pseudoknot database maintained by E. van Batenburg, Leiden University	http://wwwbio.leidenuniv.nl/~batenburg/pkb.html	see Web page
Ribonuclease P database Web site	http://jwbrown.mbio.ncsu.edu/RNaseP/home.html	Brown (1999)
Ribosomal RNA database project (RDP II)	http://www.cme.msu.edu/RDP/	Maidak et al. (1999)
Ribosomal RNA mutation databases	http://www.fandm.edu/Departments/Biology/Databases/RNA.html	Triman and Adams (1997)
RiboWeb Project–3D models of *E. coli* 30S ribosomal subunit and 16s rRNA	http://www-smi.stanford.edu/projects/helix/ribo3dmodels/index.html	Chen et al. (1997)
RNA aptamer sequence database (University of Texas)	http://speak.icmb.utexas.edu/ellington/aptamers.html	see Web site
RNA editing Web site, UCLA	http://www.lifesci.ucla.edu/RNA/index.html	Simpson et al. (1998)
RNA editing, uridine insertion/deletion	http://www.lifesci.ucla.edu/RNA/trypanosome/	Simpson et al. (1998)
RNA modification database	http://medlib.med.utah.edu/RNAmods/	Limbach et al. (1994); Rozenski et al. (1999)
RNA secondary structures, Group I introns, 16S rRNA, 23S rRNA	http://www.rna.icmb.utexas.edu	Gutell (1994); Schnare et al. (1996 and references therein)
RNA structure database	http://www.rnabase.org/	see Web page
RNA world at IMB Jena	http://www.imb-jena.de/RNA.html	Sühnel (1997)
rRNA–Database of ribosomal subunit sequences	http://rrna.uia.ac.be/	De Rijk et al. (1992, 1999)
Signal recognition particle database	http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html	Samuelsson and Zwieb (2000)
Small RNA database	http://mbcr.bcm.tmc.edu/smallRNA/smallrna.html	see Web page
snoRNA database for *S. cerevisiae*	http://rna.wustl.edu/snoRNAdb/	Lowe and Eddy (1999)
tmRNA[a] database	http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html	Wower and Zwieb (1999)
tmRNA[a] Web site	http://www.indiana.edu/~tmrna/	Williams (1999)
tRNAscan-SE search server	http://www.genetics.wustl.edu/eddy/tRNAscan-SE/	Lowe and Eddy (1997)
tRNA and tRNA gene sequences	http://www.uni-bayreuth.de/departments/biochemie/sprinzl/trna/	Sprinzl et al. (1998)
u RNA database	http://psyche.uthct.edu/dbs/uRNADB/uRNADB.html	Zwieb (1997)
Vienna RNA package for RNA secondary structure prediction and comparison	http://www.tbi.univie.ac.at/~ivo/RNA/	Hofacker et al. (1998); Wuchty et al. (1999)
Viroid and viroid-like RNA sequences	http://www.callisto.si.usherb.ca/~jpperra	Lafontaine et al. (1999)

[a] tmRNA adds a carboxy-terminal peptide tag to the incomplete protein product from a broken mRNA molecule and thereby targets the protein for proteolysis.

A list of RNA Web sites and databases is available at http://bioinfo.math.rpi.edu/~zukerm/ and at http://pundit.colorado.edu:8080/.

and guest sites for RNA analysis or for access to databases of RNA molecules and sequences. These molecules perform a variety of important biochemical functions, including translation; RNA splicing, processing, and editing; and cellular localization. As with proteins, RNA-specifying genes may be identified by using the unknown gene as a query sequence for DNA sequence similarity searches, as described in Chapter 7. If a significant match to the sequence of an RNA molecule of known structure and function is found, then the query molecule should have a similar role. For some small molecules, the amount of sequence variation necessitates the use of more complex search methods, described later in this chapter.

RNA STRUCTURE PREDICTION BASICS

A computational method for predicting the most likely regions of base-pairing in an RNA molecule has been designed, just given the sequence, thus providing an ab initio prediction of secondary structure. From the many possible choices of complementary sequences that can potentially base-pair, the compatible sets that provide the most energetically stable molecules are chosen. Structures with energies almost as stable as the most stable one may also be produced, and regions whose predictions are the most reliable can be identified from such an analysis. Sequence variations found in related sequences may also be used to predict which base pairs are likely to be found in each of the molecules. One variation of RNA structure prediction methods will predict a set of sequences that are able to form a particular structure. Methods for predicting three-dimensional structures from sequence are also being developed (see http://bioinfo.math.rpi.edu/~zuker/rna/).

Another type of RNA secondary structure prediction method takes into account conserved patterns of base-pairing that are conserved during evolution of a given class of RNA molecules. Sequence positions that base-pair are found to vary at the same time during evolution of RNA molecules so that structural integrity is maintained. For example, if two positions G and C form a base pair in a given type of molecule, then sequences that have C and G reversed, or A and U or U and A at the corresponding positions, would be considered reasonable matches. These patterns of covariation in RNA molecules are a manifestation of secondary structure that lead to a structural prediction. The computational challenge is to discover these covariable positions against the background of other sequence changes.

FEATURES OF RNA SECONDARY STRUCTURE

Like protein secondary structure, RNA secondary structure can be conveniently viewed as an intermediate step in the formation of a three-dimensional structure. RNA secondary structure is composed primarily of double-stranded RNA regions formed by folding the single-stranded molecule back on itself. To produce such double-stranded regions, a run of bases downstream in the RNA sequence must be complementary to another upstream run so that Watson–Crick base-pairing between the complementary nucleotides G/C and A/U (analogous to the G/C and A/T base pairs in DNA) can occur. In addition, however, G/U wobble pairs may be produced in these double-stranded regions. As in DNA, the G/C base pairs contribute the greatest energetic stability to the molecule, with A/U base pairs contributing less stability than G/C, and G/U wobble base pairs contributing the least. From the RNA structures that have been solved, these base pairs and a number of addi-

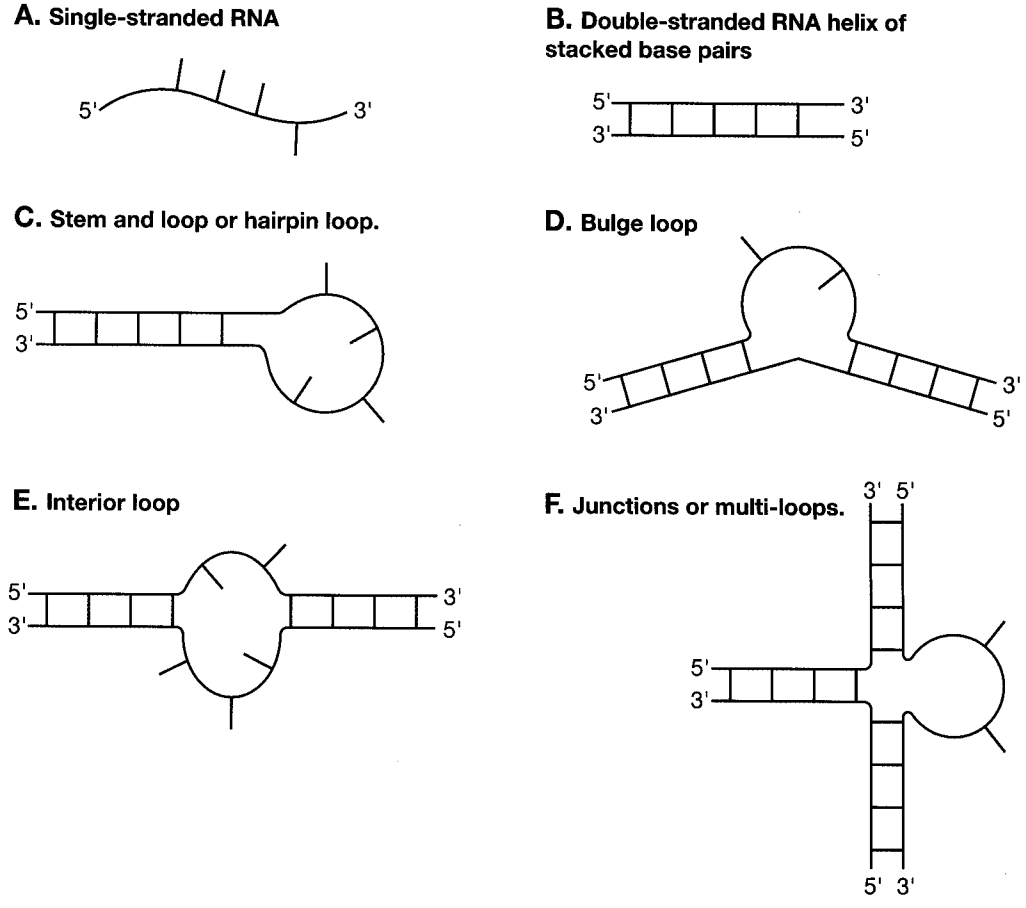


Figure 5.2. Types of single- and double-stranded regions in RNA secondary structures. Single-stranded RNA molecules fold back on themselves and produce double-stranded helices where complementary sequences are present. A particular base may either not be paired, as in *A*, or paired with another base, as in *B*. The double-stranded regions will most likely form where a series of bases in the sequence can pair with a complementary set elsewhere in the sequence. The stacking energy of the base pairs provides increased energetic stability. Combinations of double-stranded and single-stranded regions produce the types of structures shown in *C–F*, with the single-stranded regions destabilizing neighboring double-stranded regions. The loop of the stem and loop in *C* must generally be at least four bases long to avoid steric hindrance with base-pairing in the stem part of the structure. The stem and loop reverses the chemical direction of the RNA molecule. Interior loops, as in *D*, form when the bases in a double-stranded region cannot form base pairs, and may be asymmetric with a different number of base pairs on each side of the loop, as shown in *E*, or symmetric with the same number on each side. Junctions, as in *F*, may include two or more double-stranded regions converging to form a closed structure. The RNA backbone is red, and both unpaired and paired bases are blue. The types of loop structures can be represented mathematically, thereby aiding in the prediction of secondary structure (Sankoff et al. 1983; Zuker and Sankoff 1984). (Adapted from Burkhard et al. 1999b.)

tional ones (see Burkhard et al. 1999a,b) have been identified. RNA structure predictions comprise base-paired and non-base-paired regions in various types of loop and junction arrangements, as shown in Figure 5.2.

In addition to secondary structural interactions in RNA, there are also tertiary interactions, illustrated by the examples in Figure 5.3. These kinds of structures are not predictable by secondary structure prediction programs. They can be found by careful covariance analysis.

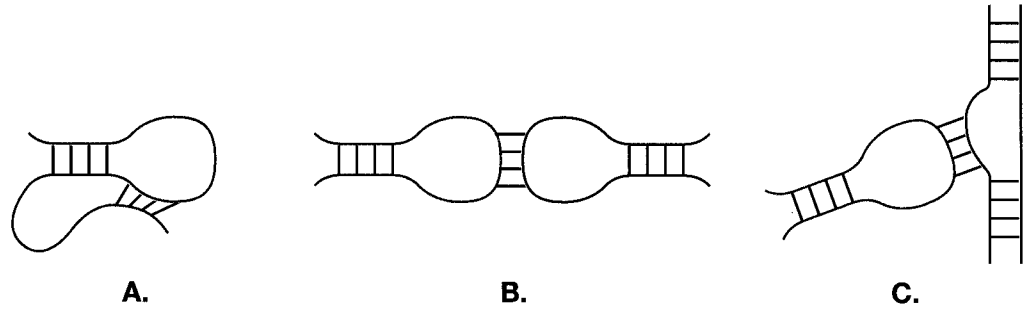


Figure 5.3. Examples of known interactions of RNA secondary structural elements. (*A*) Pseudoknot. (*B*) Kissing hairpins. (*C*) Hairpin-bulge contact. (Adapted from Burkhard et al. 1999b.)

LIMITATIONS OF PREDICTION

In predicting RNA secondary structure, some simplifying assumptions are usually made. First, the most likely structure is similar to the energetically most stable structure. Second, the energy associated with any position in the structure is only influenced by local sequence and structure. Thus, the energy associated with a particular base pair in a double-stranded region is assumed to be influenced only by the previous base pair and not by the base pairs farther down the double-stranded region or anywhere else in the structure. These energies can be reliably estimated by experimentation with small, synthetic RNA oligonucleotides (Tinoco et al. 1971, 1973; Freier et al. 1986; Turner and Sugimoto 1988; SantaLucia 1998) recently improved to include sequence dependence (Mathews et al. 1999). They are most reliable when used for standard Watson–Crick base pairs and single G-U pairs surrounded

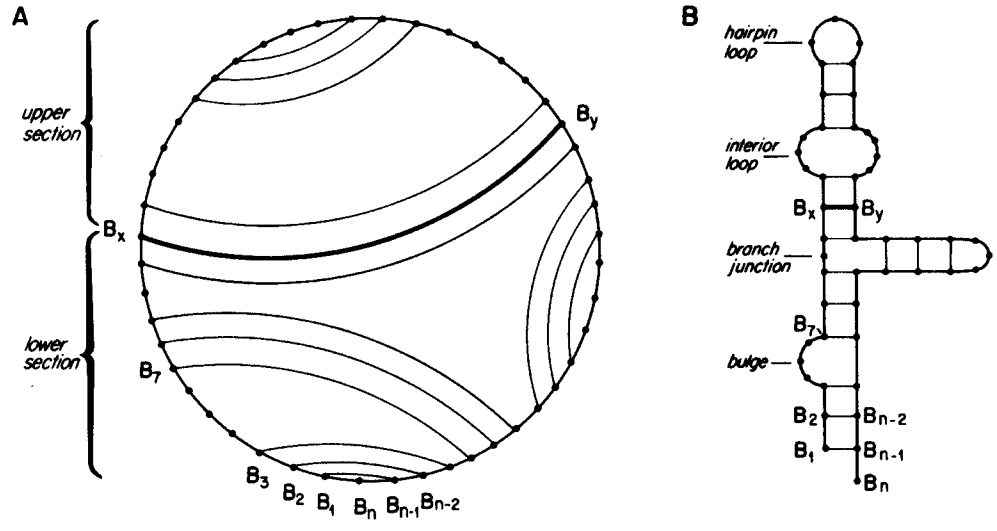


Figure 5.4. Display of base pairs in an RNA secondary structure by a circle plot. The predicted minimum free-energy structure shown in *B* is represented by a plot of the predicted base pairs as arcs connecting the bases in the sequence, which is drawn around the circumference of a circle, as shown in *A* (see Nussinov and Jacobson 1980). Note that none of the lines cross, a representation that the structure does not include any knots. (Reprinted from Nussinov and Jacobson 1980.)

by Watson–Crick pairs. Finally, the structure is assumed to be formed by folding of the chain back on itself in a manner that does not produce any knots. The best way of representing this requirement is to draw the sequence in a circular form. The paired bases are then joined by arcs. If the total structure with all predicted base pairs is to be free of knots, none of the arcs must cross (Fig. 5.4). Note, however, that if a pseudoknot (Fig. 5.3) is represented on such a diagram, the lines will cross.

DEVELOPMENT OF RNA PREDICTION METHODS

The development of methods for predicting RNA secondary structure has been reviewed by von Heijne (1987). Tinoco et al. (1971) first estimated the energy associated with regions of secondary structure by extrapolation from studies with small molecules and then attempted to predict which configurations of larger molecules were the most energetically stable. Energy estimates included the stabilizing energy associated with stacking base pairs in a double-stranded region and the destabilizing influence of regions that were not paired. Pipas and McMahon (1975) developed computer programs that listed all possible helical regions in tRNA sequences; using modified Watson–Crick base-pairing rules, they created all possible secondary structures by forming permutations of compatible helical regions, and evaluated each possible structure for total free energy. Studnicka et al. (1978) designed a method for adding compatible double-stranded regions together to produce the energetically most favorable structure. Martinez (1984) made a list of possible double-stranded regions, and these regions were then given weights in proportion to their equilibrium constants, calculated by the Boltzmann function [$\exp(-\Delta G/RT)$], where $-\Delta G$ is the free energy of the regions, R is the gas constant, and T is the temperature. The RNA molecule is folded by a Monte Carlo method in which one initial region is chosen at random from a weighted pool, similar to the method used in Gibbs sampling (see p. 177).

In the Monte Carlo method, a random drawing is made from a pool of all possible double-stranded regions, with the number of each type weighted in proportion to energetic stability.

Imagine each possible double-stranded region being represented by a marble in a bag. The number of each type of marble is weighted by the Boltzmann probability so that marbles corresponding to more energetically stable regions are more likely to be chosen. Additional compatible regions are then added sequentially by further selections from the weighted pool until no more can be added. This method generates a set of possible structures weighted by energy, but it does not take into account the destabilizing effect of unpaired regions. The Boltzmann probability function is used in more recent applications (described below) to find the most probable secondary structures (Hofacker et al. 1998; Wuchty et al. 1999).

Nussinov and Jacobson (1980) were the first to design a precise and efficient algorithm for predicting secondary structure. The algorithm generates two scoring matrices—one $M(i,j)$ to keep track of the maximum number of base pairs that can be formed in any interval i to j in the sequence and a second $K(i,j)$ to keep track of the base position k that is paired with j. From these matrices, a structure with the maximum possible number of base pairs could be deduced by a trace-back procedure similar to that used in performing sequence alignments by dynamic programming. Zuker and Stiegler (1981) used the dynamic programming algorithm and energy rules for producing the most energetically favorable structure. Their method assumes that the most energetic, and usually longest, predicted dsRNA regions are present in the molecule. Because many double-stranded regions are predictable for most RNA sequences, the number of predictions is reduced by including known biochemical or structural information to indicate which bases should be paired or not paired, by enforcing topological restraints and by requiring that the structure be in an energetically stable configuration.

MFOLD, written by Dr. Michael Zuker and colleagues, is commonly used to predict the energetically most stable structures of an RNA molecule (Jaeger et al. 1989, 1990; Zuker 1989, 1994). MFOLD provides a set of possible structures within a given energy range and provides an indication of their reliability. The program also uses covariance information from phylogenetically related sequences (Zuker et al. 1991). MFOLD includes methods for graphic display of the predicted molecules. This program is one of the most demanding on computer resources that is currently used because the algorithm is of N^3 complexity, where N is the sequence length. For each doubling of sequence length, the time taken to compute a structure increases eightfold. The program also requires a large amount of memory for storing intermediate calculations of structure energies in multiple scoring matrices. As a result, MFOLD is most often used to predict the structure of sequences less than 1000 nucleotides in length. This method is most reliable for small molecules and becomes less reliable as the length of the molecule increases.

MFOLD and many other types of useful information on RNA are found at the Web site of Dr. Michael Zuker, at http://bioinfo.math.rpi.edu/~zuker/rna/. Details of running MFOLD are not given here because the user manual for MFOLD is widely available (Jaeger et al. 1990). Recently, a new method called the partition function method for finding the most probable secondary structural configuration of an RNA molecule and the most probable base pairs has been reported by the Vienna RNA group (Wuchty et al. 1999) and is discussed below (p. 219).

One advance in the prediction of RNA structure has come from the recognition that certain RNA sequences form specific structures and that the presence of these sequences is strongly predictive of such a structure. For example, the hairpin CUUCGG occurs in different genetic contexts and forms a very stable structure (Tuerk et al. 1988). Databases of such RNA structures and RNA sequences can greatly assist in RNA structure prediction (Table 5.1).

The genetic algorithm (see Chapter 4, p. 157) has also been used to predict secondary structure (Shapiro and Navetta 1994); for aligning RNA sequences, taking into account both sequence and secondary structure and including pseudoknots (Notredame et al. 1997); and for simulation of RNA-folding pathways (Gultyaev et al. 1995). The program FOLDALIGN uses a dynamic programming algorithm to align RNAs based on sequence and secondary structure and locates the most significant motifs (Gorodkin et al. 1997). Chan et al. (1991) have described another algorithm for the same purpose, and Chetouani et al. (1997) have developed ESSA, a method for viewing and analyzing RNA secondary structure.

METHODS

SELF-COMPLEMENTARY REGIONS IN RNA SEQUENCES PREDICT SECONDARY STRUCTURE

One of the simplest types of analyses that can be performed to find stretches of sequence in RNA that are self-complementary is a dot matrix sequence comparison for self-complementary regions. For single-stranded RNA molecules, these repeats represent regions that can potentially self-hybridize to form RNA double strands (von Heijne 1987; Rice et al. 1991). All types of RNA secondary structure analysis begin by the identification of these regions, and, once identified, the compatible regions may be used to predict a minimum free-energy structure. A more advanced type of dot matrix can be used to show the most energetic parts of the molecule (see Fig. 5.8, below).

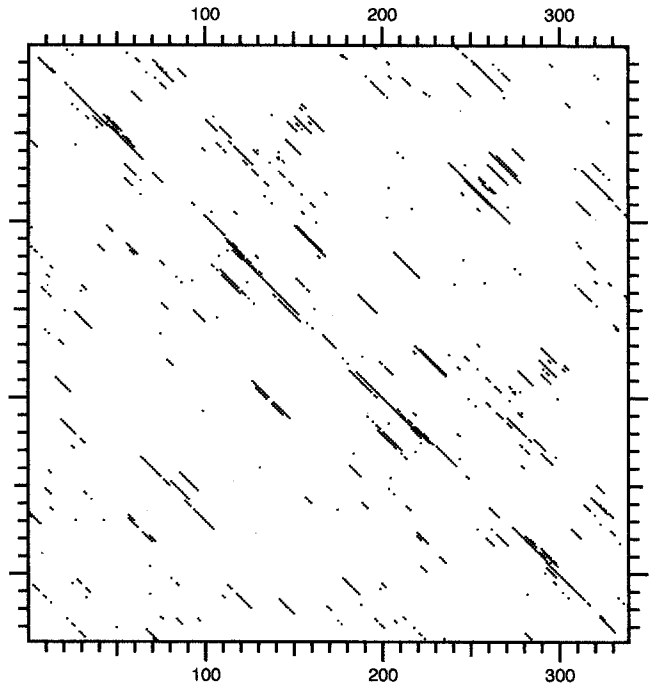


Figure 5.5. Dot matrix analysis of the potato tuber spindle viroid for RNA secondary structure using the MATRIX function of DNA Strider v. 1.2 on a Macintosh computer.

Self-complementary regions in RNA may be found by performing a dot matrix analysis with the sequence to be analyzed listed in both the horizontal and vertical axes. In one method for finding such regions, the sequence is listed in the 5′→3′ direction across the top of the page and the sequence of the complementary strand is listed down the side of the page, also in the 5′→3′ direction. The matrix is then scored for identities. Self-complementary regions appear as rows of dots going from upper left to lower right. For RNA, these regions represent sequences that can potentially form A/U and G/C base pairs. G/U base pairs will not usually be included in this simple type of analysis. As with matching DNA sequences, there are many random matches between the four bases in RNA, and the diagonals are difficult to visualize. A long window and a requirement for a large number of matches within this window are used to filter out these random matches.

An example of the RNA secondary structure analysis using a DNA matrix option of DNA Strider is shown in Figure 5.5. An analysis of the potato spindle tuber viroid is shown, using a window of 15 and a required match of 11. Note the appearance of a diagonal running from the center of the matrix to the upper left, and a mirror image of this diagonal running to the lower right. The presence of this diagonal indicates the occurrence of a large self-complementary sequence such that the entire molecule can potentially fold into a hairpin structure. An alternative dot matrix method for finding RNA secondary structure is to list the given RNA sequence across the top of the page and also down the side of the page and then to score matches of complementary bases (G/C, A/U, and G /U). Diagonals indicating complementary regions will go from upper right to lower left in this type of matrix. This is the kind of matrix used to produce an energy matrix (see Fig. 5.8, below).

MINIMUM FREE-ENERGY METHOD FOR RNA SECONDARY STRUCTURE PREDICTION

To predict RNA secondary structure, every base is first compared to every other base by a type of analysis very similar to the dot matrix analysis. The sequence is listed across the top and down the side of the page, and G/C, A/U, and G/U base pairs are scored (for an example using a dot matrix method to find hairpins, see Fig. 5.5). Just as a diagonal in a two-sequence comparison indicates a range of sequence similarity, a row of matches in the RNA matrix indicates a succession of complementary nucleotides that can potentially form a double-stranded region. The energy of each predicted structure is estimated by the nearest-neighbor rule by summing the negative base-stacking energies for each pair of bases in double-stranded regions and by adding the estimated positive energies of destabilizing regions such as loops at the end of hairpins, bulges within hairpins, internal bulges, and other unpaired regions. Representative examples of the energy values that are currently used are given in Table 5.2. To evaluate all the different possible configurations and to find the most energetically favorable, several types of scoring matrices are used. The complementary regions are evaluated by a dynamic programming algorithm to predict the most energetically stable molecule. The method is similar to the dynamic programming method used for sequence alignment (see Chapter 3).

To calculate the stacking energy of a row of base pairs in the molecule, the stacking energies similar to those shown in Table 5.2 are used. An illustrative example for evaluation of energy in a double-stranded region is shown in Figure 5.6. The sequence is listed down the side of the matrix, and a portion of the same sequence is also listed across the top of the matrix; matching base pairs have been identified within the matrix. The object is to find a diagonal row of matches that goes from upper right to lower left, and such a row is shown in the example. In Figure 5.6, a match of four complementary bases in a row produces a molecule of free energy −6.4 kcal/mole. In general, each matrix value is obtained by considering the minimum energy values obtained by all previous complementary pairs

Table 5.2. *Predicted free-energy values (kcal/mole at 37°C) for base pairs and other features of predicted RNA secondary structures*

A. Stacking energies for base pairs

	A/U	C/G	G/C	U/A	G/U	U/G
A/U	−0.9	−1.8	−2.3	−1.1	−1.1	−0.8
C/G	−1.7	−2.9	−3.4	−2.3	−2.1	−1.4
G/C	−2.1	−2.0	−2.9	−1.8	−1.9	−1.2
U/A	−0.9	−1.7	−2.1	−0.9	−1.0	−0.5
G/U	−0.5	−1.2	−1.4	−0.8	−0.4	−0.2
U/G	−1.0	−1.9	−2.1	−1.1	−1.5	−0.4

B. Destabilizing energies for loops

Number of bases	1	5	10	20	30
Internal	–	5.3	6.6	7.0	7.4
Bulge	3.9	4.8	5.5	6.3	6.7
Hairpin	–	4.4	5.3	6.1	6.5

(*Upper*) Stacking energy in double-stranded region when base pair listed in left column is followed by base pair listed in top row. C/G followed by U/A is therefore the dinucleotide 5′ CU 3′ paired to 5′ AG 3′. (*Lower*) Destabilizing energies associated with loops. Hairpin loops occur at the end of a double-stranded region, internal loops are unpaired regions flanked by paired regions, and a bulge loop is a bulge of one strand in an otherwise paired region (Fig. 5.2). An updated and more detailed list of energy parameters may be found at the Web site of M. Zuker (http://bioinfo.math.rpi.edu/~zuker/rna/energy/).

From Turner and Sugimoto (1988); Serra and Turner (1995).

A. Base comparisons

5'	A	C	G	U	3'
A					
C					
G					
U					
–					
–					
G		C/G		U/G	
C			G/C		
G		C/G		U/G	
U	A/U	C/U	G/U		
3'					

B. Free energy calculations

5'	A	C	G	U	3'
A					
C					
G					
U					
–					
–					
G				−6.4	
C			−5.2		
G		−1.8			
U					
3'					

Figure 5.6. Evaluation of secondary structure in RNA sequence by the method described in the text. The sequence is listed down the first column of *A* and *B* in the 5′→3′ orientation, and the first four bases of the sequence are also listed in the first row of the tables in the 5′→3′ direction. Several complementary base pairs between the first and last four bases that could lead to secondary structure are shown in *A*. The most 5′ base is listed first in each pair. The diagonal set of base pairs A/U, C/G, G/C, and U/G reveals the presence of a potential double-stranded region between the first and last four bases. The free energy associated with such a row of base pairs is shown in *B*. A C/G base pair following an A/U base pair has a base stacking energy of −1.8 kcal/mole (Turner and Sugimoto 1988). This value is placed in the corresponding position in *B*. Similarly, a C/G base pair followed by a G/C provides energy of −3.4, and a G/C followed by a U/G, −1.2 kcal/mole. Hence, the energy accumulated after stacking of these additional two base pairs is −5.2 and −6.4. The energy of this double-stranded structure will continue to decrease (become more stable) as more base pairs are added, but will be increased if the structure is interrupted by noncomplementary base pairs.

decreased by the stacking energy of any additional complementary base pairs or increased by the destabilizing energy associated with noncomplementary bases. The increase depends on the type and length of loop that is introduced by the noncomplementary base pair, whether internal loop, bulge loop, or hairpin loop, as shown in Table 5.2. This comparison of all possible matches and energy values is continued until all nucleotides have been compared. The pattern followed in comparing bases within the RNA molecule is illustrated in Figure 5.7.

SUBOPTIMAL STRUCTURE PREDICTIONS BY MFOLD AND THE USE OF ENERGY PLOTS

Originally, the FOLD program of M. Zuker predicted only one structure having the minimum free energy. However, changes in a single nucleotide can result in drastic changes in the predicted structure. A later version, called MFOLD, has improved prediction of non-base-paired interactions and predicts several structures having energies close to the minimum free energy. These predictions accurately reflect structures of related RNA molecules derived from comparative sequence analysis (Jaeger et al. 1989; Zuker 1989, 1994; Zuker et al. 1991; Zuker and Jacobson 1995). To find these suboptimal structures, the dynamic programming method was modified (Zuker 1989, 1991) to evaluate parts of a new scoring matrix in which the

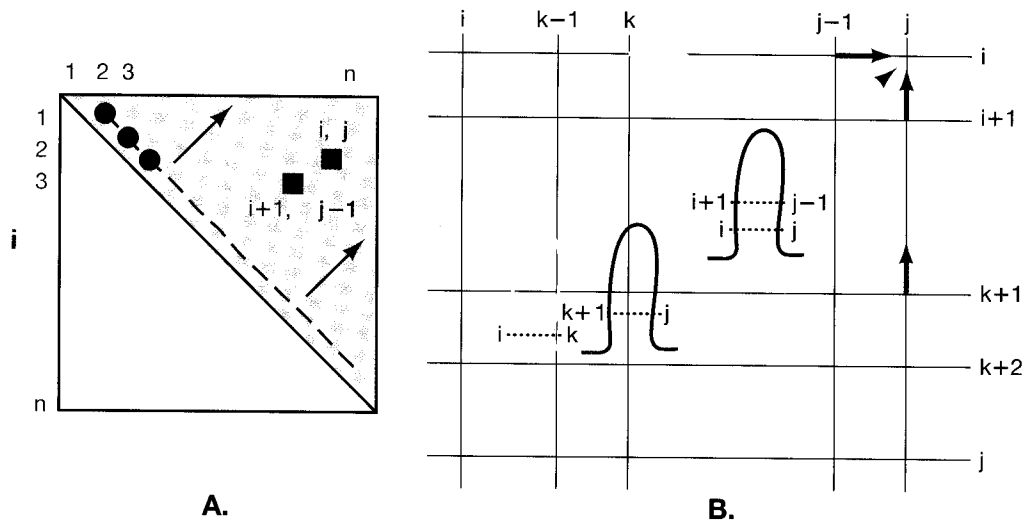


Figure 5.7. Method used in dynamic programming analysis for identifying the most energetically favorable configuration of a linear RNA molecule. (*A*) The sequence of an RNA molecule of length *n* bases is listed across the top of the page and down the side. The index of the sequence across the top is *j* and that down the side is *i*. The search only includes the upper right part of the matrix shown in gray and begins at the first diagonal line for matching base pairs. First positions $i = 1$ and $j = 2$ are compared for potential base-pairing, and if pairing can occur, an energy value is placed in an energy matrix *W* at position 1,2. Then, $i = 2$ and $j = 3$ base are compared, and so on, until all base combinations along the dashed diagonal have been made. Then, comparisons are made along the next upper right diagonal. As each pair of bases is compared, an energy calculation is made that is the optimal one up to that point in the comparison. In the simplest case, if $i+1$ pairs with $j-1$, and *i* pairs with *j*, and if this structure is the most favorable up to that point, the energy of the *i/j* base pair will be added to that of the $i+1/j-1$ base pair. Other cases are illustrated in *B*. The process of obtaining the most stable energy value at each matrix position is repeated following the direction of the arrows until the last position, $i = 1$ and $j = n$, has been compared and the energy value placed at this position in matrix *W*, the value entered in *W(1,n)*, will be the energy of the most energetically stable structure. The structure is then found by a trace-back procedure through the matrices similar to that used for sequence alignments. The method used is a combination of a search for all possible double-stranded regions and an energy calculation based on energy values similar to those in Table 5.2. The search for the most energetic structure uses an algorithm (Zuker and Stiegler 1981) similar to that for finding the structure with maximum base-pairing (Nussinov and Jacobson 1980). These authors recognized that there are three possible ways, illustrated here by the colored arrows, of choosing the best energy value at position *i,j* in an energy matrix *W*. The simplest calculation (*red arrow*) is to use the energy value found up to position $i-1, j-1$ diagonally below *i,j*. If *i* and *j* can form a base pair (and if there are at least four bases between them in order to allow enough sequence for a hairpin) and $i+1$ and $j-1$ also pair, then the stacking energy of *i/j* upon $i+1/j-1$ will reduce the energy value at $i+1, j-1$, producing a more stable structure, and the new value can be considered a candidate for the energy value entered at position *i,j*. If *i* and *j* do not pair, then another choice for the energy at *i,j* is to use the values at positions *i*, $j-1$ or $i+1$, *j* illustrated by the blue arrows. *i* and *j* then become parts of loop structures. Finally, *i* and *j* may each be paired with two other bases, *i* with *k* and *j* with $k+1$, where *k* is between *i* and *j* ($i < k < j$), illustrated by the structure shown in yellow and green, reflecting the location of the paired bases. The minimum free-energy value for all values of *k* must be considered to locate the best choice as a candidate value at *i,j*. Finally, of the three possible choices for the minimum free-energy value at *i,j* indicated by the four colored arrows, the best energy value is placed at position *W(i,j)*. The procedure is repeated for all values of *i* and *j*, as illustrated in *A*. Besides the main energy scoring matrix *W*, additional scoring matrices are used to keep track of auxiliary information such as the best energy up to *i,j* where *i* and *j* form a pair, and the influence of bulge loops, interior loops, and other destabilizing energies. An essential second matrix is *V(i,j)*, which keeps track of all substructures in the interval *i,j* in which *i* forms a base pair with *j*. Some values in the *W* matrix are derived from values in the *V* matrix and vice versa (Zuker and Stiegler 1981).

sequence is represented in two tandem copies on both the vertical and horizontal axes. The regions from $i = 1$ to n and $j = 1$ to n are used to calculate an energy $V(i,j)$ for the best structure that includes an i,j base pair and is called the included region. A second region, the excluded region, is used to calculate the energy of the best structure that includes i,j but is not derived from the structure at $i+1, j-1$ (Fig. 5.7). After certain corrections are made, the difference between the included and excluded values is the most energetic structure that includes the base pair i,j. All complementary base pairs can be sampled in this fashion to determine which are present in a suboptimal structure that is within a certain range of the optimal one.

An energy dot plot is produced showing the locations of alternative base pairs that produce the most stable or suboptimally stable structures, as illustrated in Figure 5.8. The program may be instructed to find structures within a certain percentage of the minimum free energy. Parameter d provides a measure of similarity between two structures. When MFOLD is established on a suitable local host machine, the window is interactive, and clicking a part of the display will lead to program output of the corresponding structure. The dot plot may be filtered so that only suboptimal regions with helices of a certain minimal length are shown. One of the predicted structures is shown in Figure 5.9.

Reliability of Secondary Structure Prediction

Three scores, *P*num (i), *H*num (i,j), and Ssum, have been derived to assist with a determination of the reliability of a secondary structure prediction for a particular base i or a base pair i,j. *P*num(i) is the total number of energy dots regardless of color in the ith row and ith column of the energy dot plot, and represents in an unfiltered dot plot the number of base pairs that the ith base can form with all other base pairs in structures within the defined energy range. The lower this value, the more well defined or "well determined" the local structure because there are few competitive foldings. *H*num(i,j) is the sum of *P*num(i) and *P*num(j) less 1 and is the total number of dots in the ith row and jth column and represents the total number of base pairs with the ith or jth base in the predicted structures. The *H*num for a double-stranded region is the average *H*num value for the base pairs in that helix. The lower this number, the more well determined the double-stranded region. In an analysis of tRNAs, 5S RNAs, ribosomal RNAs, and other published secondary structure models based on sequence variation (Jaeger et al. 1990; Zuker and Jacobson 1995), these methods correctly predict about 70% of the double-stranded regions. *S*num, also called ss-count, is the number of foldings in which base i is single-stranded divided by m, the number of foldings, and gives the probability that base i is single-stranded. If *S*num is approximately 1, then base i is probably in a single-stranded region, and if *S*num is approximately 0, then base i is probably not in such a region. This reliability information has been used to annotate output files of MFOLD and other RNA display programs (Zuker and Jacobsen 1998). Plots of these values against sequence position are given by the MFOLD program and the Zuker Web site.

OTHER ALGORITHMS FOR SUBOPTIMAL FOLDING OF RNA MOLECULES

A limitation of the Zuker method and other methods (Nakaya et al. 1995) for computing suboptimal RNA structures is that they do not compute all the structures within a given energy range of the minimum free-energy structure. For example, no alternative structures

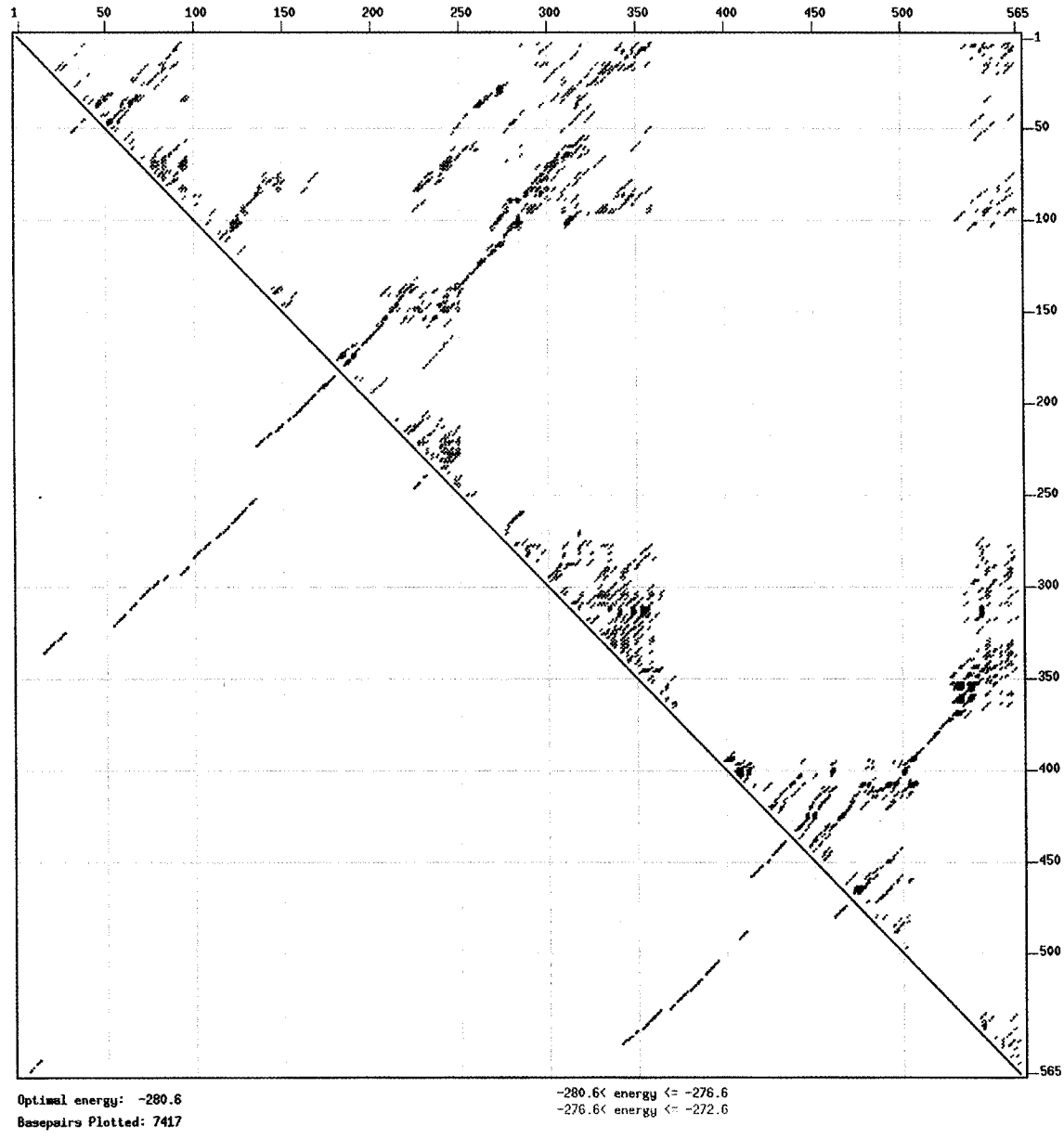


Figure 5.8. The energy dot plot (boxplot) of alternative choices of base pairs of an RNA molecule (Jacobson and Zuker 1993). The sequence is that of a human adenovirus pre-terminal protein (GenBank U52533) that is given by M. Zuker as an example on his Web site at http://bioinfo.math.rpi.edu/~zukerm. Foldings were computed using the default parameters of the MFOLD program at http://bioinfo.math.edu/~mfold/rna/form1.cgi (Mathews et al. 1999) using the thermodynamic values of SantaLucia (1998). The minimum energy of the molecule is −280.6 kcal/mole and the maximum energy increment is 12 kcal/mole. Black dots indicate base pairs in the minimum free-energy structure and are shown both above and the mirror image below the main diagonal. Red, blue, and yellow dots are base pairs in foldings of increasing 4, 8, and 12 kcal/mole energies greater than the minimum energy, respectively. A region with very few alternative base pairs such as the pairing of 370–395 with 530–505 is considered to be strongly predictive, whereas regions with many alternative base pairs such as the base-pairing in the region of 340–370 with 570–530 are much less predictive.

are produced that have the absence of base pairs in the best structure, and, if two substructures are joined by a stretch of unpaired bases, no structures are produced that are suboptimal for both structures. These factors limit the number of alternative structures predicted compared to known variations based on sequence variations in tRNAs (Wuchty et al. 1999).

These limitations have been largely overcome by using an algorithm originally described by Waterman and Byers (1985) for finding sequence alignments within a certain range of the optimal one by modifications of the trace-back procedure used in dynamic programming. This method efficiently calculates a large number of alternative structures, up to a very large number, within a given energy range of the minimum free-energy structure (see Fig. 5.10). The method has been used to demonstrate that natural tRNA sequences can form many alternative structures which are close to the minimum free-energy structure and that base modification plays a major role in this energetic stability (Wuchty et al. 1999). The method may also be used to assess the thermodynamic stability of RNA structures given expected changes in energies associated with base pairs and loops as a function of temperature. The RNA secondary structure prediction and comparison Web site at http://www.tbi.univie.ac.at/~ivo/RNA/ will fold molecules of length > 300 bases, and the Vienna RNA Package software for folding larger molecules on a local machine is available from this site.

PREDICTION OF MOST PROBABLE RNA SECONDARY STRUCTURE

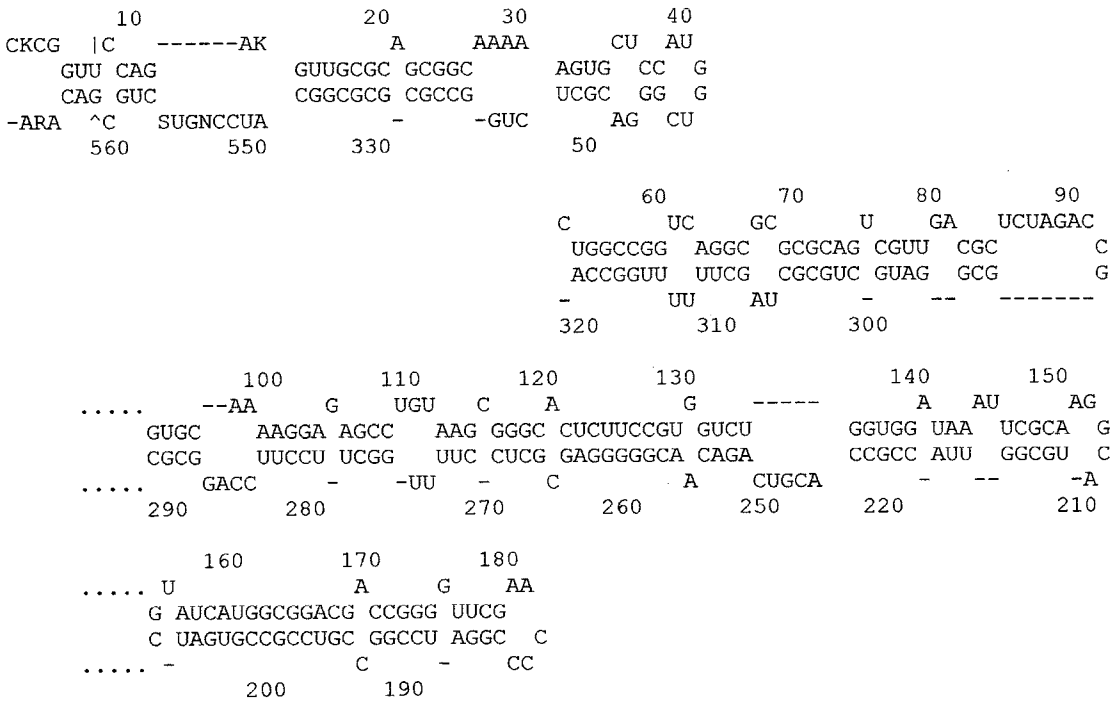
In the above types of analyses, the energy associated with predicted double-stranded regions in RNA is used to produce a secondary structure. Stabilizing energies associated with base-paired regions and destabilizing energies associated with loops are summed to produce the most stable structure or suboptimal RNA secondary structure. A different way of predicting the structures is to consider the probability that each base-paired region will form based on principles of thermodynamics and statistical mechanics. The probability of forming a region with free energy ΔG is expressed by the Boltzmann distribution, which states that the likelihood of finding a structure with free energy $-\Delta G$ is proportional to [exp $(-\Delta G/kT)$] where k is the Boltzmann gas constant and T is the absolute temperature.

The Boltzmann constant k is 8.314510 J/mole/degree K.

Note that the more stable a structure, the lower the value of ΔG. Since ΔG is a negative number, the value of $\exp(-\Delta G/kT)$ increases for more stable structures and also grows exponentially with a decrease in energy. The probability of these regions forming increases in the same manner. Conversely, the effect of destabilizing loops that have a positive ΔG is to decrease the probability of formation. By using these probability calculations and a dynamic programming method similar to that used in MFOLD, it is possible to predict the most probable RNA secondary structures and to assess the probability of the base pairs that contribute energetic stability to this structure.

For a set of possible structural states, the likelihood of each may be calculated using this formula, and the sum of these likelihoods provides a partition function that can be used to normalize each individual likelihood, providing a probability that each will occur. Thus, probability of structure A of energy $-\Delta G_a$ is [exp $(-\Delta G_a/kT)$] divided by the partition function Q, where $Q = \Sigma_s$ [exp $(-\Delta G_s/kT)$], the sum of probabilities of all possible struc-

A.



B.

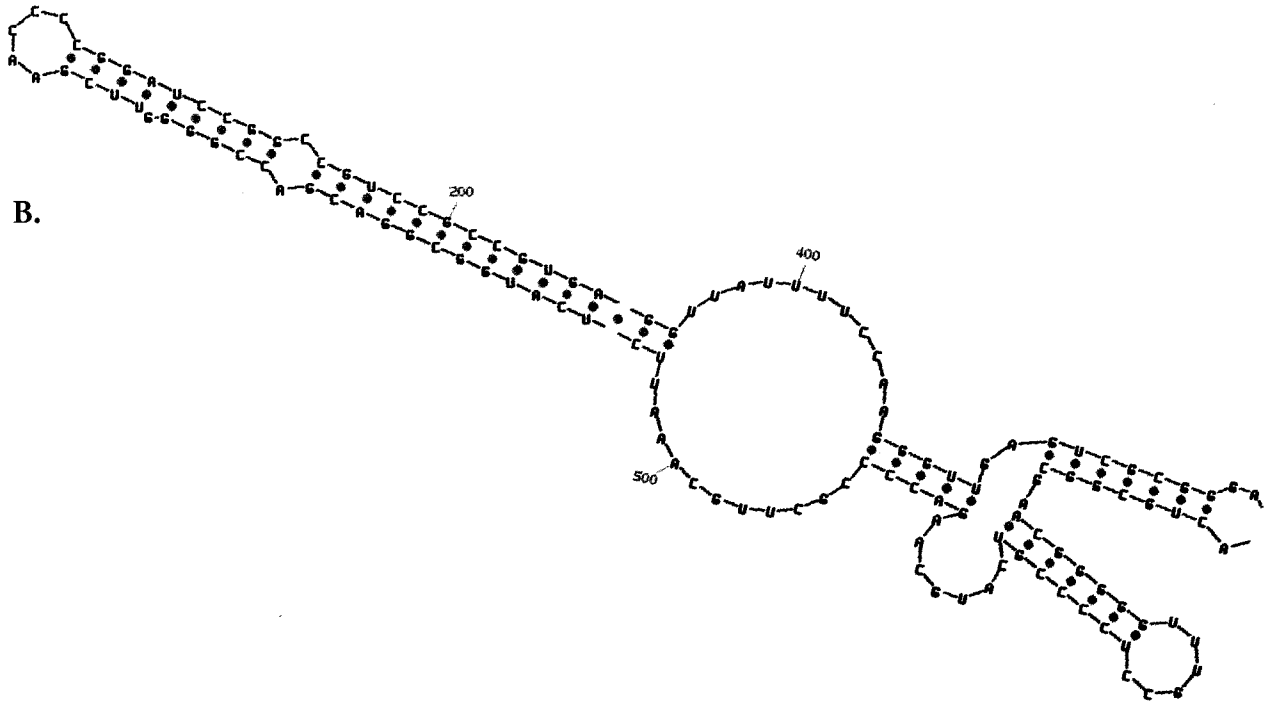


Figure 5.9. Model of RNA secondary structure of the human adenovirus pre-terminal protein. This model is one of several alternative structures represented by the above energy plot and provided as an output by the current versions of MFOLD. (*A*) Simple text representation of one of the predicted structures. Each stem-and-loop structure is shown separately and the left end of each structure is placed below the point of connection to the one above. (*B*) More detailed rendition of one part of the predicted structures. The structure continues beyond the right side of the page.

tures, s. This kind of analysis allows one to calculate the probability of a certain base pair forming.

The key to this analysis is the calculation of the partition function Q. A dynamic programming algorithm for calculating this function exactly for RNA secondary structure has been developed (McCaskill 1990). The algorithm is very similar to that used for computing an optimal folding by MFOLD. Complexity similarly increases as the cube of the sequence length, and the energy values used for base pairs and loops are also the same except that structures with very large interior loops are ignored. Just as the minimum free-energy value is given at $W(1,n)$ in the Zuker MFOLD algorithm, the value of the partition function is given at matrix position $Q(1,n)$ in the corresponding partition matrix.

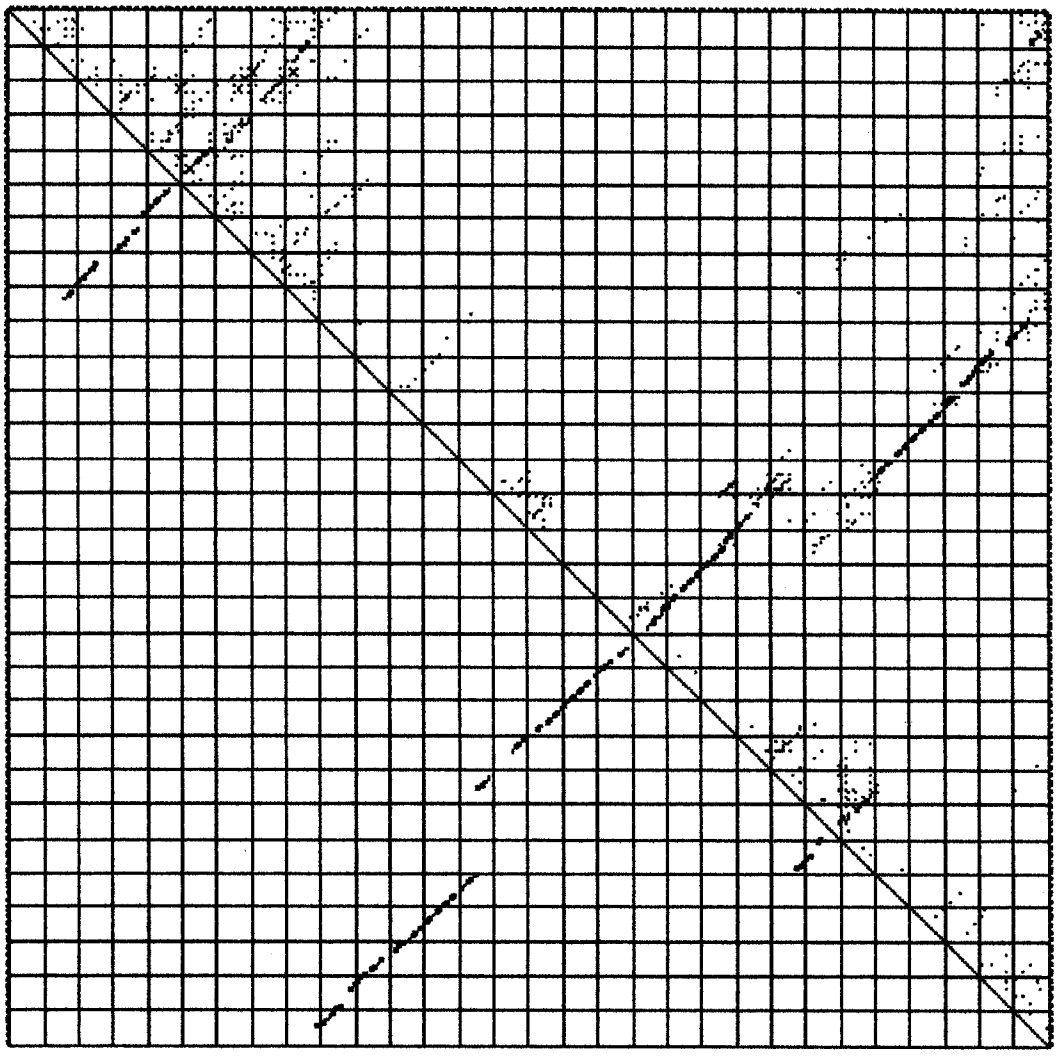
As indicated above, the partition function is calculated as the sum of the probabilities of each possible secondary structure. Because there are a very large possible number of structures, the calculation is simplified by calculating an auxiliary function, $Q^b(i,j)$, which is the sum of the probabilities of all structures that include the base pair *i,j*. The partition function $Q(i,j)$ includes both these structures and the additional ones where *i* is not paired with *j*. An example illustrating the difference between the minimum free energy and the partition function methods should be instructive. Suppose that the bases at positions $i+1, j-1$ and *i,j* can both form base pairs. They then form a stack of two base pairs. In the minimum free-energy method, the energy of the *i,j* pair stacked on the $i+1, j-1$ pair will be added to $V(i+1, j-1)$ to give $V(i,j)$, where *V* is a scoring matrix that keeps track of the best structure that includes an *i,j* base pair. In contrast, the value for $Q^b(i,j)$ will be calculated by multiplying the matrix value $Q^b(i+1, j-1)$ by the probability of the base pair *i,j* given by the Boltzmann probability [exp $(-\Delta G/kT)$], where ΔG is the negative stacking energy of the *i,j* base pair on the $i+1, j-1$ base pair, and will be a large number reflecting the probability given the stability of the base-paired region.

For a hairpin structure with a row of successive base pairs, the probability will be the product of the Boltzmann factors associated with the stacked pair, giving a high number for the relative likelihood of formation. The procedure followed by the partition function algorithm is to calculate $Q^b(i,j)$ and $Q(i,j)$ iteratively in a scoring matrix similar to that illustrated in Figure 5.7A until $Q(1,n)$ is reached. This matrix position contains the value of the full partition function Q.

Both the partition function and the probabilities of all base pairs are computed by this algorithm, and the most probable structural model is thereby found. Information about intermediate structures, base-pair opening and slippage, and the temperature dependence of the partition function may also be determined. The latter calculation provides information about the melting behavior of the secondary structure.

A suite of RNA-folding programs available from the Vienna RNA secondary structure prediction Web site (http://www.tbi.univie.ac.at/~ivo/RNA/) uses this methodology to predict the most probable and alternative RNA secondary structures. An example of the folding of a 300-base RNA molecule is given in Figure 5.10. The probability of forming each base pair is shown in a dot matrix display in which the dots are squares of increasing size reflecting the probability of the base pair formed by the bases in the horizontal and vertical positions of the matrix. Secondary structure prediction is done by two kinds of dynamic programming algorithms: the minimum free-energy algorithm of Zuker and Stiegler (1981) and the partition function algorithm of McCaskill (1990).

A.



B.

```
CKCGGUUCCAGAKGUUGCGCAGCGGCAAAAAGUGCUCCAUGGUCGGGACGCUCUGGCCGGUC
AGGCGCGCGCAGUCGUUGACGCUCUAGACCGUGCAAAAGGAGAGCCUGUAAGCGGGCACUCU
UCCGUGGUCUGGUGGAUAAAUUCGCAAGGGUAUCAUGGCGGACGACCGGGGUUCGAACCCCG
GAUCCGGCCGUCCGCCGUGAUCCAUGCGGUUACCGCCCGCGUGUCGAACCCAGGUGUGCGAC
GUCAGACAACGGGGGAGCGCUCCUUUUGGCUUCCUUCCAGGCGCGGCGGAUG
.................(((((((((.....((((.((.(((((((.....)))))))
....)).))))...)))))))..)))....(.((((((..(((((.(((((...(((.(((
(.(((((((((.(((((((((.....((((..((.(((((((((((((.(((((..(
((.....))).)))))..)))))))))))))))..)))))).......)))).(((((.((.
...)).)))))..)))))..)))))))))..)))))))..)))))))))....)))))))
.....
```

Figure 5.10. Suboptimal foldings of an RNA sequence using probability distributions of base-pairings. The first 300 bases of the same adenovirus sequence used in Fig. 5.8 was submitted to the Vienna Web server. (*A*) The region shown represents structures within the range of bases 150–300 and may be compared to the same region in Fig. 5.8. The minimum free energy of this thermodynamic ensemble is −134.85 kcal/mole, compared to a minimum free energy of 125.46 kcal/mole. The size of the square box at highlighted matrix positions indicates the probability of the base pair and decreases in steps of 10-fold; i.e., order of magnitude decreases. The size variations shown in the diagram cover a range of ~4–6 orders of magnitude. Calculations of base-pair probabilities are discussed in the text. (*B*) The minimum free-energy structure representing base pairs as pairs of nested parentheses. A low-resolution picture was also produced (not shown).

USING SEQUENCE COVARIATION TO PREDICT STRUCTURE

The second major method that has been used to make RNA secondary structure predictions (Woese et al. 1983) and also tertiary structure analyses such as those shown in Figure 5.3 (Gutell et al. 1986) is RNA sequence covariation analysis. This method examines sequences of the same RNA molecules from different species for positions that vary together in a manner that would allow them to produce a base pair in all of the molecules. The idea is quite simple. On the one hand, for double-stranded regions in RNA molecules, sequence changes that take place in evolution should maintain the base-pairing. On the other hand, sequence changes in loops and single-stranded regions should not have such a constraint. The method of analysis is to look for sequence positions at which covariation maintains the base-pairing properties. The justification for this method is that these types of joint substitutions or covariations actually are found to occur during evolution of such genes. As shown in Figure 5.11, when one position corresponding to a base pair is changed, another position corresponding to the base-pairing partner will also change. For example, if two positions G and C form a base pair, then sequences that have C and G reversed, or A and T or T and A at the corresponding positions, would also be considered reasonable matches. Sequence covariability has been used to improve thermodynamic structure prediction as described in the above section (Hofacker et al. 1998). An example of using covariation analysis to decipher base-pair interactions in tRNA is shown in Figure 5.12.

One method of covariation analysis also examines which phylogenetic groups exhibit change at a given position. For each position, the base that generally predominates in one particular part of the tree is determined. These methods have required manual examination of sequences and structures for covariation, but automatic methods have also been devised and demonstrated to produce reliable predictions (Winker et al. 1990; Han and Kim 1993; see box below).

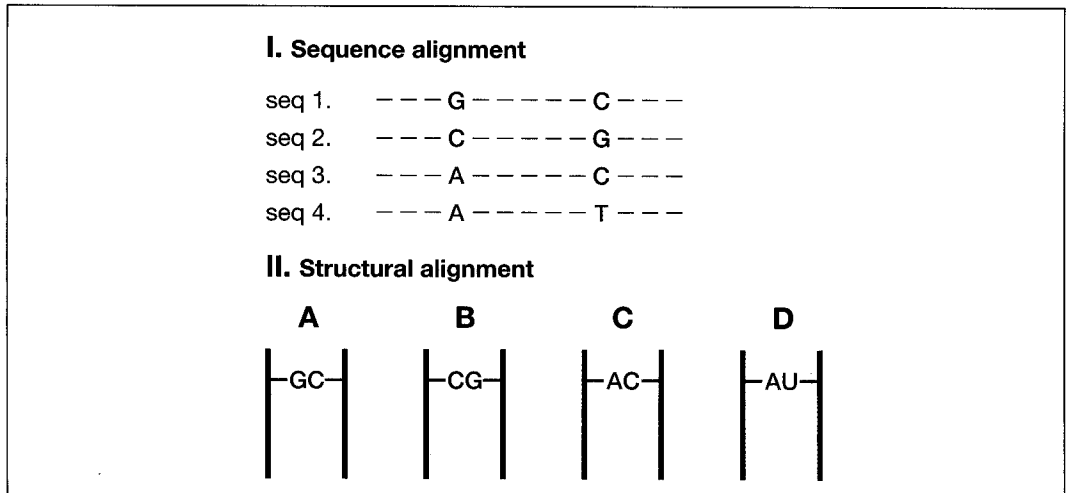


Figure 5.11. Conservation of base pairs in homologous RNA molecules influences structure prediction. The predicted structure takes into account sequence covariation found at aligned sequence positions, and may also use information about conserved positions in components of a phylogenetic tree. In the example shown, sequence covariations in A, B, and D found in sequences 1, 2, and 4, respectively, permit Watson–Crick base and G-U base-pairing in the corresponding structure, but variation C found in sequence 3 is not compatible. Sometimes correlations will be found that suggest other types of base interactions, or the occurrence of a common gap in a multiple sequence alignment may be considered a match. Positions with greater covariation are given greater weight in structure prediction. Molecules with only one of the two sequence changes necessary for conservation of the base-paired position may be functionally deleterious.

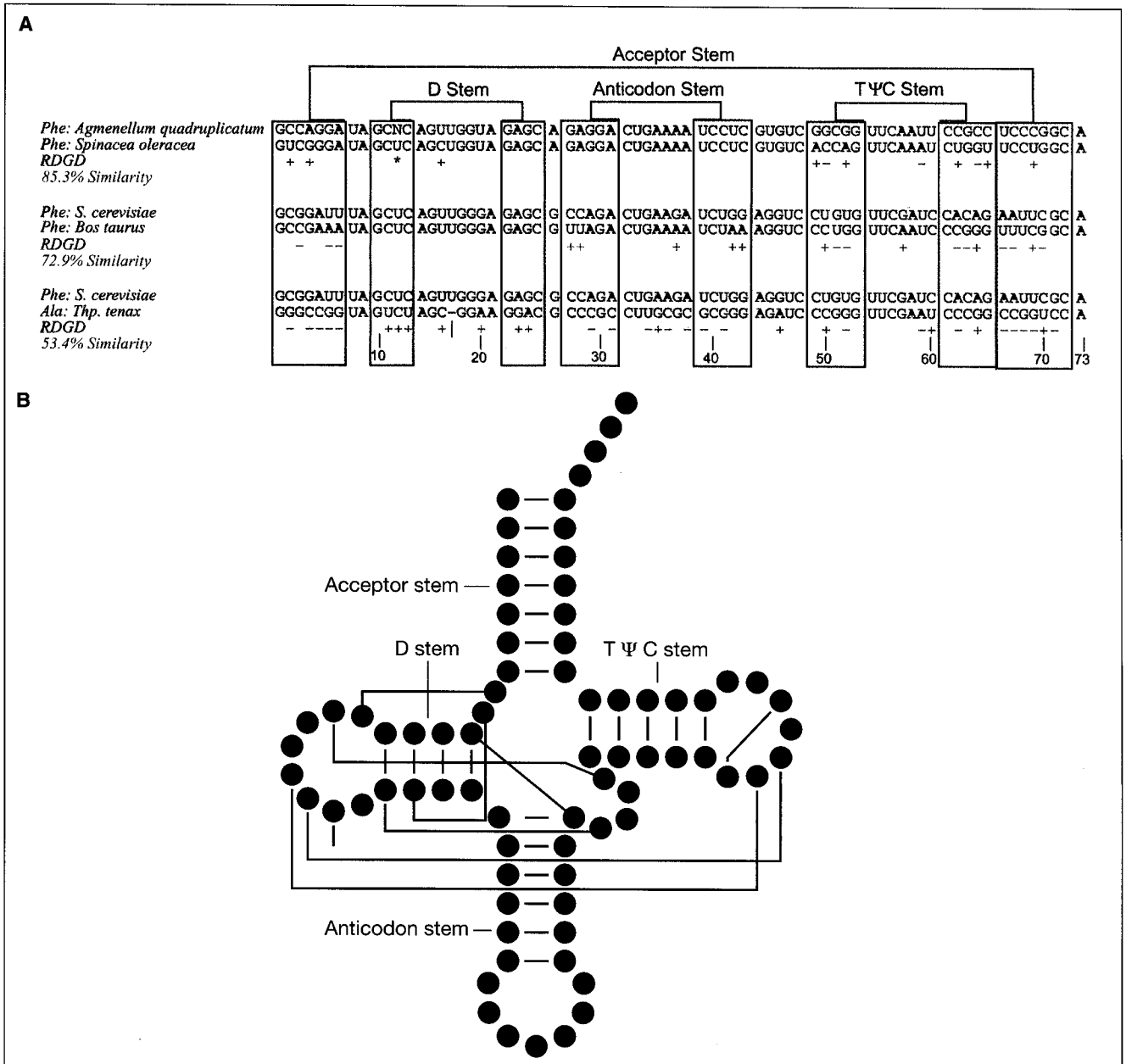


Figure 5.12. Covariation found in tRNA sequences reveals base interactions in tRNA secondary and tertiary structure. (*A*) Alignment of tRNA sequences showing regions of interacting base pairs. (+) Transition; (−) transversions; (|) deletion; (*) ambiguous nucleotide. (*B*) Diagram of tRNA structure illustrating base–base interactions revealed by a covariance analysis. Adapted from the Web site of R. Gutell at http://www.rna.icmb.utexas.edu.

Methods of Covariation Analysis in RNA Sequences

Secondary and tertiary features of RNA structure may be determined by analyzing a group of related sequences for covariation. Two sequence positions that covary in a manner that frequently maintains base-pairing between them provides evidence that the bases interact in the structure. Combinations of the following methods have been used to locate such covarying sites in RNA sequences (see R. Gutell for additional details and at http://www.rna.icmb.utexas.edu/METHODS/menu.html).

1. Optimally align pairs of sequence to locate conserved primary sequence, mark transitions and transversions from a reference sequence, and then visually examine these changes to identify complementary patterns that represent potential secondary structure.
2. Perform a multiple sequence alignment, highlight differences using one of the sequences as a reference, and visually examine for complementary patterns.
3. Mark variable columns in the multiple sequence alignment by numbers that mark changes (e.g., transitions or transversions) from a reference sequence; examine marked columns for a similar or identical number pattern that can represent potential secondary structure.
4. Perform a statistical analysis (Chi-square test) of the number of observations of a particular base pair in columns i and j of the multiple sequence alignment, compared to the expected number based on the frequencies of the two bases.
5. Calculate the mutual information score ($mixy$) for each pair of columns in the alignment, as described in the text and illustrated in Figure 5.13.
6. Score the number of changes in each pair of columns in the alignment divided by the total number of changes (the ec score), examine the phylogenetic context of these changes to determine the number of times the changes have occurred during evolution, and choose the highest scores that are representative of multiple changes.
7. Measure the covariance of each pair of positions in the alignment by counting the numbers of all 16 possible base-pair combinations and dividing by the expected number of each combination (number of sequence × frequency of base in first position × frequency of base in second position), choose the most prevalent pair, and examine remaining combinations for additional covariation; then sum frequency of all independently covarying sites to obtain covary score.

Mutual Information Content

A method used to locate covariant positions in a multiple sequence alignment is the mutual information content of two columns. First, for each column in the alignment, the frequency of each base is calculated. Thus, the frequencies in column m, $f_m(B_1)$, are $f_m(A)$, $f_m(U)$, $f_m(G)$, and $f_m(C)$ and those for column n, $f_n(B_2)$, are $f_n(A)$, $f_n(U)$, $f_n(G)$, and $f_n(C)$. Second, the 16 joint frequencies of two nucleotides, $f_{m,n}(B_1,B_2)$ one base B_1 in column m and the same or another base B_2 in column n are calculated. If the base frequencies in any two columns are independent of each other, then the

ratio of $f_{m,n}(B_1,B_2) / [f_m(B_1) \times f_n(B_2)]$ is expected to equal 1, and if the frequencies are correlated, then this ratio will be greater than 1. If they are perfectly covariant, then $f_{m,n}(B_1,B_2) = f_m(B_1) = f_n(B_2)$. To calculate the mutual information content $H(m,n)$ in bits between the two columns m and n, the logarithm of this ratio is calculated and summed over all possible 16 base-pair combinations.

$$H(m,n) = \Sigma_{B1,B2}\, f_{m,n}(B_1,B_2) \times \log_2 \{f_{m,n}(B_1,B_2) / [f_m(B_1)\, f_n(B_2)]\}$$

$H(m,n)$ varies from the value of 0 bits of mutual information representing no correlation to that of 2 bits of mutual information, representing perfect correlation (Eddy and Durbin 1994).

The mutual information content may be plotted on a motif logo (Gorodkin et al. 1997), similar to that described in Chapter 4, page 196, for illustrating a sequence motif. The example shown in Figure 5.13 shows the mutual information content M superimposed on the information content of each sequence position in an RNA alignment.

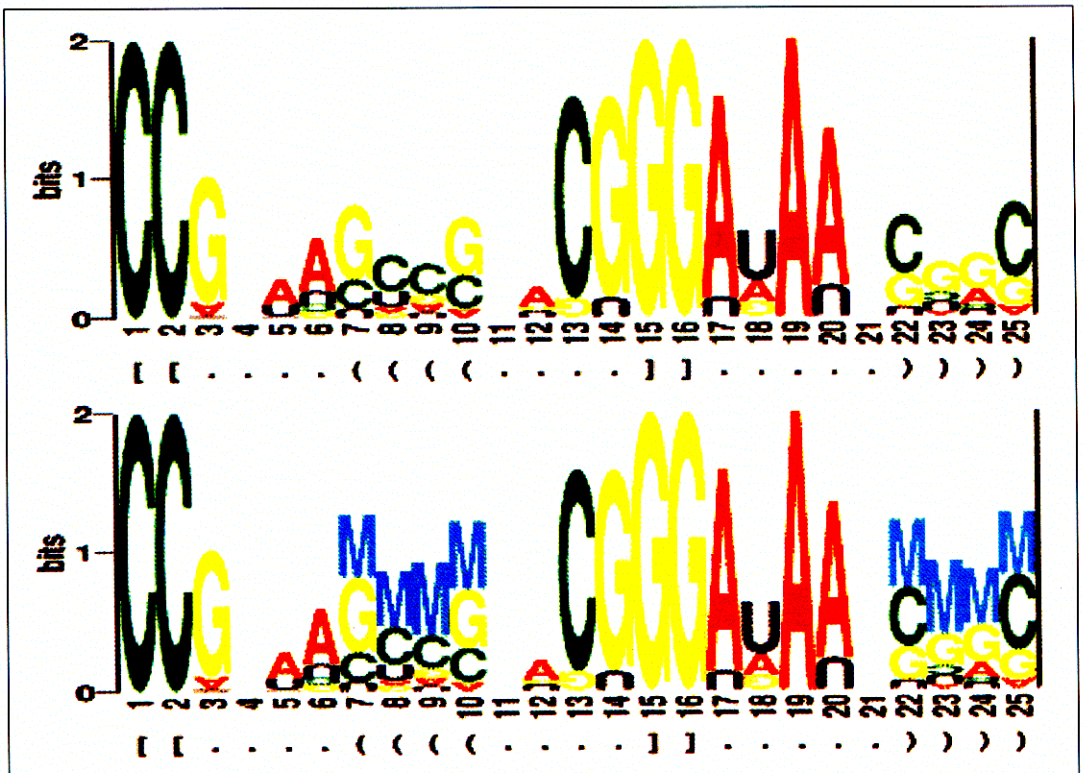


Figure 5.13. RNA structure logo. The top panel is the normal sequence logo showing the size of each base in proportion to the contribution of that base to the amount of information in that column of the multiple sequence alignment. The relative entropy method is used in which the frequency of bases in each column is compared to the background frequency of each base. Inverted sequence characters indicate a less than background frequency (see Chapter 4, page 196). The bottom panel includes the same information plus the mutual information content in pairs of columns. The amount of information is indicated by the letter M, and the matching columns are shown by nested sets of brackets and parentheses. All sequences have a C in column 1 and a matching G in column 16. Similar columns 2 and 15 can form a second base pair stacked upon the first. Columns 7–10 and 25–22 also can form G/C base pairs most of the time. Sequences with a G in column 7 frequently have a C in column 25, and those with a C in column 7 may have a G in column 25. Thus, there is mutual information in these two columns (Gorodkin et al. 1997 [using data of Tuerk and Gold 1990]).

A formal covariance model has been devised by Eddy and Durbin (1994). Although very accurate when used for identifying tRNA genes, the algorithm is extremely slow and unsuitable for searching through large genomes. Instead, the method has been used to screen through putative tRNA genes previously identified by faster methods (Lowe and Eddy 1997). The difficulty that is faced in modeling RNA molecules is to identify the potential base pairs in a set of related RNA molecules based on covariation at two sites. Recall from Chapter 4 that the hidden Markov model is used for capturing the types of variations observed in a sequence profile, including matches, mismatches, insertions, and deletions. This type of model assumes each sequence can be predicted by a series of states in the model, one after the other, as in a series of independent events in a Markov chain. The hidden Markov model does not analyze joint variations at sequence positions such as occur in RNA molecules. The model that is used for analyzing RNA secondary structure (but not tertiary structure) is an ordered tree model. A simplified tree representation of RNA secondary structure is shown in Figure 5.14.

The above assumes that we know which bases are paired in a model of RNA secondary structure, whereas the goal is to build a model that discovers this information. The task is achieved by constructing a more general model, training the model with a set of sequences,

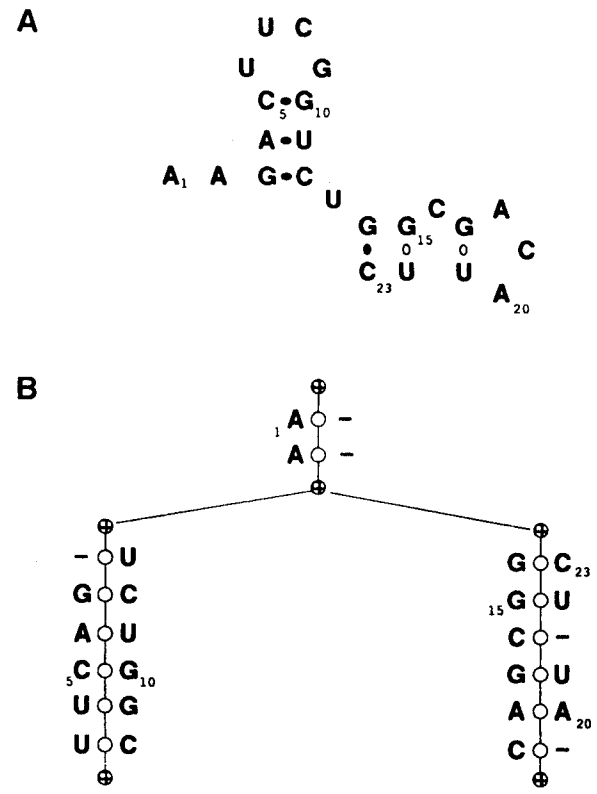


Figure 5.14. Tree model of RNA secondary structure. The model in *A* is represented by the ordered binary tree shown in *B*. This model attempts to capture both the sequence and the secondary structure of the RNA molecule. The tree is read like a sequence starting at the root node at the top of the model, then moving down the main branch to the bifurcation mode. Along the main trunk are nodes that represent matched or unmatched base pairs. Shown are two A's matching a "-," indicating no pairing with these bases. After the bifurcation mode, one then moves down the most leftward branch to the end node. Along the branch are unmatched bases, matched base pairs, and mismatched pairs. After the end node is reached, go back to the previous bifurcation node and follow the right branch. (Reprinted, with permission of Oxford University Press, from Eddy and Durbin 1994.)

and then having the model reveal the most likely base-paired regions. The approach is similar to training a hidden Markov model for proteins to recognize a family of protein sequences, thereby producing the most probable multiple sequence alignment. In the case of RNA secondary structure, a tree model is trained by the sequences, and the model may then be used to predict the most probable secondary structure. In addition, the model may also be used to search a database for sequences that produce a high score when aligned to the model. These sequences are likely to encode a similar type of RNA molecule such as tRNA or 5S RNA. Each model is derived by training a more general tree model with the sequences.

The general tree model needs to represent the types of variations that are found in aligning a series of related sequences, such as insertions, deletions, and mismatches. To allow for such variations, each node in the tree is replaced by a set of states that correspond to all of the possible sequence variations that might be encountered at that position. These states are illustrated in Figure 5.15.

The mutual information content of all sequence positions is used in designing the model, and the expectation maximization method (Chapter 4) is used to optimize the parameters of the model. A dynamic programming method is used to find a model that maximizes the amount of covariation. The structure of the model may subsequently be altered during training. Once a covariance model suitable for an RNA molecule has been established, the model is trained by the sequences. The methodology is similar to that of hidden Markov models and is described in detail in Chapter 4. Basically, the model is initialized by giving starting values to the base and dinucleotide frequencies in each MATCH and INS state and to the transition probabilities. All possible paths through the model are found for each sequence in the training set. The frequencies and transition probabilities are modified each time a particular path in the model is used. The base pairs are found from MATP (see Fig. 5.15), which gives probabilities to the 16 possible dinucleotides.

Once the model has been trained, the most probable path for each sequence provides a consensus structural alignment of the sequences. A dynamic programming algorithm is used that matches subsequence alignments to the nodes of the covariance model. The result is a log odds score of the sequence matching the covariance model. A similar method may be used to find sequences in a genomic database with high matching scores to the covariance model. The method was used to predict the structural alignment of representative sets of tRNA sequences, and it provided alignments that closely matched actual structural alignments based on other methods. The software for the COVELS program is available by request from the authors (Eddy and Durbin 1994).

STOCHASTIC CONTEXT-FREE GRAMMARS FOR MODELING RNA SECONDARY STRUCTURE

In the above section, we discussed the need to have models for RNA secondary structure that reflect the interaction among base pairs. Simpler models of sequence variation treat sequences as simple strings of characters without such interactions and are therefore not suitable for RNA. A general theory for modeling strings of symbols, such as bases in DNA sequences, has been developed by linguists. There is a hierarchy of these so-called transformational grammars that deal with situations of increasing complexity. The application of these grammars to sequence analysis has been extensively discussed elsewhere (Durbin et al. 1998). The context-free grammar is suitable for finding groups of symbols in different parts of the input sequence that thus are not in the same context. Complementary regions in sequences, such as those in RNA that will form secondary structures, are an

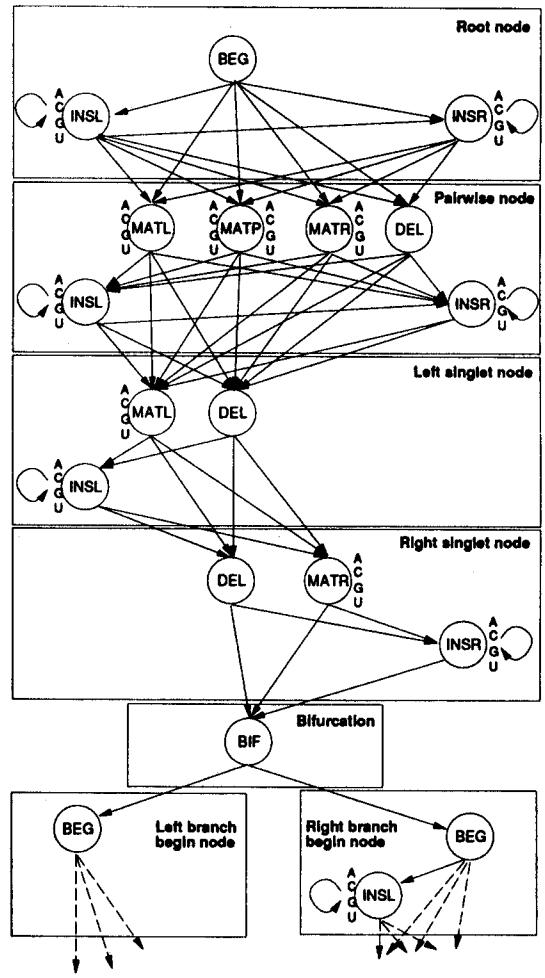


Figure 5.15. Details of tree model for RNA secondary structure. Each type of node in the tree shown in Fig. 5.14 is replaced by a pattern of states corresponding to the types of sequence variations that are expected in a family of related RNA sequences. These states each store a table of frequencies of 4 bases or of 16 possible dinucleotides. The seven different types of nodes are illustrated. BEG node includes insert states for sequence of any length on the right or left side of the node. The pair-wise node includes a state MATP for storing the 16 possible dinucleotide frequencies; MATL and MATR states for storing single base frequencies on either the left or right side of the node, respectively; a DEL state for allowing deletions; and INSL and INSR states that allow for insertions of any length on the left or right of the node. DEL does not store information. The other five node types have the same types of states. Each state is joined to other states by a set of transition probabilities shown by the arrows. These probabilities are similar to those used in hidden Markov models. BIF is a bifurcation state with transition probabilities entering the state from above and then leaving to one or the other of two branches. (Reprinted, with permission of Oxford University Press, from Eddy and Durbin 1994.)

example of such context-free sequences. Stochastic context-free grammars (SCFG) introduce uncertainty into the definition of such regions, allowing them to use alternative symbols as found in the evolution of RNA molecules. Thus, SCFGs can help define both the types of base interactions in specific classes of RNA molecules and the sequence variations at those positions. SCFGs have been used to model tRNA secondary structure (Sakakibara et al. 1994). Although SCFGs are computationally complex (Durbin et al. 1998), they are likely to play an important future role in identifying specific types of RNA molecules.

The application of SCFGs to RNA secondary structure analysis is very similar in form to the probabilistic covariance models described in the above section. For RNA, the symbols of the alphabet are A, C, G, and U. The context-free grammar establishes a set of rules called productions for generating the sequence from the alphabet, in this case an RNA molecule with sections that can base-pair and others that cannot base-pair. In addition to the sequence symbols (named terminal symbols because they end up in the sequence), another set of symbols (nonterminal symbols) designated S_0, S_1, S_2 . . . , determines intermediate production stages. The initial symbol is S_0 by convention. The next terminal symbol S_1 is produced by modifying S_0 in some fashion by productions indicated by an arrow. For example, the productions $S_0 \rightarrow S_1$, $S_1 \rightarrow C\ S_2\ G$ generate the sequence $C\ S_2\ G$ where S_2 has to be defined further by additional productions. The example shown in Figure 5.16 (from Sakakibara et al. 1994) shows a set of productions for generating the sequence CAUCAGGGAAGAUCUCUUG and also the secondary structure of this molecule. The productions chosen describe both features.

In this example of a context-free grammar, only one sequence is produced at each production level. In a SCFG, each production of a nonterminal symbol has an associated probability for giving rise to the resulting product, and there are a set of productions, each giving a different result. For example, the production $S_1 \rightarrow C\ S_2\ G$ could also be represented by 15 other base-pair combinations, and each of these has a corresponding probability. Thus, each production can be considered to be represented by a probability distribution over the possible outcomes. Note the identity of the SCFG representation of the predicted structure to that shown for the tree representation of the covariance model in Figure 5.14. The use of SCFGs in RNA secondary structure production analysis is in fact very similar to that of the covariance model, with the grammatical productions resembling the nodes in the ordered binary tree. As with hidden Markov models, the probability distribution of each production must be derived by training with known sequences. The algorithms used for training the SCFG and for aligning a sequence with the SCFG are somewhat different from those used with hidden Markov models, and the time and memory requirements are greater (Sakakibara et al. 1994: Durbin et al 1998).

SEARCHING GENOMES FOR RNA-SPECIFYING GENES

One goal in RNA research has been to design methods to identify sequences in genomes that encode small RNA molecules. Larger, highly conserved molecules can simply be identified based on their sequence similarity with already-known sequences. For smaller sequences with more sequence variation, this method does not work. A number of methods for finding small RNA genes have been described and are available on the Web (Table 5.1). A major problem with these methods in searches of large genomes is that a small false-positive rate becomes quite unacceptable because there are so many false positives to check out.

One of the first methods used to find tRNA genes was to search for sequences that are self-complementary and can fold into a hairpin like the three found in tRNAs (Staden 1980).

Figure 5.16. A set of transformation rules for generating an RNA sequence and the secondary structure of the sequence from the RNA alphabet (ACGU). (*A*) The set of production rules for producing the sequence and the secondary structure. These rules reveal which bases are paired and which are not paired. (*B*) Derivation of the sequence. (*C*) A parse tree showing another method for displaying the derivation of the sequence in *B*. (*D*) Secondary structure from applying the rules. (Redrawn, with permission of Oxford University Press, from Sakakibara et al. 1994.)

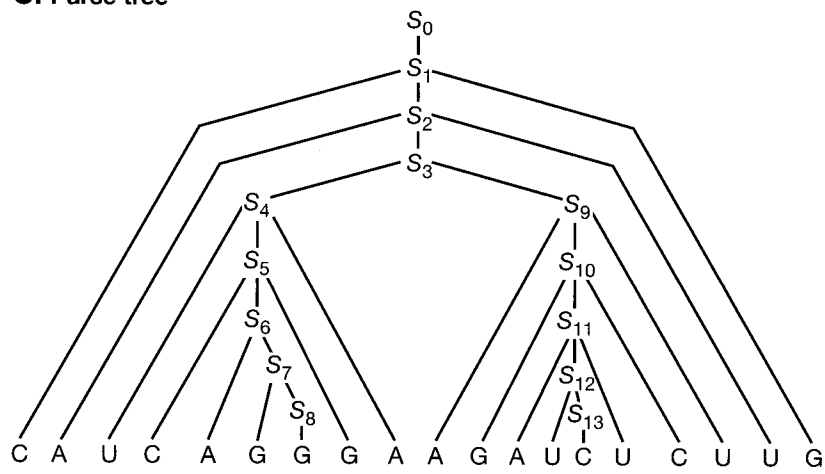
A. Productions

$P = \{ S_0 \rightarrow S_1,$ $S_7 \rightarrow G\,S_8,$
$S_1 \rightarrow C\,S_2\,G,$ $S_8 \rightarrow G,$
$S_2 \rightarrow A\,S_3\,U,$ $S_9 \rightarrow A\,S_{10}\,U,$
$S_3 \rightarrow S_4\ S_9,$ $S_{10} \rightarrow G\,S_{11}\,C,$
$S_4 \rightarrow U\,S_5\,A,$ $S_{11} \rightarrow A\,S_{12}\,U,$
$S_5 \rightarrow C\,S_6\,G,$ $S_{12} \rightarrow U\,S_{13},$
$S_6 \rightarrow A\,S_7,$ $S_{13} \rightarrow C$ $\}$

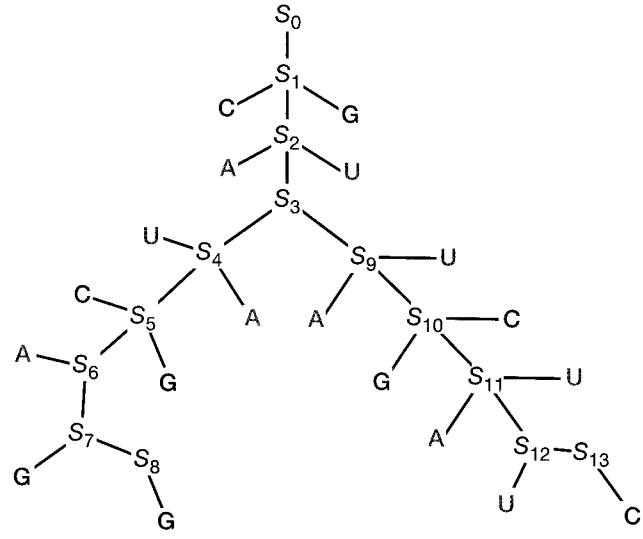
B. Derivation

$S_0 \rightarrow S_1 \rightarrow CS_2G \rightarrow CAS_3UG \rightarrow CAS_4S_9UG$
$\rightarrow CAUS_5AS_9UG \rightarrow CAUCS_6GAS_9UG$
$\rightarrow CAUCAS_7GAS_9UG \rightarrow CAUCAGS_8GAS_9UG$
$\rightarrow CAUCAGGGAS_9UG \rightarrow CAUCAGGGAAS_{10}UUG$
$\rightarrow CAUCAGGGAAGS_{11}CUUG$
$\rightarrow CAUCAGGGAAGAS_{12}UCUUG$
$\rightarrow CAUCAGGGAAGAUS_{13}UCUUG$
$\rightarrow CAUCAGGGAAGAUCUCUUG.$

C. Parse tree



D. Secondary structure



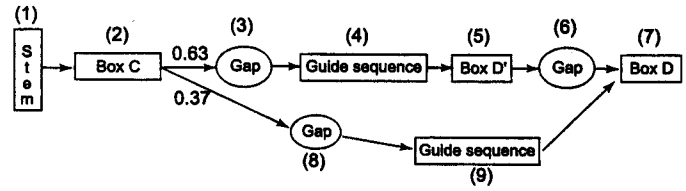


Figure 5.17. Probabilistic model of snoRNAs. The numbered boxes and ovals represent conserved sequence and structural features that have been modeled by training on snoRNAs. Secondary structural features of Stem were modeled with an SCFG. Boxes with ungapped hidden Markov models, the guide sequence with a hidden Markov model, and gapped regions (spacers) are shown by ovals. The guide sequence interacts with methylation sites on rRNA and is targeted in each search to a complementary sequence near one of those sites. The alignment of this model produces a log odds score that provides an indication of the reliability of the match. The transition probabilities are 1, except where the model bifurcates to allow identification of two types of target sequences. The model is highly specific and seldom identifies incorrect matches in random sequences. (Reprinted, with permission, from Lowe and Eddy 1999 [copyright AAAS, Washington, D.C.].)

Fichant and Burks (1991) described a program, tRNAscan, that searches a genomic sequence with a sliding window searching simultaneously for matches to a set of invariant bases and conserved self-complementary regions in tRNAs with an accuracy of 97.5%. Pavesi et al. (1994) derived a method for finding the RNA polymerase III transcriptional control regions of tRNA genes using a scoring matrix derived from known control regions that is also very accurate. Finally, Lowe and Eddy (1997) have devised a search algorithm tRNAscan-SE that uses a combination of three methods to find tRNA genes in genomic sequences—tRNAscan, the Pavesi algorithm, and the COVELS program based on sequence covariance analysis (Eddy and Durbin 1994). This method is reportedly 99–100% accurate with an extremely low rate of false positives.

The probabilistic model shown in Figure 5.17 was used to identify small nucleolar (sno) RNAs in the yeast genome that methylate ribosomal RNA. The model is not used to search genomic sequences directly. Instead, a list of candidate sequences is first found by searching for patterns that match the sequences in the model (Lowe and Eddy 1999). The probability model was a hybrid combination of HMMs and SCFGs trained on snoRNAs. These RNAs vary sufficiently in sequence and structure that they are not found by straightforward similarity searches. The RNAs found were shown to be snoRNAs by insertional mutagenesis.

APPLICATIONS OF RNA STRUCTURE MODELING

In summary, methods for predicting the structure of RNA molecules include (1) an analysis of all possible combinations of potential double-stranded regions by energy minimization methods and (2) identification of base covariation that maintains secondary and tertiary structure of an RNA molecule during evolution. Energy minimization methods have been so well refined that a series of energetically feasible models and the most thermodynamically probable structural models may be computed. Covariation analysis by C. Woese led to his building of detailed structural models for rRNAs. By examining the evolutionary variation in these structures, he was able to predict three domains of life—the Bacteria, the Eukarya, and a newly identified Archaea. Although a large amount of horizontal transfer among evolutionary lineages of other genes has added a great deal of noise to the evolutionary signal, the rRNA-based prediction is supported by other types of

genomic analyses. In addition to these uses of rRNA structural analysis, excellent probabilistic models of two small RNA molecules, tRNA and snoRNA, have been built, and these models may be used to search reliably through genomic sequences for genes that encode these RNA molecules. The successful analysis of these types of RNA molecules should be readily extensible to other classes of RNA molecules.

REFERENCES

Berman H.M., Zardecki C., and Westbrook J. 1998. The nucleic acid database: A resource for nucleic acid science. *Acta Crystallogr. D Biol. Crystallogr.* **54:** 1095–1104.

Brown J.W. 1999. The ribonuclease P database. *Nucleic Acids Res.* **27:** 314.

Burkhard M.E., Turner D.H., and Tinoco I., Jr. 1999a. The interactions that shape RNA secondary structure. In *The RNA world*, 2nd edition (ed. R.F. Gesteland et al.), pp. 233–264. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

———. 1999b. Appendix 2: Schematic diagrams of secondary and tertiary structure elements. In *The RNA world*, 2nd edition (ed. R.F. Gesteland et al.), pp. 681–685. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Ceci L.R., Volpicella M., Liuni S., Volpetti V., Licciulli F., and Gallerani R. 1999. PLMItRNA, a database for higher plant mitochondrial tRNAs and tRNA genes. *Nucleic Acids Res.* **27:** 156–157.

Chan L., Zuker M., and Jacobson A.B. 1991. A computer method for finding common base paired helices in aligned sequences: Application to the analysis of random sequences. *Nucleic Acids Res.* **19:** 353–358.

Chen R.O., Felciano R., and Altman R.B. 1997. RIBOWEB: Linking structural computations to a knowledge base of published experimental data. *Ismb* **5:** 84–87.

Chetouani F., Monestié P., Thébault P., Gaspin C., and Michot B. 1997. ESSA: An integrated and interactive computer tool for analysing RNA secondary structure. *Nucleic Acids Res.* **25:** 3514–3522.

De Rijk P., Neefs J.M., Van de Peer Y., and De Wachter R. 1992. Compilation of small ribosomal subunit RNA sequences. *Nucleic Acids Res.* **20:** 2075–2089.

De Rijk P., Robbrecht E., de Hoog S., Caers A., Van de Peer Y., and De Wachter R. 1999. Database on the structure of large subunit ribosomal RNA. *Nucleic Acids Res.* **27:** 174–178.

Dong S. and Searls D.B. 1994. Gene structure prediction by linguistic methods. *Genomics* **23:** 540–551.

Durbin R., Eddy S., Krogh A., and Mitchison G., Eds. 1998. *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*, chapters 9 and 10. Cambridge University Press, Cambridge, United Kingdom.

Eddy S. and Durbin R. 1994. RNA sequence analysis using covariance models. *Nucleic Acids Res.* **22:** 2079–2088.

Fichant G.A. and Burks C. 1991. Identifying potential tRNA genes in genomic DNA sequences. *J. Mol. Biol.* **220:** 659–671.

Freier S.M., Kierzek R., Jaeger J.A., Sugimoto N., Caruthers M.H., Neilson T., and Turner D.H. 1986. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci.* **83:** 9373–9377.

Gorodkin J., Heyer L.J., Brunak S., and Stormo G.D. 1997. Displaying the information contents of structural RNA alignments: The structure logos. *Comput. Appl. Biosci.* **13:** 583–586.

Gultyaev A.P., van Batenburg F.H., and Pleij C.W. 1995. The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.* **250:** 37–51.

Gutell R.R. 1994. Collection of small subunit (16S- and 16S-like) ribosomal RNA structures *Nucleic Acids Res.* **22:** 3502–3507.

Gutell R.R., Noller H.F., and Woese C.R. 1986. Higher order structure in ribosomal RNA. *EMBO J.* **5:** 1111–1113.

Han K. and Kim H.-J. 1993. Prediction of common folding structures of homologous RNAs. *Nucleic Acids Res.* **21:** 1251–1257.

Hofacker I.L., Fekete M., Flamm C., Huynen M.A., Rauscher S., Stolorz P.E., and Stadler P.F. 1998. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res.* **26:** 3825–3836.

Jacobson A.B. and Zuker M. 1993. Structural analysis by energy dot plot of a large mRNA. *J. Mol. Biol.* **233:** 261–269.

Jaeger J.A., Turner D.H., and Zuker M. 1989. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci.* **86:** 7706–7710.

———. 1990. Predicting optimal and suboptimal secondary structure for RNA. *Methods Enzymol.* **183:** 281–306.

Korab-Laskowska M., Rioux P., Brossard N., Littlejohn T.G., Gray M.W., Lang B.F., and Burger G. 1998. The organelle genome database project (GOBASE). *Nucleic Acids Res.* **26:** 138–144.

Lafontaine D.A., Deschenes P., Bussiere F., Poisson V., and Perreault J.P. 1999. The viroid and viroid-like RNA database. *Nucleic Acids Res.* **27:** 186–187.

Limbach P.A., Crain P.F., and McCloskey J.A. 1994. Summary: The modified nucleosides of RNA. *Nucleic Acids Res.* **22:** 2183–2196.

Lowe T.M. and Eddy S.R. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25:** 955–964.

———. 1999. A computational screen for methylation guide snoRNAs in yeast. *Science* **283:** 1168–1171.

Maidak B.L., Cole J.R., Parker C.T., Jr., Garrity G.M., Larsen N., Li B., Lilburn T.G., McCaughey M.J., Olsen G.J., Overbeek R., Pramanik S., Schmidt T.M., Tiedje J.M., and Woese C.R. 1999. A new version of the RDP (ribosomal database project). *Nucleic Acids Res.* **27:** 171–173.

Martinez H.M. 1984. An RNA folding rule. *Nucleic Acids Res.* **12:** 323–334.

Mathews D.H., Sabina J., Zuker M., and Turner D.H. 1999. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.* **288:** 911–940.

McCaskill J.S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29:** 1105–1119.

Nakaya A., Yamamoto K., and Yonezawa A. 1995. RNA secondary structure prediction using highly parallel computers. *Comput. Appl. Biosci.* **11:** 685–692.

Notredame C., O'Brien E.A., and Higgins D.G. 1997. RAGA: RNA sequence alignment by genetic algorithm. *Nucleic Acids Res.* **25:** 4570–4580.

Nussinov R. and Jacobson A.B. 1980. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci.* **77:** 6903–6913.

Pavesi A., Conterio F., Bolchi A., Dieci G., and Ottonello S. 1994. Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. *Nucleic Acids Res.* **122:** 1247–1256.

Pipas J.M. and McMahon J.E. 1975. Method for predicting RNA secondary structure. *Proc. Natl. Acad. Sci.* **72:** 2017–2021.

Rice P.M., Elliston K., and Gribskov M. 1991. DNA. In *Sequence analysis primer* (ed. M. Gribskov and J. Devereux), pp. 51–57. Stockton Press, New York.

Rozenski J., Crain P.F., and McCloskey J.A. 1999. The RNA modification database: 1999 update. *Nucleic Acids Res.* **27:** 196–197.

Sakakibara Y., Brown M., Hughey R., Mian I.S., Sjölander K., Underwood R.C., and Haussler D. 1994. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.* **22:** 5112–5120.

Samuelsson T. and Zwieb C. 2000. SRPDB (signal recognition particle database). *Nucleic Acids Res.* **28:** 171–172.

Sankoff D., Kruskal J.B., Mainville S., and Cedergren R.J. 1983. Fast algorithms to determine RNA secondary structures containing multiple loops. In *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison* (ed. D. Sankoff and J.B. Kruskal), chap. 3, pp. 93–120. Addison-Wesley, Reading, Massachusetts.

SantaLucia J., Jr. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci.* **95:** 1460–1465.

Schnare M.N., Damberger S.H., Gray M.W., and Gutell R.R. 1996. Comprehensive comparison of structural characteristics in eukaryotic cytoplasmic large subunit (23 S-like) ribosomal RNA. *J. Mol. Biol.* **256:** 701–719.

Serra M.J. and Turner D.H. 1995. Predicting thermodynamic properties of RNA. *Methods Enzymol.* **259:** 242–261.

Shapiro B.A. and Navetta J. 1994. A massively parallel genetic algorithm for RNA secondary structure prediction. *J. Supercomput.* **8:** 195–207.

Shumyatsky G. and Reddy R. 1993. Compilation of small RNA sequences. *Nucleic Acids Res.* **21:** 3017.
Simpson L., Wang S.H., Thiemann O.H., Alfonzo J.D., Maslov D.A., and Avila H.A. 1998. U-insertion/deletion edited sequence database. *Nucleic Acids Res.* **26:** 170–176.
Souza A.E. and Göringer H.U. 1998. The guide RNA database. *Nucleic Acids Res.* **26:** 168–169.
Spingola M., Grate L., Haussler D., and Ares M., Jr. 1999. Genome-wide bioinformatic and molecular analysis of introns of *Saccharomyces cerevisiae*. *RNA* **5:** 221–234.
Sprinzl M., Horn C., Brown M., Ioudovitch A., and Steinberg S. 1998. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **26:** 148–153.
Staden R. 1980. A computer program to search for tRNA genes. *Nucleic Acids Res.* **8:** 817–825.
Studnicka G.M., Rahn G.M., Cummings I.W., and Salser W.A. 1978. Computer method for predicting the secondary structure of single-stranded RNA. *Nucleic Acids Res.* **5:** 3365–3387.
Sühnel J. 1997. Views of RNA on the world wide web. *Trends Genet.* **13:** 206–207.
Szymanski M., Barciszewska M.Z., Barciszewski J., and Erdmann V.A. 1999. 5S ribosomal RNA Data Bank. *Nucleic Acids Res.* **27:** 158–160.
Tinoco I., Jr., Uhlenbeck O.C., and Levine M.D. 1971. Estimation of secondary structure in ribonucleic acids. *Nature* **230:** 362–367.
Tinoco I., Jr., Borer P.N., Dengler B., Levine M.D., Uhlenbeck O.C., Crothers D.M., and Gralla J. 1973. Improved estimation of secondary structure in ribonucleic acids. *Nat. New Biol.* **246:** 40–41.
Triman K.L. and Adams B.J. 1997. Expansion of the 16S and 23S ribosomal RNA mutation databases (16SMDB and 23SMDB). *Nucleic Acids Res.* **25:** 188–191.
Tuerk C. and Gold L. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249:** 505–510.
Tuerk C., Gauss P., Thermes C., Groebe D.R., Gayle M., Guild N., Stormo G., d'Aubenton-Carafa Y., Uhlenbeck O.C., Tinoco I., Jr., et al. 1988. CUUCGG hairpins: Extraordinarily stable RNA secondary structures associated with various biochemical processes. *Proc. Natl. Acad. Sci.* **85:** 1364–1368.
Turner D.H. and Sugimoto N. 1988. RNA structure prediction. *Annu. Rev. Biophys. Biophys. Chem.* **17:** 167–192.
von Heijne G. 1987. *Sequence analysis in molecular biology — Treasure trove or trivial pursuit*, pp. 58–72. Academic Press, San Diego, California.
Waterman M.S. and Byers T.H. 1985. A dynamic programming algorithm to find all solutions in a neighborhood of the optimum. *Math. Biosci.* **77:** 179–188.
Williams K.P. 1999. The tmRNA website. *Nucleic Acids Res.* **27:** 165–166.
Winker R., Overbeek R., Woese C., Olsen G.J., and Pfluger N. 1990. Structure detection through automated covariance search. *Comput. Appl. Biosci.* **6:** 365–371.
Woese C.R., Gutell R., Gupta R., and Noller H.F. 1983. Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. *Microbiol. Rev.* **47:** 621–669.
Wower J. and Zwieb C. 1999. The tmRNA database (tmRDB). *Nucleic Acids Res.* **27:** 167.
Wuchty S., Fontana W., Hofacker I.L., and Schuster P. 1999. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **49:** 145–165.
Zuker M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science* **244:** 48–52.
———. 1991. Suboptimal sequence alignment in molecular biology. Alignment with error analysis. *J. Mol. Biol.* **221:** 403–420.
———. 1994. Predicting optimal and suboptimal secondary structure for RNA. *Methods Mol. Biol.* **25:** 267–294.
Zuker M. and Jacobson A.B. 1995. "Well-determined" regions in RNA secondary structure prediction: Analysis of small subunit ribosomal RNA. *Nucleic Acids Res.* **23:** 2791–2798.
———. 1998. Using reliability information to annotate RNA secondary structures. *RNA* **4:** 669–679.
Zuker M. and Sankoff D. 1984. RNA secondary structures and their prediction. *Bull. Math. Biol.* **46:** 591–621.
Zuker M. and Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9:** 133–148.
Zuker M., Jaeger J.A., and Turner D.H. 1991. A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nucleic Acids Res.* **19:** 2707–2714.
Zwieb C. 1997. The uRNA database. *Nucleic Acids Res.* **25:** 102–103.