CHAPTER SIX

Predictive Methods Using RNA Sequences

DAVID MATHEWS

MICHAEL ZUKER

Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Third Edition, edited by Andreas D. Baxevanis and B.F. Francis Ouellette.
ISBN 0-471-47878-4

INTRODUCTION

RNA is a versatile biopolymer that plays many roles beyond simply carrying and recognizing genetic information as messenger RNA (mRNA) and transfer RNA (tRNA). It has been known for two decades that RNA sequences can catalyze phosphodiester bond cleavage and ligation (Doudna & Cech, 2002) and that RNA is an important component in the signal recognition particle (Walter & Blobel, 1982). More recently, other roles have been discovered for RNA, including roles in development (Lagos-Quintana et al., 2001; Lau et al., 2001), the immune system (Cullen, 2002), and peptide bond catalysis (Hansen et al., 2002; Nissen et al., 2000). Furthermore, RNA can be made to evolve *in vitro* to catalyze reactions that do not naturally occur (Bittker et al., 2002). RNA is also an important target and agent for the pharmaceutical industry. In the ribosome, RNA is the target of several classes of antibiotics. mRNA is the target of drugs that work on the antisense principle (Dias & Stein, 2002) or by redirecting alternative splicing (Sazani & Kole, 2003). RNA sequences can also be tailored to catalyze therapeutic reactions, such as gene repair (Long et al., 2003).

To understand fully its mechanism of action or to target an RNA sequence, the structure of RNA needs to be understood. RNA structure has three levels of organization, as shown in Figure 6.1. The first level, primary structure, is the linear sequence of nucleotides. Secondary structure is the collection of canonical base pairs (meaning both Watson-Crick pairs and G-U pairs) in the RNA structure. Finally, tertiary structure is the three-dimensional arrangement of the atoms in the RNA sequence, and hence includes all of the noncanonical contacts.

Often, the secondary structure of an RNA sequence is solved before its tertiary structure because there are accurate methods for determining the secondary structure of an RNA sequence and because the knowledge of the secondary structure often is helpful in designing constructs for tertiary structure determination. A typical RNA secondary structure, illustrated in Figure 6.2, is composed of both helical and loop regions. The helical

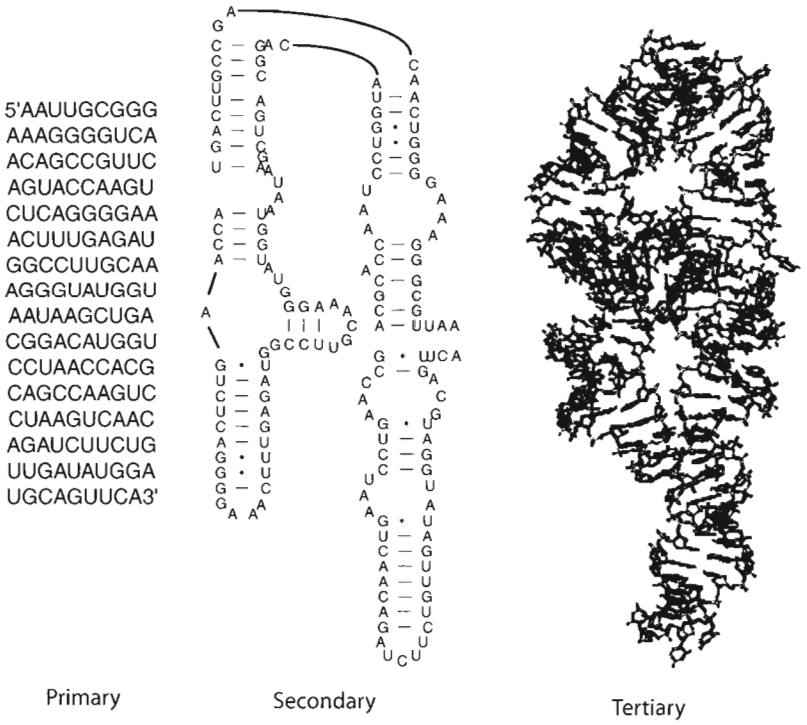

FIGURE 6.1 The three levels of organization of RNA structure. From left to right are the primary sequence, the secondary structure (Cannone et al., 2002), and the tertiary structure (Cate et al., 1996) of a domain of the group I intron from *Tetrahymena*. The secondary structure illustrates the canonical base pairs, and the tertiary structure is the actual three-dimensional arrangement of atoms.

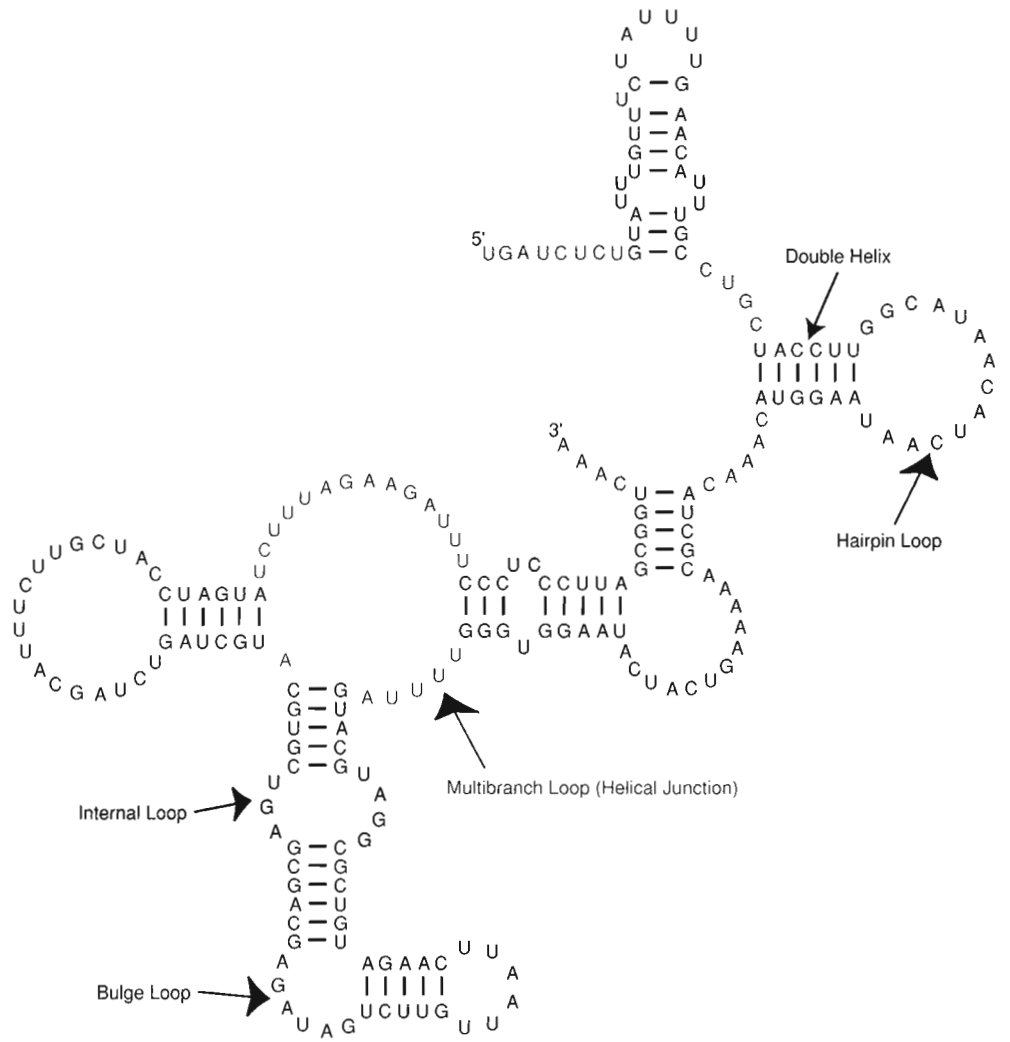

FIGURE 6.2 The RNA secondary structure of the 3′ UTR from the *D. sucinea* R2 element (Lathe & Eickbush, 1997; Mathews et al., 1997). Base pairs in nonhelical regions, known as loops, are colored by type of loop.

regions are held together by canonical base pairs. The loop regions fall into four broad categories: hairpin loops, in which the backbone makes a 180° bend; internal loops, in which the pairing of both strands is interrupted; bulge loops, in which the pairing of one strand is interrupted; and multibranch loops (also called helical junctions), from which more than two helices exit. Although secondary structure diagrams often do not illustrate explicitly the specific nucleotide interactions within the loop regions, these are sites of many noncanonical interactions that stabilize the structure.

The gold standard for predicting the placement of loops and helices, in the absence of a tertiary structure, is comparative sequence analysis, which uses evolutionary evidence found in sequence alignments to determine base pairs (Pace et al., 1999). Base pairs predicted by comparative sequence analysis for large and small subunit rRNA are 97% accurate when compared with high-resolution crystal structures (Gutell et al., 2002).

This chapter presents current methods for RNA secondary structure prediction, including methods applicable to a single sequence and methods applicable to multiple available sequences. To that end, RNA folding thermodynamics and dynamic programming are introduced. A detailed example for applying secondary structure prediction to a single sequence is drawn from the R2 retrotransposon 3′ untranslated region (UTR) RNA sequences (Eickbush, 2002). This chapter concludes with a brief introduction to the methods used for RNA tertiary structure prediction.

RNA SECONDARY STRUCTURE THERMODYNAMICS

Most methods for RNA secondary structure prediction rely on free energy minimization using nearest-neighbor parameters for predicting the stability of an RNA secondary structure, in terms of Gibbs free energy at 37°C (ΔG°_{37}; Mathews et al., 1999b; Turner, 2000; Xia et al., 1999; Xia et al., 1998). The rules for predicting stability are called *nearest-neighbor parameters* because the stability of each base pair depends only on the most adjacent pairs; the total free energy is the sum of each contribution.

An example of a nearest-neighbor stability calculation is shown in Figure 6.3. Terms for helical stacking, loop initiation, and unpaired nucleotide stacking contribute to the total conformational free energy. Favorable free energy increments are less than zero. The free energy increments of base pairs are counted as stacks of adjacent pairs. The consecutive CG base pairs, for example, are worth −3.3 kcal/mol (Xia et al., 1998). Note that the loop regions have unfavorable increments called *loop initiation energies* that largely reflect an entropic cost for constraining the nucleotides in the loop. For example, the hairpin loop of four nucleotides has an initiation of 5.6 kcal/mol (Mathews et al., 1999b). Unpaired nucleotides in loops can provide favorable energy increments as either stacked nucleotides or as mismatched pairs. The 3′-most G, called a *dangling end*, stacks on the terminal base pair and provides −1.3 kcal/mol of stability. The first mismatch in the hairpin loop with this sequence context is worth −1.1 kcal/mol.

The Gibbs free energy of formation for an RNA structure (ΔG°) quantifies the equilibrium stability of

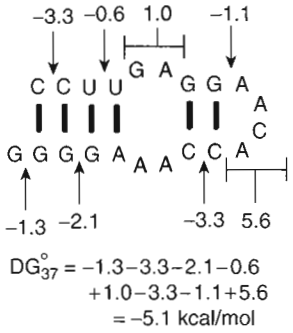

FIGURE 6.3 Prediction of conformational free energy for a conformation of rCCUUGAGGAACACCAAAGGGG. Each contributing free energy increment is labeled. The total free energy is the sum of each increment.

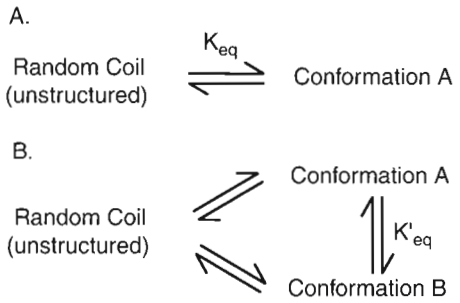

FIGURE 6.4 Equilibria of structures in solution. (a) The equilibrium between conformation *A* and the random coil structure. K_{eq}, related to the ΔG°_{37}, describes the equilibrium. (b) The equilibrium between two conformations, *A* and *B*, and the random coil. K'_{eq}, which is related to the free energy of folding for both *A* and *B*, describes the population of conformation *A* versus conformation *B*.

that structure at a specific temperature. For example, consider the RNA structure *A* at equilibrium with the random-coil (i.e., unstructured) conformation. The relative concentration of each conformation is governed by the equilibrium constant, K_{eq}, as illustrated in Figure 6.4a. K_{eq} is related to Gibbs free energy by the relationship shown in Equation 6.1:

$$K_{eq} = \frac{[\text{Conformation } A]}{[\text{Random Coil}]} = e^{-\Delta G^\circ/RT}, \qquad (6.1)$$

where R is the universal gas constant and T is the absolute temperature. For the example in Figure 6.3, with a predicted stability of −5.1 kcal/mol, there is a population of 3900 folded strands to every unfolded one (K_{eq} = 3900).

Furthermore, for multiple alternative conformations *A* and *B* for which there is an equilibrium distribution of conformations, K'_{eq}, as shown in Figure 6.4b, describes the distribution of strands between the structures. In this case, the free energy of each conformation relative to random coil also describes the population of each conformation, as shown in Equation 6.2:

$$K'_{eq} = \frac{[\text{Conformation } A]}{[\text{Conformation } B]} = e^{-(\Delta G^\circ_A - \Delta G^\circ_B)/RT}. \qquad (6.2)$$

This generalizes to any number of conformations. Therefore, the lowest free energy conformation is the most probable conformation for an RNA at equilibrium.

The nearest-neighbor free energy parameters use sequence-dependent terms for predicting the free

energy increments of loop regions (Mathews et al., 1999b) to reflect experimental observations. For example, a symmetric 2 × 2 internal loop can vary in stability from −2.6 to +2.8 kcal/mol, depending on the sequence of the closing pair and mismatches (Schroeder et al., 1999), corresponding to a K_{eq} of 6.4×10^3.

DYNAMIC PROGRAMMING

In the last section, the concept that the lowest free energy structure is the most likely structure for an RNA sequence at equilibrium was introduced. Given that there are nearest-neighbor parameters for predicting the free energy of a given sequence and structure, how, then, is the secondary structure predicted? The naïve approach would be to generate each possible conformation explicitly, to evaluate the free energy of each conformation, and then to choose the conformation that had the lowest free energy.

One estimate is that there are $(1.8)^N$ secondary structures possible for a sequence of N nucleotides (Zuker & Sankoff, 1984). This translates to 3×10^{25} structures for a modest length sequence of 100 nucleotides. Given that a fast computer can calculate the free energy for 10,000 structures in a second, this approach would still require 1.6×10^{14} CPU years! Clearly, a faster solution is needed for this problem.

The most commonly used solution for computationally intensive problems such as this is dynamic programming, which uses recursion to speed the calculation

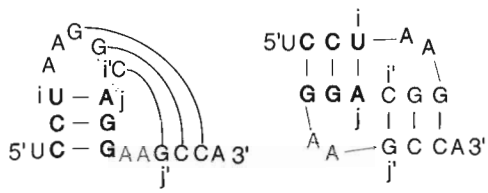

FIGURE 6.5 A simple RNA pseudoknot. This figure illustrates two representations of the same simple, H-type pseudoknot. A pseudoknot is defined by two base pairs such that i–j and i′–j′ are two pairs with ordering $i < i' < j < j'$. The base pair between nucleotides i and j defines an enclosed region. The base pair i′ and j′ spans the enclosed region and an adjacent region, making the pseudoknot.

(Nussinov & Jacobson, 1980; Zuker & Stiegler, 1981). Appendix 6.1 describes this method in detail for the interested reader. Modern implementations (Mathews et al., 1999b; Wuchty et al., 1999) of the dynamic programming algorithm for RNA secondary structure prediction also predict structures with free energy greater than the lowest free energy structure. These are called *suboptimal structures* (Zuker, 1989).

The dynamic programming algorithm for secondary structure prediction is $O(N^3)$ in time and $O(N^2)$ in storage when pseudoknots are excluded from the calculation (see Box 6.1). A pseudoknot, illustrated in Figure 6.5, occurs when there are nonnested base pairs. For example, the simplest pseudoknot occurs for which there are two base pairs i–j and i′–j′ such that $i < i' < j < j'$. It had been assumed that pseudoknots could not be predicted by a polynomial time dynamic programming until Rivas and Eddy (1999) presented a polynomial time dynamic programming algorithm that can predict structures containing a certain class of pseudoknots that is sufficiently rich to cover all cases of practical importance. Their algorithm, however, is $O(N^6)$ in time and $O(N^4)$ in storage, making the calculation impractical for sequences longer than approximately 300 nucleotides.

BOX 6.1 Algorithm Complexity

Algorithm complexity describes the scaling of a calculation in the worst-case scenario. It is expressed using the "Big-O" notation, which can read as "order." Algorithms that are $O(N)$ in time require a linear increase in time as the size parameter, N, lengthens. $O(N^2)$ and $O(N^3)$ algorithms scale by the square and cube of the parameter N. Therefore, the dynamic programming algorithm for RNA secondary structure prediction, which is $O(N^3)$, where N is the number of nucleotides, requires roughly eight times the execution time for a sequence twice as long. This is a fairly expensive calculation as compared to sorting a list, which can generally be accomplished in $O(N \log(N))$ time.

The Big-O notation also applies to the scaling of memory (also called storage) used by an algorithm. Secondary structure prediction requires two-dimensional arrays of size N × N. Therefore, in storage, the secondary structure prediction algorithm is $O(N^2)$.

ACCURACY OF RNA SECONDARY STRUCTURE PREDICTION

The accuracy of RNA secondary structure can be assessed by predicting structures for RNA sequences with known secondary structures, as determined by comparative sequence analysis. For a collection of structures assembled to test the accuracy of prediction, which included small subunit rRNA (Cannone et al., 2002),

large subunit rRNA (Cannone et al., 2002), 5S rRNA (Szymanski et al., 2000), group I introns (Cannone et al., 2002), group II introns (Michel et al., 1989), RNase P RNA (Brown, 1999), SRP RNA (Larsen et al., 1998), and tRNA (Sprinzl et al., 1998), 73% of base pairs in the known structure can, on average, be correctly predicted (Mathews et al., 1999a). For these calculations, the small and large subunit rRNA are divided into domains of fewer than 700 nucleotides, based on the known structure (Mathews et al., 1999b).

It has been demonstrated that the prediction accuracy can be improved by constraining secondary structure prediction with enzymatic constraints. Enzymes are used to determine nucleotides that are single or double stranded (Knapp, 1989). For the 5S rRNA sequence from *Escherichia coli*, which is poorly predicted without experimental constraints, the accuracy improves from 26% to 87% when enzymatic cleavage data are included (Mathews et al., 1999b; Speek & Lind, 1982; Szymanski et al., 2000).

PROGRAMS AVAILABLE FOR RNA SECONDARY STRUCTURE PREDICTION OF A SINGLE SEQUENCE

MFold

Mfold is an RNA secondary structure prediction package available through a Web frontend and as code for compilation on Unix and Linux machines (Mathews et al., 1999b; Zuker, 2003). It uses the current set of nearest neighbor parameters for free energies at 37°C (Mathews et al., 1999b). Minimum free energy and suboptimal secondary structures, sampled heuristically (Zuker, 1989), are predicted. Predicted suboptimal structures represent alternative structures to the lowest free energy structure and reflect both the possibility that an RNA sequence may have more than a single structure (Schultes & Bartel, 2000) and the fact that the energy rules contain some uncertainty (Mathews et al., 1999b). *Mfold* also predicts energy dot plots, which display the lowest free energy conformation possible for each possible base pair (Zuker & Jacobson, 1995). These plots conveniently demonstrate all possible base pairs within a user-specified increment of the lowest free energy structure, and predicted structures can be color annotated to demonstrate regions in the structure for which many folding alternatives exist (Zuker & Jacobson, 1998).

Figure 6.6 shows the input form for the *mfold* RNA server. A separate server for secondary structure prediction of DNA, using DNA folding free energies (SantaLucia, 1998), is available by following the link to the *DNA mfold server*. A sequence name can be entered in the box labeled *Enter a name for your sequence* and the sequence is typed (or pasted from the clipboard) in the box labeled *Enter the Sequence to be folded in the box*. As the caption explains, blanks and nonalphabetic characters are ignored and do not interfere with sequence interpretation. For example, the form shows the tRNA sequence (Sprinzl et al., 1998), RD1140, pasted into the sequence field. The remainder of the form has default values that can be changed by advanced users. The next box provides the option of constraining structure prediction with auxiliary evidence derived from enzymatic cleavage (Knapp, 1989), comparative sequence analysis (Pace et al., 1999), or intuition. Next, the default is for linear RNA sequence folding, although circular sequences also can be folded by changing the option from `linear` to `circular`. Note that the folding temperature is fixed at 37°C using the current parameters. An older, less complete set of parameters allows secondary structure prediction at other temperatures (Jaeger et al., 1989), but it is recommended that the current parameters be used for most applications. The older parameters can be used for folding by following the link at the top of the page to *RNA mfold version 2.3 server* (not shown in Figure 6.6). The percent suboptimality number, 5 by default, is the maximum percent difference in free energy from the lowest free energy structure that is allowed when generating suboptimal secondary structures. The upper bound on the computed foldings (default = 50) is the maximum number of suboptimal secondary structures to be predicted. The window parameter controls how different each suboptimal structure must be from all others. It defaults to a value based on the length of the sequence that is shown by following the link at *Window*. For example, the tRNA used here is 77 nucleotides long and will have a default window of 2. A smaller window allows for more suboptimal structures and a larger window yields greater differences between the predicted structures. The smallest window size allowed is zero. The maximum number of unpaired nucleotides in bulge or internal loops is limited to 30, by default. The maximum asymmetry in internal loops (the difference in length in unpaired nucleotides on each strand) is also 30 by default. The maximum distance allowed between paired nucleotides defaults to `no limit`. These values can be modified, as appropriate.

The remaining options control the server output. Currently, sequences of 800 or fewer nucleotides can be folded and the results returned as an "immediate job." Longer sequences must be folded as a batch job, requiring that the default option be changed from `An immediate` to `A batch` job. Batch jobs also require that

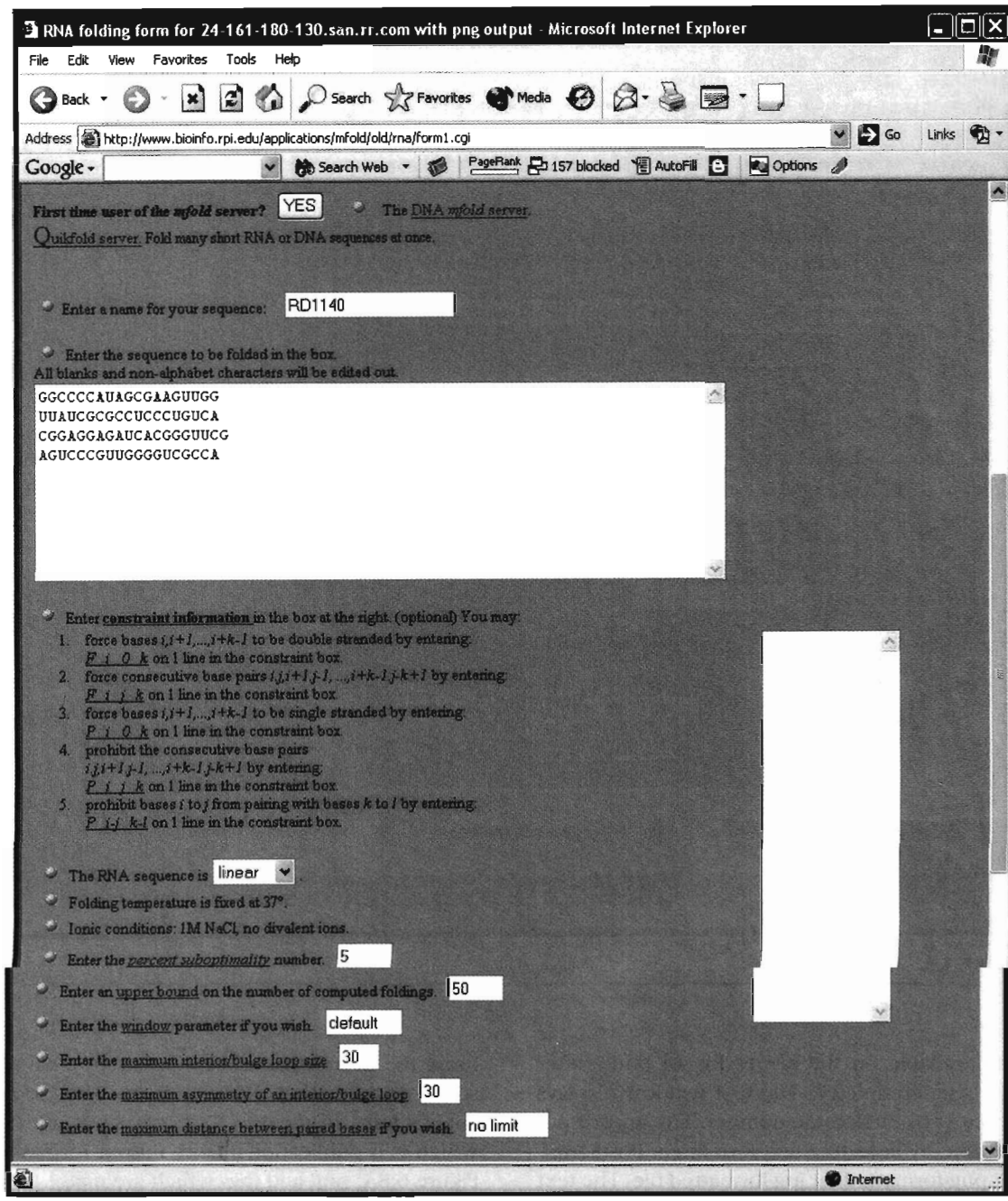

FIGURE 6.6 The input form for the version 3.1 *mfold* server. (a and b) The top and the bottom of the form, respectively. Default parameters are shown, with the exceptions noted in the text.

the user enter an E-mail address for receiving notification that the calculation is complete. The tRNA in this example is short, so the default of `An immediate` job will be used. The remaining options control the way the server generates output. Each of these options has link to a Web page that describes each parameter. For this example, color annotation by p-num is turned on to show regions in the predicted structure that having alternative low energy base pairs to those in the minimum free energy structure. By default, color annotation is not included. The button labeled *Fold RNA* is clicked to start the calculation.

Figure 6.7 shows the *mfold* server output form for the secondary structure prediction of the RD1140 tRNA.

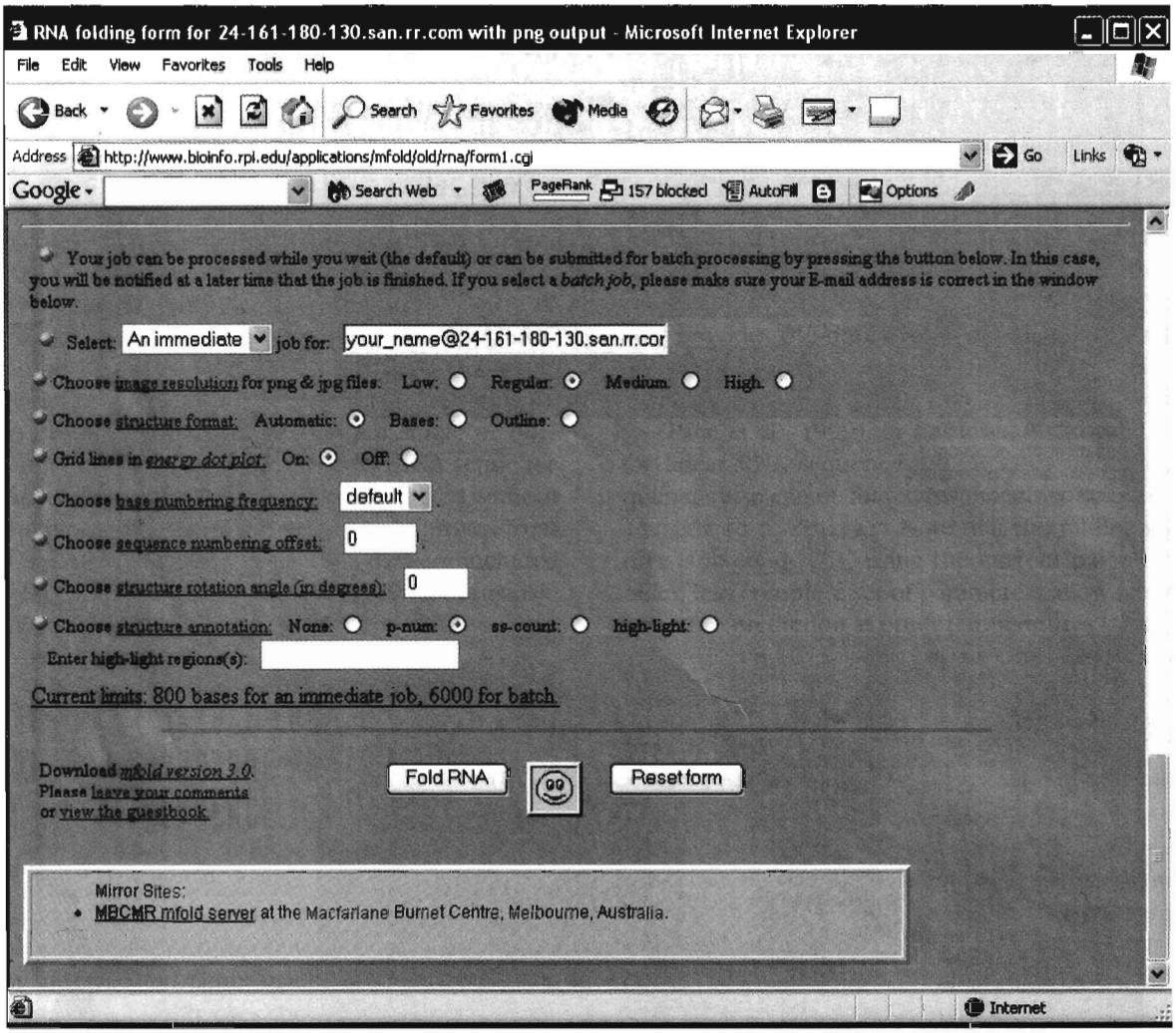

FIGURE 6.6 *Continued*

Results are available on the server for 24 hours after the job has been completed. The first window displays the sequence with nucleotide number. The energy dot plot is available by links to a text-formatted, PostScript-formatted, png-formatted, or jpg-formatted file. The text format is suitable for analysis in other software programs. PostScript is a publication-quality output and is shown in Figure 6.8b. Png and jpg both link to interactive pages that allow the user to zoom to regions, change the energy increment and number of colors, and click on individual base pairs to determine the exact energy. An RNAML-formatted output file is available for exchanging information with other RNAML-compliant programs. This is an xml file format that promises eventually to allow seamless information exchange between RNA analysis programs (Waugh et al., 2002). A diagram of each predicted secondary structure is available in a variety of formats. For this example, only a single structure is predicted using the default parameters for suboptimal secondary structure prediction. The commonly used formats, available by links adjacent to `Structure 1`, are PostScript, which is a publication-quality output format shown in Figure 6.8A; png and jpg, which are image formats that allow user interaction; and RNAViz CT and XRNAss formats, which are export formats for secondary structure drawing tools, explained below.

Figure 6.8 demonstrates sample output for the *mfold* server using the tRNA sequence for RD1140 (Sprinzl et al., 1998). The predicted secondary structure (Figure 6.8a) is color annotated according to the number of competing pairs in the energy dot plot (Figure 6.8b). Nucleotides outlined in red are in well-defined regions with no competing base pairs. The stem with black outlined pairs is less well-defined than the other stems, according to the dot plot. In the dot plot, each dot represents a base pair between nucleotides indicated on the x-axis

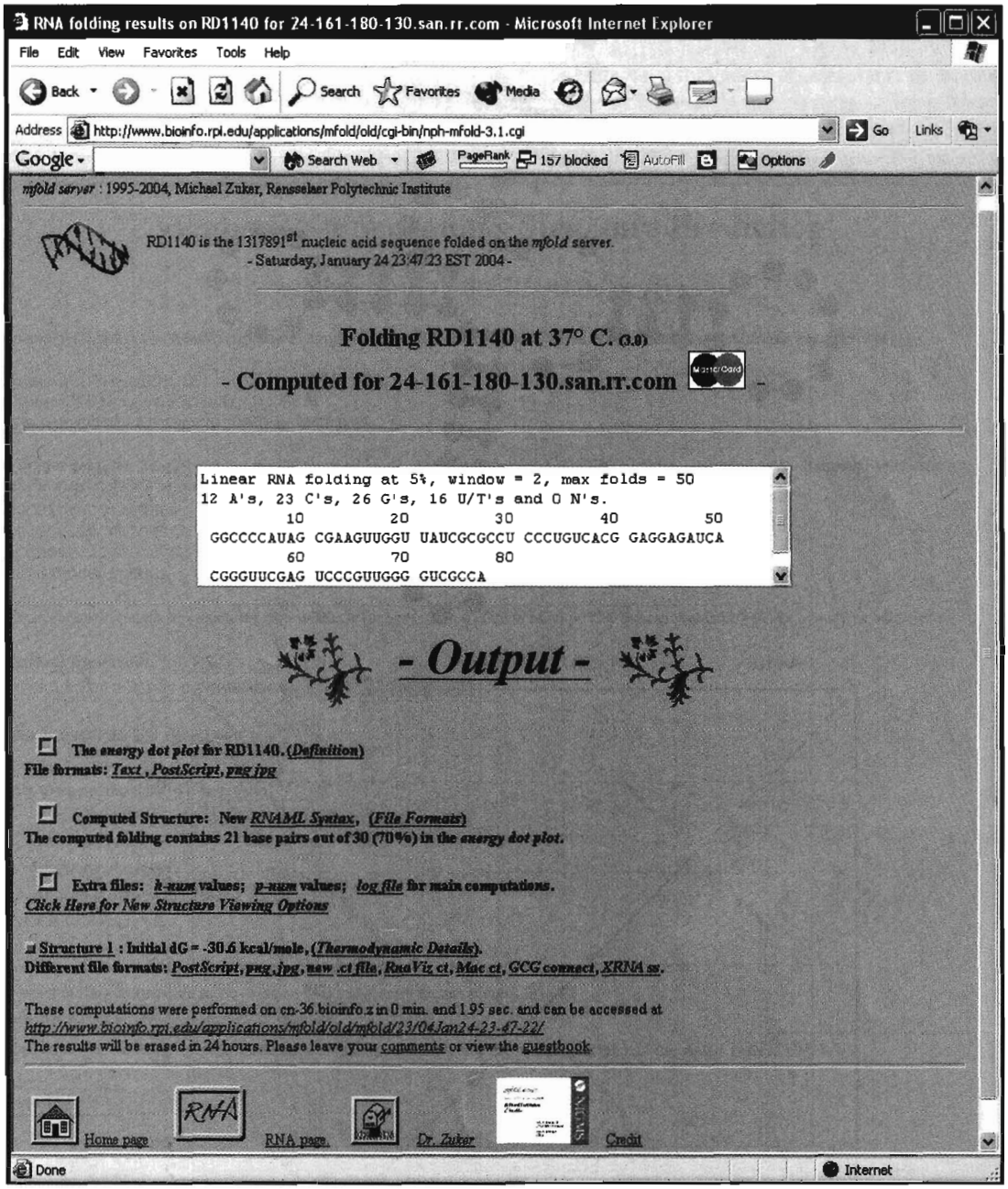

FIGURE 6.7 The output page for the *mfold* server. See main text for details.

and y-axis, and its color indicates the best energy for a structure that contains that pair. The energy dot plot is divided into two triangles. The upper triangle is the energy plot including suboptimal pairs and the lower triangle is the location of base pairs in the predicted minimum free energy structure. The energy dot plot in Figure 6.8 shows that there are alternative base pairs contained in structures with free energies between −29.7 and −30.1 kcal/mol, a separation of less than 0.5 kcal/mol from the lowest free energy structure.

Vienna RNA Package

The Vienna RNA Package can be used to predict RNA secondary structures via either a Web interface or by compilation onto Unix and Linux machines (Hofacker, 2003; Hofacker et al., 1994). It uses a dynamic programming approach and the current set of thermodynamic parameters (Mathews et al., 1999b). The Vienna Package also implements an algorithm that calculates the partition function for RNA folding, which predicts the base

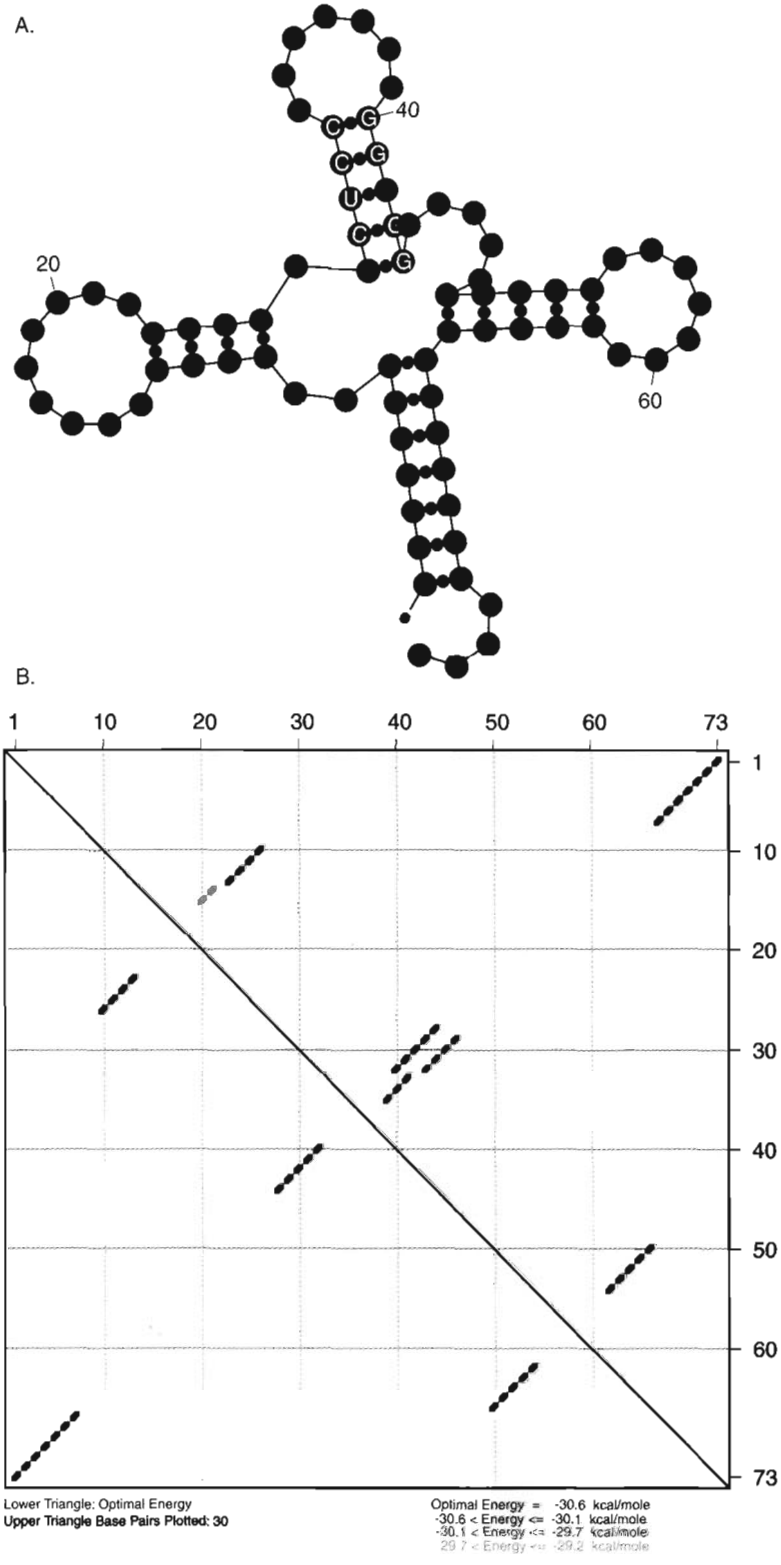

FIGURE 6.8 Sample output from the *mfold* Web server, version 3.1. (a) The secondary structure predicted for the tRNA, RD1140 (Sprinzl et al., 1998). (b) Color annotation based on the energy dot plot.

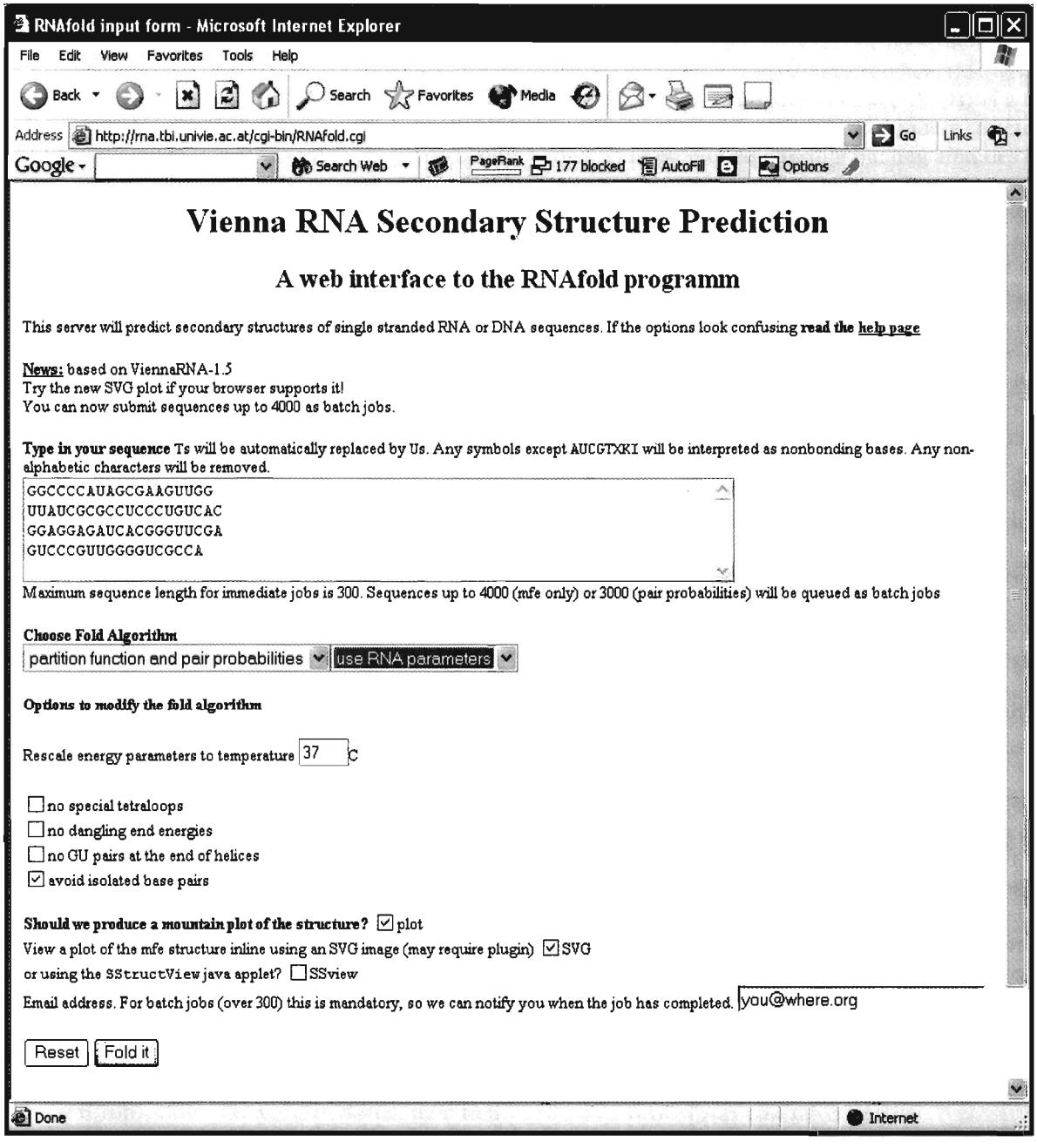

FIGURE 6.9 Vienna Server input form. A discussion of the available options can be found in the main text.

pairing probability for each possible base pair in a sequence. The partition function prediction algorithm, first implemented by McCaskill (1990), is also a dynamic programming algorithm. The calculated base pair probabilities are commonly displayed in a probability dot plot, analogous to the energy dot plots from *mfold*. Additionally, the Vienna Package includes software for the generation of all suboptimal secondary structures within a given energy increment of the lowest free energy structure (Wuchty et al., 1999). The number of secondary structures grows exponentially with increasing size of the energy increment.

Figure 6.9 shows the input form for the Vienna Package Web server. The link to the *help page* can be followed for an explanation of the fields. The sequence is typed or pasted from the clipboard in the box below *Type your sequence*. The tRNA sequence, RD1140 (Sprinzl et al., 1998), is shown in the sequence box. Nonalphabetic characters are ignored automatically by the server. The default fold algorithm is partition function and pair probabilities, although the partition function calculation can be turned off by changing to minimum free energy only. The parameter set is chosen on this form as either use RNA parameters, old RNA parameters, or use

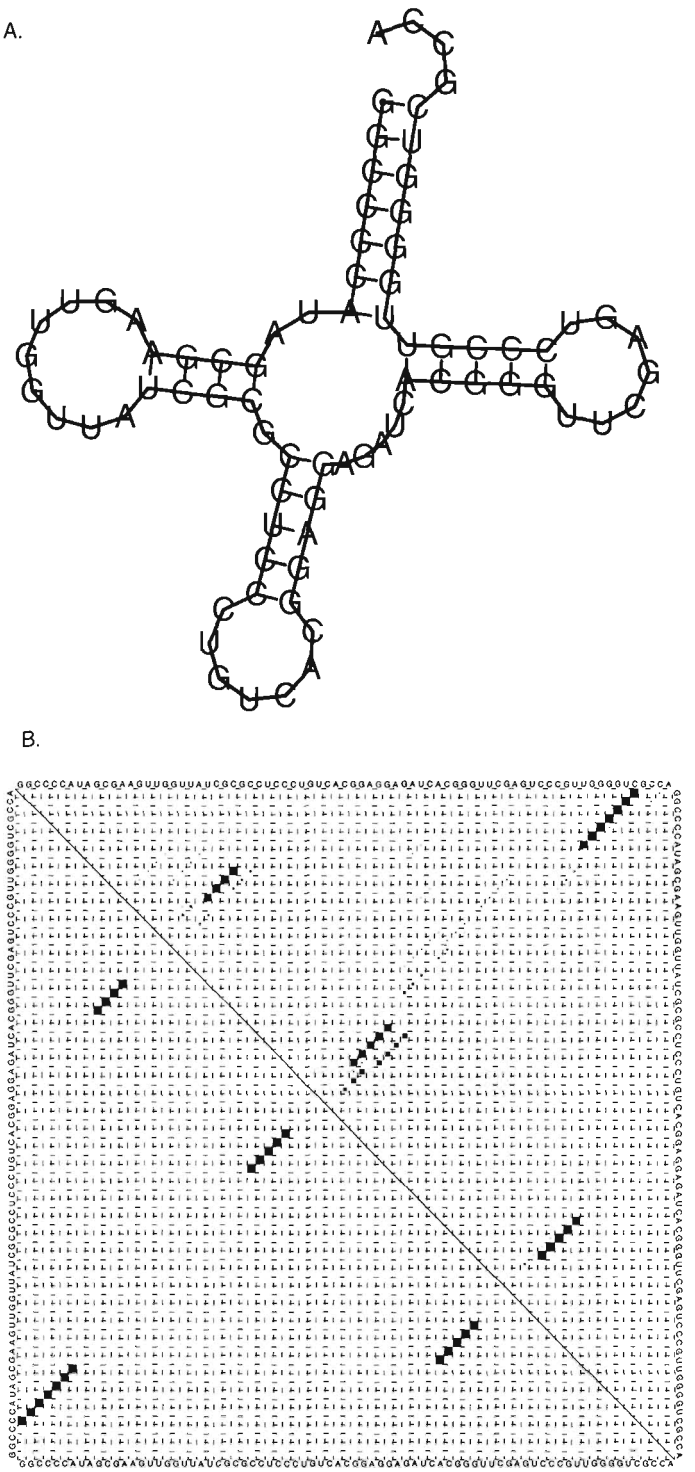

FIGURE 6.10 Sample output from the Vienna Package Web server, version 1.5. (a) The predicted minimum free energy secondary structure for the tRNA RD1140 (Sprinzl et al., 1998). (b) The probability dot plot for the same sequence.

DNA parameters. The old RNA parameters are available (Walter et al., 1994) also, so that previous predictions can be reproduced. The temperature of folding can be changed from the default of 37°C, but for similar temperatures it is recommended that the default be used. Folding at temperatures other than 37°C use an older set of thermodynamic parameters (Jaeger et al., 1989) that are based on fewer experiments than the current set of parameters (Mathews et al., 1999b). The next parameters, *no special tetraloops*, *no dangling end energies*, *no GU pairs at the ends of helices*, and *avoid isolated base pairs*, modify the energy rules. The default is to check *avoid isolated pairs*, and this will reproduce the behavior of the *mfold* server. The other choices can be modified by advanced users. The checkboxes at the bottom of the form (*plot*, *SVG*, and *SSView*) control the output formats. The default options, shown in Figure 6.9, are suitable for most users. Finally, for batch folding (required for sequences longer than 300 nucleotides) the user must enter an E-mail address to receive notification that the calculation is complete. For shorter sequences, an immediate job can be performed without providing an E-mail address. The calculation is started by clicking *Fold it*.

Figure 6.10 demonstrates the output of the Vienna Package Web server, based on the input shown in Figure 6.9. The probability dot plot (Figure 6.10b) is obtained by following the link to PostScript *dot plot* on the output form. The Adobe SVG Viewer, downloaded for free by following the link on the Vienna Package Web server output form, is required to view the predicted structures. Note that the predicted secondary structure (Figure 6.10a) of the RD1140 tRNA sequence (Sprinzl et al., 1998) is identical to that predicted by the *mfold* server (Figure 6.8a). The structure is drawn counterclockwise, with the ends of the sequence at the top of the figure, whereas the *mfold* server draws the structure clockwise, with the ends at the bottom, but the predicted base pairs are identical. The probability dot plot (Figure 6.10b) shows the predicted minimum free energy structure base pairs in the lower triangle. In the upper triangle, the area of a square dot is proportional to the probability of the corresponding base pair, indicated by the nucleotides on the x- and y-axes. The probability dot plot for this sequence also indicates pairs of lower probability competing with those in the stem starting with the pair of nucleotides 28 and 44.

RNAstructure

RNAstructure is a secondary structure prediction dynamic programming algorithm for the Microsoft Windows (Redmond, WA) environment that uses the current set of thermodynamic parameters for RNA secondary structure prediction (Mathews et al., 1999b). Detailed instructions for predicting a secondary structure are available in the online help file and elsewhere (Mathews et al., 2000). OligoWalk (Mathews et al., 1999a) is a component of RNAstructure that uses secondary structure prediction to predict equilibrium binding affinities of complementary DNA or RNA oligonucleotides to an RNA target. OligoWalk considers all $N - L + 1$ fully complementary oligonucleotides of length L to a target of length N. The equilibrium shown in Figure 6.11 is considered by these programs in which a complementary oligomer pairs to the target, but self-structure in both the oligomer and target can reduce the free energy of binding. Oligomers can be either RNA or DNA, where the thermodynamic parameters for DNA oligomers are derived from nearest neighbors for DNA-DNA (SantaLucia, 1998) and DNA-RNA (Sugimoto et al., 1995) base pairing. Predicted free energy parameters for oligomer-target binding correlate with cell-based measures of antisense efficacy (Mathews et al., 1999a; Matveeva et al., 2003). It is likely that this equilibrium also will be important for the design of sequences for RNAi.

Figure 6.12 is a screen shot of the RNAstructure predicted minimum free energy structure for the tRNA sequence RD1140 (Sprinzl et al., 1998), using the default suboptimal structure parameters. This structure is equivalent to that predicted using the programs discussed earlier (Figures 6.8 and 6.10). Figure 6.13 shows a screen shot of the OligoWalk input form for predicting the affinity of complementary oligonucleotides to this sequence. The user clicks the button labeled *CT File* to choose the file that contains the predicted structure of the target. A default output file name, called a report file, is then generated, but the file name can be changed by clicking the *Report File* button. One of three modes is

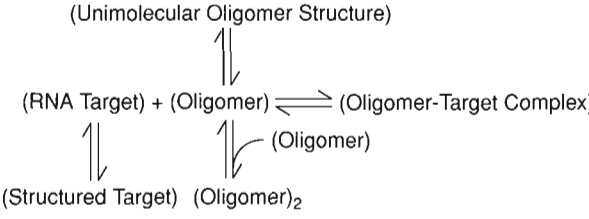

FIGURE 6.11 The equilibrium used by OligoWalk for predicting the affinity of an oligomer to an RNA target. The oligomer binds by Watson-Crick base pairing to the RNA target to form the oligomer-target complex. Competing with this basepairing are unimolecular self-structure in the target, unimolecular self-structure in the oligomer, and bimolecular oligomer self-structure (Mathews et al., 1999a).

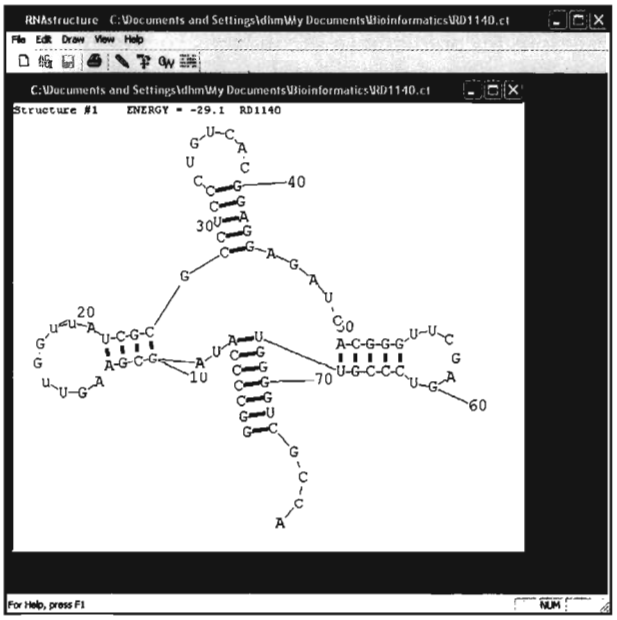

FIGURE 6.12 Screen shot of secondary structure prediction by RNAstructure, version 3.71. This is the predicted minimum free energy structure for the tRNA sequence RD1140 (Sprinzl et al., 1998).

chosen (Mathews et al., 1999a). The *Break Local Structure* mode assumes that the base pairs in the target do not reequilibrate after the oligonucleotide binds and are a suitable default. The user chooses whether to *Include Target Suboptimal Target Structures* in the calculation. For the short tRNA target here, not including suboptimal target structures is suitable. For long targets with a large number of suboptimal structures within a small energy increment from the lowest free energy structure, including suboptimal target structures is preferred because this can help overcome some of the drawbacks of limited prediction accuracy. The user then needs to choose the length of the oligonucleotides, whether the oligomers are DNA or RNA, and the concentration of the oligonucleotides. The default is to look at all complementary oligonucleotides from the first to the last nucleotide. This default can be changes by modifying the `Start` and `Stop` locations. The calculation is started by clicking *Start OligoWalk*.

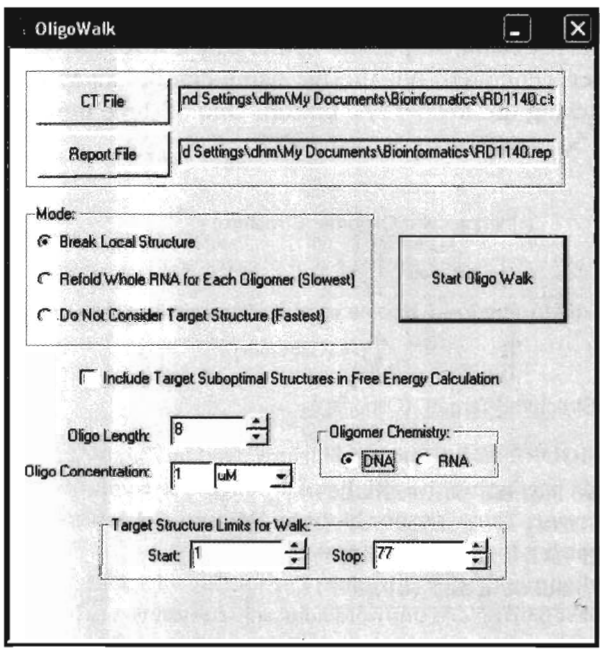

FIGURE 6.13 The input window for an OligoWalk run. The input options are discussed in the main text.

Figure 6.14 shows the output of this calculation in the graphical user interface. The target sequence, the tRNA RD1140, is displayed from 5′ to 3′ horizontally along the center of the window with nucleotides predicted to be base paired in the lowest free energy structure in red. The current oligonucleotide, selectable by the user with the right and left arrow keys or by clicking the < or > buttons, is displayed above from 3′ to 5′. The *Go...* button can be used to jump to a specific oligonucleotide or to jump to the highest affinity oligonucleotide. The oligonucleotide predicted to have the highest affinity, at a $\Delta G^{\circ}_{37overall}$ of −5.7 kcal/mol, is shown on the current display. At the top of the display is cost for opening base pairs in the target (−0.8 kcal/mol), the cost for opening the oligonucleotide bimolecular self structure (−0.6 kcal/mol), and the cost for opening oligonucleotide unimolecular self structure (0 kcal/mol).

Sfold

Sfold is an implementation of the partition function calculation that predicts secondary structures using a stochastic sampling procedure (Ding & Lawrence, 2001; Ding & Lawrence, 1999; Ding & Lawrence, 2003). The sampling procedure guarantees that structures are sampled with true statistical weight. Sfold is available for use through a Web interface.

Sfold has been shown to predict unpaired regions that correlate to regions accessible to antisense oligonucleotide targeting (Ding & Lawrence, 2001). Because the secondary structures are sampled statistically, the fraction of occurrences that a nucleotide is unpaired in a set of sampled structures is the predicted probability for being unpaired.

Figure 6.15 contains sample output from the Sfold Web server for the tRNA sequence RD1140 (Sprinzl et al., 1998). Figure 6.15a shows the predicted most-probable structure, which is the same lowest free energy structure

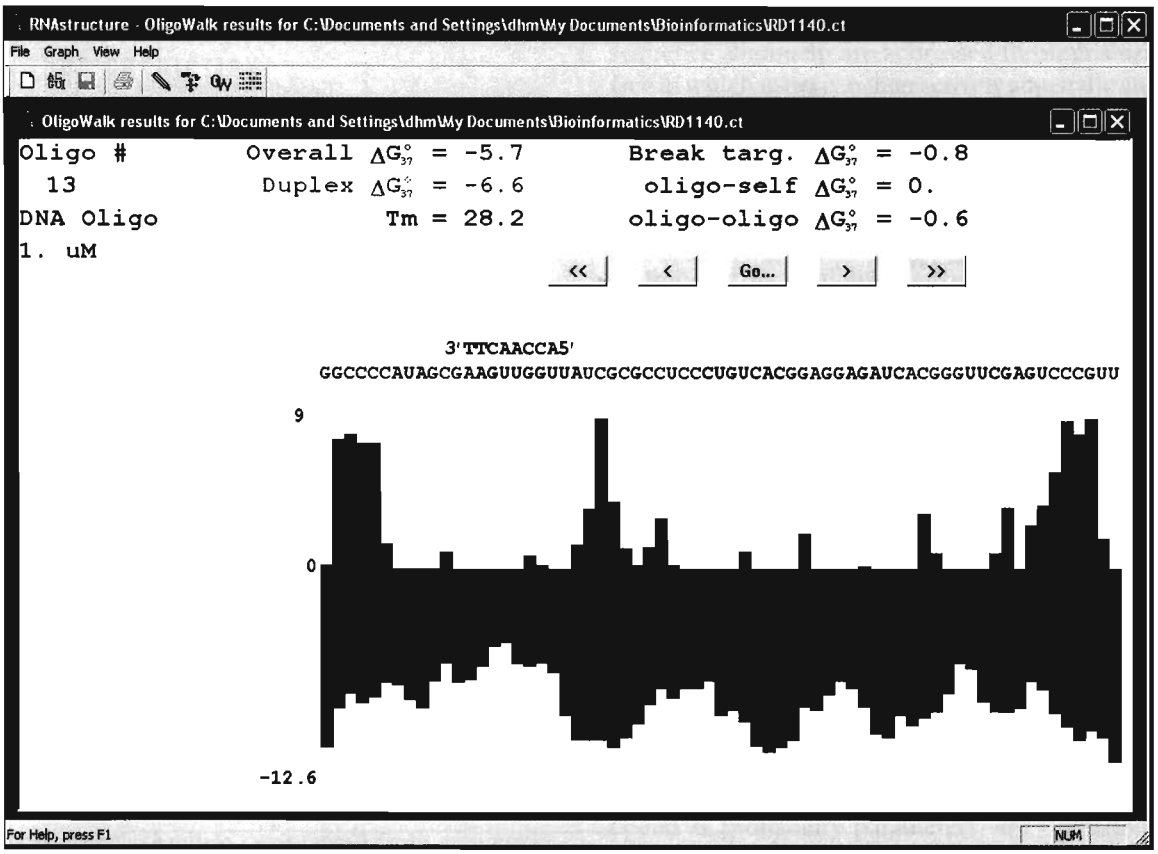

FIGURE 6.14 A screen shot of OligoWalk from RNAstructure, version 3.71. This screen shot shows the affinity predictions of 8-mer DNA oligomers to the target RNA sequence.

predicted by the other programs previously discussed. Figure 6.15b shows the probability of pairing, analogous to the probability dot plot produced by the Vienna Package. The areas of the dots correspond to base pairing probability of the nucleotides from the x-axis and y-axis. Figure 6.15c shows the probability profile for the sequence, showing the probability that a single nucleotide in the RNA is unpaired. Nucleotides that have low probability of being base paired are more suitable targets from a thermodynamic perspective.

COMPARISON OF DYNAMIC PROGRAMMING SECONDARY STRUCTURE METHODS

The software packages listed here (*mfold*, the Vienna Package, RNAstructure, and Sfold) each predict secondary structures and alternative secondary structures, and each uses the current set of free energy parameters assembled by Turner et al. (Mathews et al., 1999b). *Mfold*, the Vienna Package, and PFold are freely available through Web interfaces. Additionally, *mfold* and the Vienna Package are available for compilation on Unix and Linux machines. RNAstructure, however, is a Microsoft Windows program for installation on personal desktop computers. Each package has its own unique features, as described above.

For the example used here, the tRNA sequence of RD1140 (Sprinzl et al., 1998), all of the software packages predicted the same secondary structure. Although all packages are based on the same set of thermodynamic parameters, in general, they do not guarantee identical results. Each program uses a slightly different method for calculating the free energy of multibranch loops. The partition function in the Vienna Package assumes 3′ and 5′ dangling ends at the end of each helix. SFold explicitly checks for 3′ and 5′ dangling ends at the end of each helix, but assumes that a nucleotide will stack preferentially as a 3′ dangling end if both possibilities exist. *Mfold* and RNAstructure explicitly find the optimal stacking of 3′ dangling ends, 5′ dangling ends, or both at the end of each helix in a multibranch loop. Coaxial stacking, the end-to-end stacking of two helices, is included in a second-step calculation that recalculates the free energy of predicted structures, called *efn2*. RNAstructure and *Mfold* differ slightly in the use of

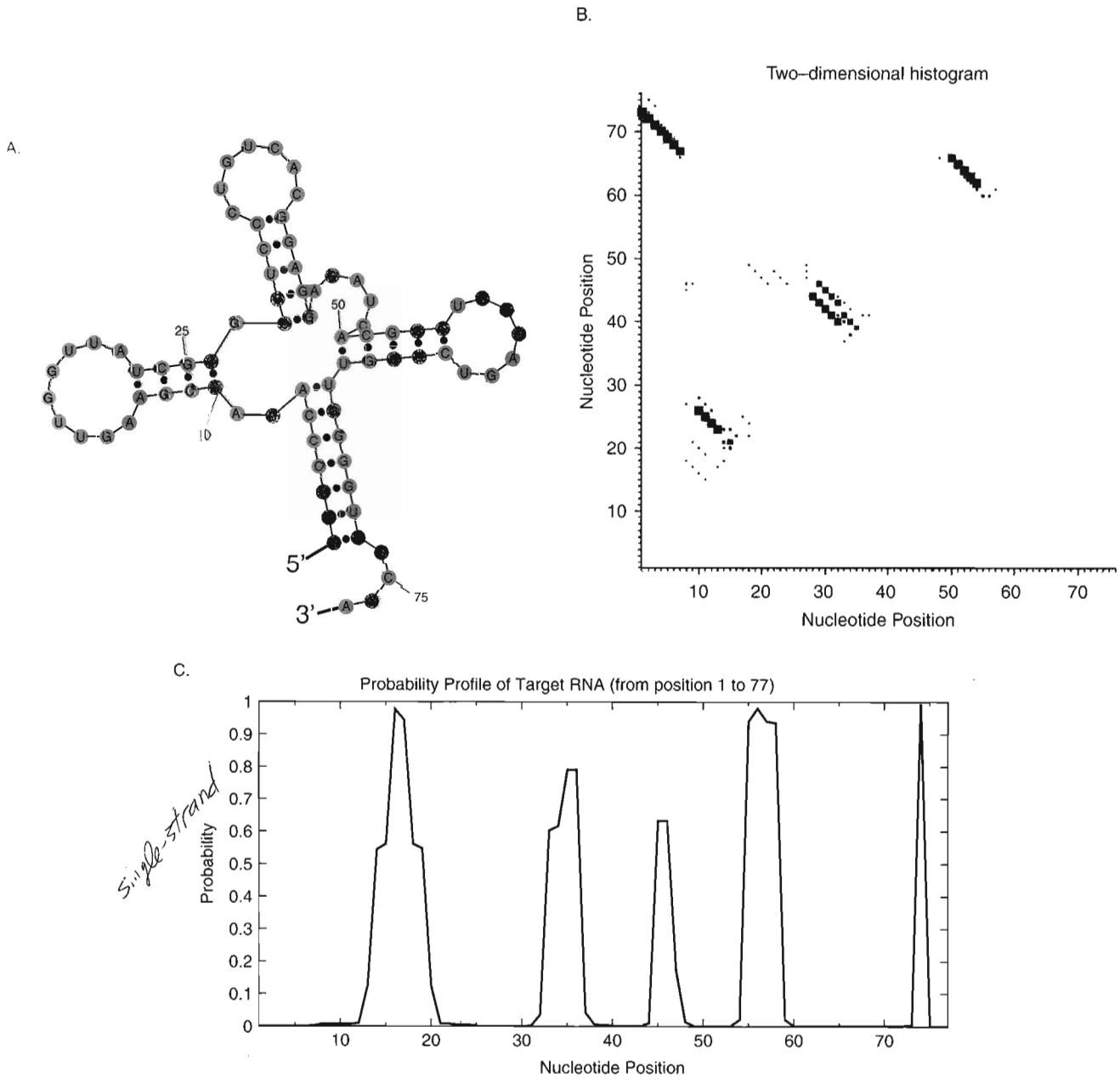

FIGURE 6.15 Sample output from the SFold server. (a) The most probable secondary structure for the RD1140 tRNA sequence. (b) The probability of pairing for all possible pairs with the largest dots indicating the most probable pairs. (c) The probability profile for a nucleotide being single stranded.

efn2, creating subtle differences in the predicted structures. Finally, the free energy minimization algorithm from the Vienna Package, RNAfold, explicitly can include the terminal stacking calculations and a subset of the known coaxial stacking interactions in the dynamic programming algorithm. The partition function calculations use the simplified energy rules because of increased computational overhead as compared with free energy minimization programs. Given that there are many secondary structures within a small energy increment of the predicted minimum free energy structure (Wuchty et al., 1999), these subtle differences can result in different structure predictions. Differing structure predictions are more likely the longer the sequence being studied. No systematic studies have been carried out to examine how crucial each of these terms for multibranch loop stability is for the accuracy of secondary structure prediction.

GENETIC ALGORITHM FOR RNA SECONDARY STRUCTURE PREDICTION

Other computational methods have been explored for RNA secondary structure prediction. For example, a genetic algorithm, which uses random mutations of structure and selection of the most fit solutions, is available in STAR (Gultyaev et al., 1995; Van Batenburg et al., 1995). This algorithm determines fitness based on conformational free energy (Mathews et al., 1999b). The algorithm is executed with the sequence lengthening from 5′ to 3′ end to simulate a pathway of RNA folding. Also, because the algorithm is not based in dynamic programming, it is capable of including pseudoknots explicitly in a computationally reasonable time. However, the drawbacks to simulations like genetic algorithms are that they do not guarantee the optimal solution and that they can provide different results with repeated calculations on the same sequence.

PREDICTING THE SECONDARY STRUCTURE COMMON TO MULTIPLE RNA SEQUENCES

The basis of comparative sequence analysis is the detection of conserved structure, as inferred from sequence differences between species or between sequences discovered by *in vitro* evolution (Pace et al., 1999). The assumption of a conserved secondary structure eliminates from consideration the many possible secondary structures for a single sequence that the ensemble of sequences together cannot adopt. That is, taken together, the multiple sequences constrain the possible secondary structure. These constraints can also be used as auxiliary information in the prediction of secondary structure.

RNA secondary structure prediction algorithms that incorporate information from multiple sequences can be divided between those that are constrained by an initial sequence alignment and those that are not. In general, those methods that are constrained by an initial alignment are not as robust because of the limitations in the alignment, but they are computationally faster.

Algorithms That Are Constrained by an Initial Alignment

Several programs have been developed for finding the secondary structure common to a set of aligned sequences (Hofacker et al., 2002; Juan & Wilson, 1999; Lück et al., 1999; Lück et al., 1996). One approach, called *ConStruct*, uses base pairing probabilities determined by a partition function calculation for each sequence (Lück et al., 1996). These probabilities are then summed according to the alignment to give a consensus probability matrix. The limitations imposed by the sequence alignment are addressed through a user interface in which users can interactively adjust the alignment to improve the consensus probability (Lück et al., 1999; Lück et al., 1996). The source code for ConStruct can be downloaded for compilation.

A second program, called *alifold*, uses a sequence alignment to constrain secondary structure prediction by free energy minimization or to constrain the calculation of the partition function (Hofacker et al., 2002). Additional energy terms are added to the conformational free energy to favor compensating base changes and sequence conservation. This program is available as part of the Vienna Package and through a Web server. Figure 6.16 shows the output for an alifold run for three tRNA sequences; Figure 6.16a shows the consensus secondary structure; and Figure 6.16b shows the probability dot plot. Note that, by including three sequences, the lower-probability base pairs that had competed with one of the stems (e.g., Figure 6.11b) are no longer possible.

A third program for finding a structure common to multiple sequences, called *Pfold*, uses a stochastic context-free grammar (Knudsen & Hein, 1999). The grammar defines rules for emitting a random sequence together with a secondary structure. These rules, encoded as probability parameters, are estimated from a sequence alignment and known, common secondary structures of a number of tRNAs and large ribosomal subunit (LSU) rRNAs. These sequences and structures are referred to as the *training set*. A given sequence is folded using a dynamic programming algorithm that determines a structure with a maximum probability of being emitted by the stochastic context-free grammar. Pfold is available through a Web interface, and sample output for three tRNA sequences is shown in Figure 6.17. The same consensus structure is found as with the alifold server (Figure 6.16).

Algorithms That Are Not Constrained by the Initial Alignment

A genetic algorithm has been developed for finding an alignment and common secondary structure for multiple sequences (Chen et al., 2000). This program makes random mutations on S sequences to make a set of m structures. Alternately, the free energy of conformations and the similarity of conformations are used as fitness criteria for selecting structures for future rounds of mutation and selection. Overall, the algorithm scales as $O(n^2m^2S^2)$, where n is the maximum number of stems allowed in a structure. The authors looked at test cases drawn from tRNA, 5S rRNA, and *rev* response elements of human immunodeficiency virus (HIV) and simian immunodeficiency virus (SIV) (Chen et al., 2000).

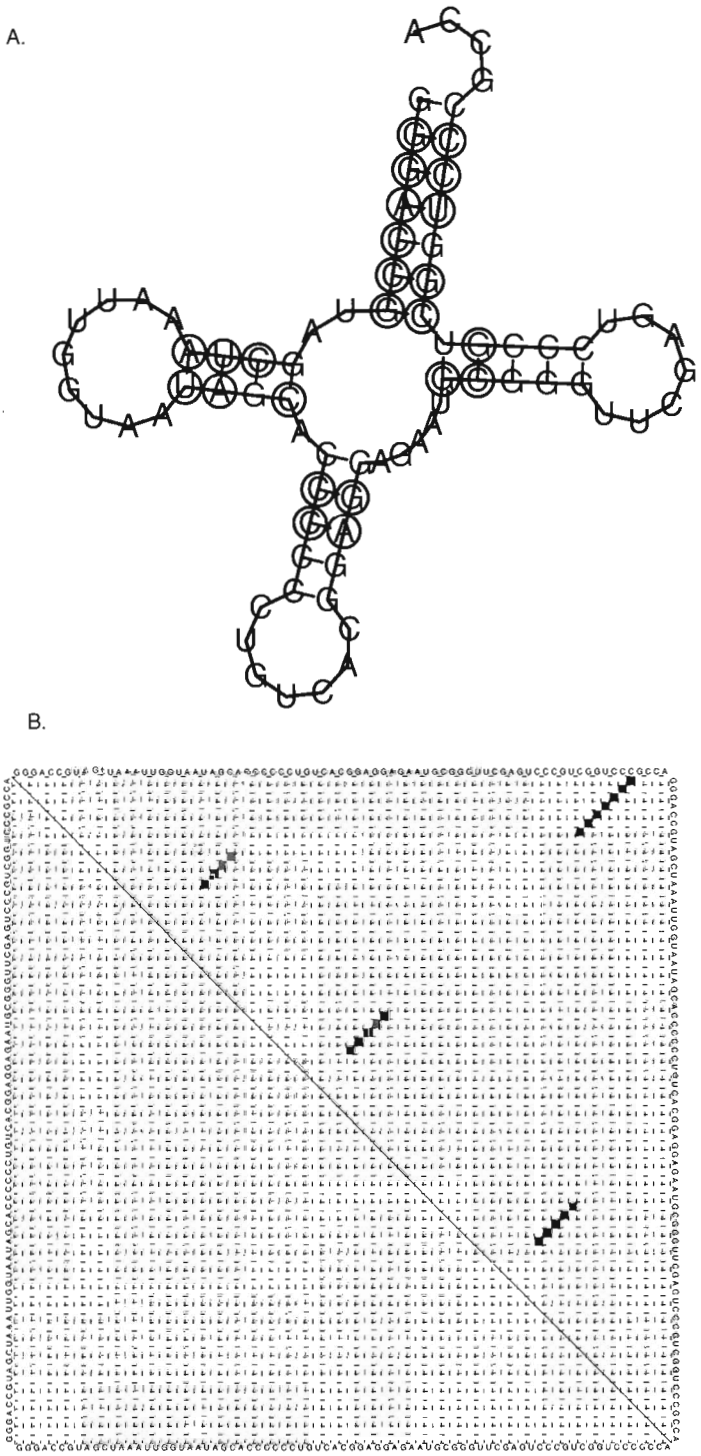

FIGURE 6.16 Sample output from the alifold server. The default parameters were used with three tRNA sequences, RD0260, RD1140, and RD2640, with the alignment taken from the Sprinzl database (Sprinzl et al., 1998). (a) The consensus secondary structure. Circled nucleotides are positions of mutation. (b) The probability dot plot in which dots are color coded to indicate base pair types. Red pairs are conserved, green pairs are positions of more than one compensating base changes, and ochre pairs are positions with a single compensating base change or a neutral change, for example, GC to GU change.

FIGURE 6.17 Sample output from the Pfold server. The structures of the three tRNA sequences are shown with bracket notation above each sequence. The parentheses indicate paired nucleotides and the direction of pairing. Every nucleotide with a "(" above is base paired to a downstream nucleotide with a ")" above. The structure predicted for the first sequence, RD0260, is identical to the structure predicted by other methods (e.g., Figure 6.10). Note that base pairing confidences are reported under the structures for each homologous base pair. The input alignment was taken from the Sprinzl database (Sprinzl et al., 1998).

Dynamic programming can be used to predict simultaneously the sequence alignment and common secondary structure for multiple sequences (Sankoff, 1985). In general, this approach is $O(N_1^3 N_2^3 N_3^3 \ldots)$ in time, where N_1 is the length of the first sequence, N_2 is the length of the second sequences, and so on, making it computationally impractical. Two computer programs are available that use dynamic programming, limited to two sequences at most. The first, called FOLDALIGN, finds the local alignment and common structure of two sequences, using a simple scoring scheme (Gorodkin et al., 1997). This scoring scheme favors base pairs, sequence conservation, and compensating base changes. It constructs a multiple sequence alignment from pairwise comparisons using a method similar to that used by ClustalW for building alignments based on sequence matching (Thompson et al., 1994). The algorithm is $O(L^4)$ in time, where L is the maximum motif size. This scaling is achieved by not allowing multibranch loops or pseudoknots. Because multibranch loops are not included, FOLDALIGN is designed to be a screening tool for finding common helices, that is, a first step in comparative sequence analysis.

The second dynamic programming algorithm for simultaneous secondary structure and alignment prediction of two sequences is Dynalign (Mathews & Turner, 2002). It minimizes the sum of the conformational free energy parameters for both sequences, using nearest-neighbor parameters and a term that penalizes the insertion of gaps into the sequence alignment. The gap insertion penalties were calibrated to folding free energies by optimizing the accuracy of pairwise structure predictions for a set of 5S rRNA sequences. Because there are no terms in the optimization for matching sequence in the alignment, Dynalign does not require any sequence similarity in two sequences. Structural bifurcations (i.e., multibranch loops) are allowed. Algorithm scaling is improved by restricting the possible sequence alignments by limiting the maximum separation between nucleotides in the alignment with a parameter M. This results in an algorithm that is $O(N^3 M^3)$ in time and $O(N^2 M^2)$ in storage, where N is the number of nucleotides in the shorter sequence. Dynalign is available as part of the RNAstructure package for Microsoft Windows or as C++ code for local compilation. An application of Dynalign is one of the worked examples at the end of the chapter.

COMPARISON OF METHODS

No single algorithm is yet available that can replace comparative sequence analysis. Each algorithm provides results that are useful for constructing a secondary

structure model for multiple sequences. Dynalign can be helpful for aligning sequences that are too dissimilar to be aligned by primary sequence without referring to secondary structure (Mathews & Turner, 2002). Alternatively, FOLDALIGN can be used for sequences too long for Dynalign (Gorodkin et al., 1997). The methods for finding secondary structure in multiple sequence alignments are best used as screening tools to find common helices, which can be used to anchor portions of a sequence alignment when making revisions for further rounds of analysis. The ConStruct tool provides one such convenient user interface for doing the alignment revisions (Lück et al., 1999).

INTERACTIVELY DRAWING RNA SECONDARY STRUCTURES

Software packages for secondary structure prediction come with programs to display predicted structures automatically. These diagrams usually are acceptable for looking at results, but generally are not of high enough quality for publication without substantial revision.

Three software packages are available for editing diagrams of RNA secondary structures. The first, xrna, is available from the University of California at Santa Cruz RNA Center. It is written in Java, and therefore should function on any platform that supports the current Java implementation. Figure 6.2 was drawn interactively with xrna on a computer using Microsoft Windows. The second program, RnaViz, is available as executable programs for Windows and Linux (De Rijk et al., 2003). The third program, sir_graph, by D. Stewart and M. Zuker, is written in C and is available, together with source code, for Unix, Linux, Mac OS X (Darwin-Fink-X11 and Darwin-Panther) and Windows (Cygwin and MingW).

PREDICTING RNA TERTIARY STRUCTURE

Although there are many automated methods for accurate RNA secondary structure prediction, tertiary structure prediction remains largely a craft that requires user input and insight. One reason for this has been the relative lack of RNA three-dimensional structures compared with secondary structures. Two- and three-dimensional nuclear magnetic resonance (NMR) methods have provided a wealth of information on the solution structure of small loops, but are limited to systems of approximately 50 nucleotides without selective nucleotide labeling. The tRNA crystal structure of yeast phenylalanine tRNA was solved more than 25 years ago (Kim et al., 1974), but few large, nonhelical crystals of RNA were solved subsequently, until more recent technological breakthroughs culminated in the publication of high-quality crystal structures of the ribosome (Ban et al., 2000; Schluenzen et al., 2000; Wimberly et al., 2000).

Several distinct computational approaches have been used successfully to model RNA tertiary structures. The first is an extension comparative sequence analysis to predict sites of tertiary contacts (Massire et al., 1998; Michel et al., 2000). This approach has its origins in the work by Levitt (1969) on tRNA sequences. In that work, three tertiary contacts, of which two were proven later to be correct, were inferred from an alignment of tRNA sequences (Levitt, 1969; Michel et al., 2000). More recently, Michel and Westhof (1990) modeled the catalytic core of the group I self-splicing intron using high-quality sequence alignment of 86 sequences with well-established secondary structure as the starting point. Nucleotide columns in the alignment not involved in canonical pairing, found to co-vary with statistical significance, are inferred to be involved in tertiary contacts. With a set of tertiary contacts, a model of the catalytic core of the *Tetrahymena* sequence was built (Michel & Westhof, 1990). A model was also built of the tertraloop–tertraloop receptor motif, with an overall orientation that was supported by a later crystal structure (Pley et al., 1994). However, most atomic details of the interaction, such as the locations of hydrogen bonds, were incorrect, suggesting that such models are coarse grained. The Westhof group makes available a computer program called MANIP, for the SGI IRIX operating system, for user assembly of structure motifs into structures (Massire & Westhof, 1998).

A second approach to tertiary structure modeling uses experimentally derived data to constrain model building with a program called MC-SYM (Major et al., 1993; Major et al., 1991). Models are constructed automatically by the stepwise assembly of nucleotides in conformations collected from known structures. Each possible model is stored until it is shown to contradict a constraint, based on experimental data or comparative analysis. The variations between all compatible models can suggest how well-determined the model is with the data used. This approach has been used to construct a model of the hairpin ribozyme using data on secondary structure, hydroxyl radical footprinting, photoaffinity cross-linking, and disulfide cross-linking (Pinard et al., 1999). A later crystal structure verified the existence of a predicted long range GC pair, although a predicted base triple, involving an A at that pair, was not observed (Rupert & Ferré-D'Amaré, 2001). Again, this suggests that the model is coarse grained, that is, many gross features are predicted correctly, although some atomic-level

interactions are incorrect. MC-SYM is available for SGI IRIX and Linux.

A third approach is homology model building using a sequence alignment and a reference tertiary structure. Homology modeling is a commonly used method for predicting the structure of proteins (see Chapter 9), but has not been a method available to the RNA community because of the lack of large tertiary structures. With the publication of the crystal structure of the 30S ribosomal subunit (Schluenzen et al., 2000; Wimberly et al., 2000), a template for homology modeling of the 16S rRNA tertiary structure became available. Tung et al. (2002) constructed a model of the 16S rRNA from *E. coli* using the crystal structure of the *Thermus thermophilus* sequence as a template. For regions of the sequence alignment that have no insertions or deletions, a direct substitution of the nucleotides was used. For the more variable regions, entire motifs, borrowed from other regions of the template structure, were inserted into the model structure. The model was found to correlate reasonably with the available cryo-EM map of the *E. coli* structure (Gabashvili et al., 2000). Similarly, a homology model of the tRNA-like domain of the tmRNA was constructed using the $tRNA^{Phe}$ (Hingerty et al., 1978) and $tRNA^{Asp}$ (Westhof et al., 1988) crystal structures as a reference (Stagg et al., 2001).

Another method applied to RNA tertiary structure modeling is low-resolution molecular mechanics calculations. The Harvey group has developed a reduced representation molecular mechanics software package, called *vammp*, that was used to model the 16 S ribosomal RNA in the context of the small ribosomal subunit (Malhotra & Harvey, 1994). The modeling was started with a representation of the RNA in which one pseudoatom was used for each helix. A random walk was performed to provide a variety of starting structures, followed by simulated annealing and energy minimization. Several possible models were retained for further refinement, starting with simulated annealing and energy minimization on a representation in which each helix was represented with five pseudoatoms. Finally, simulated annealing and energy minimization was performed with each nucleotide represented as a single pseudoatom. Constraints, derived from cross-linking and chemical modification data were modeled as pseudobonds between pseudoatoms. Each ribosomal protein in the small subunit was also considered in the calculation as a single spherical pseudoatom with very soft excluded volume constraints, allowing limited nonspherical behavior. The last step of modeling was to construct a consensus structure from the seven individual models. The regions with large structural differences between these seven models were assumed to be less well defined in the final consensus model. More recently, yammp was used to help model the tertiary structure of RNase P, constrained with cross-linking data (Chen et al., 1998).

FUTURE OF TERTIARY STRUCTURE PREDICTION

New data are becoming available with which to understand the forces that drive tertiary structure formation in RNA. At the coarse-grained level, recent studies categorized noncanonical pairs based on geometry, providing information needed for improved homology modeling and comparative sequence analysis modeling (Leontis et al., 2002). Newly solved crystal and NMR structures are providing atomic resolution models from which to study RNA structure by example (Ferre-D'Amare & Doudna, 1999; Major & Griffey, 2001; Moore, 2001; Zídek et al., 2001). Computational studies are providing an understanding of the interaction of RNA with metal ions and solvent (Auffinger et al., 2003; Auffinger & Westhof, 2000). New computational methods are also being developed that speed atomic level calculations and improve their accuracy (Kollman et al., 2000; Tsui & Case, 2001).

SUMMARY

RNA secondary structure can be predicted by free energy minimization using dynamic programming, with an average predictive accuracy of 73% for a single sequence (Mathews et al., 1999b). Several software packages, including *mfold* and the Vienna Package, are available to do this calculation (Hofacker, 2003; Zuker, 2003). These packages include algorithms that can help in the identification of base pairs that are not well determined. Secondary structure prediction has been extended to predict regions accessible to oligonucleotide binding in the programs OligoWalk and SFold (Ding & Lawrence, 2001; Mathews et al., 1999a).

Several methods are available to constrain secondary structure prediction using multiple sequences. These are divided among algorithms that are limited to an initial sequence alignment and those that are not limited to an initial alignment. ConStruct, alifold, and PFold all predict a secondary structure common to a set of aligned sequences (Hofacker et al., 2002; Knudsen & Hein, 1999; Lück et al., 1999). Dynalign, FOLDALIGN, and a genetic algorithm are capable of simultaneously predicting a common structure and sequence alignment (Chen et al., 2000; Gorodkin et al., 1997; Mathews & Turner, 2002).

RNA tertiary structure prediction requires user skill and insight. The currently available methods build

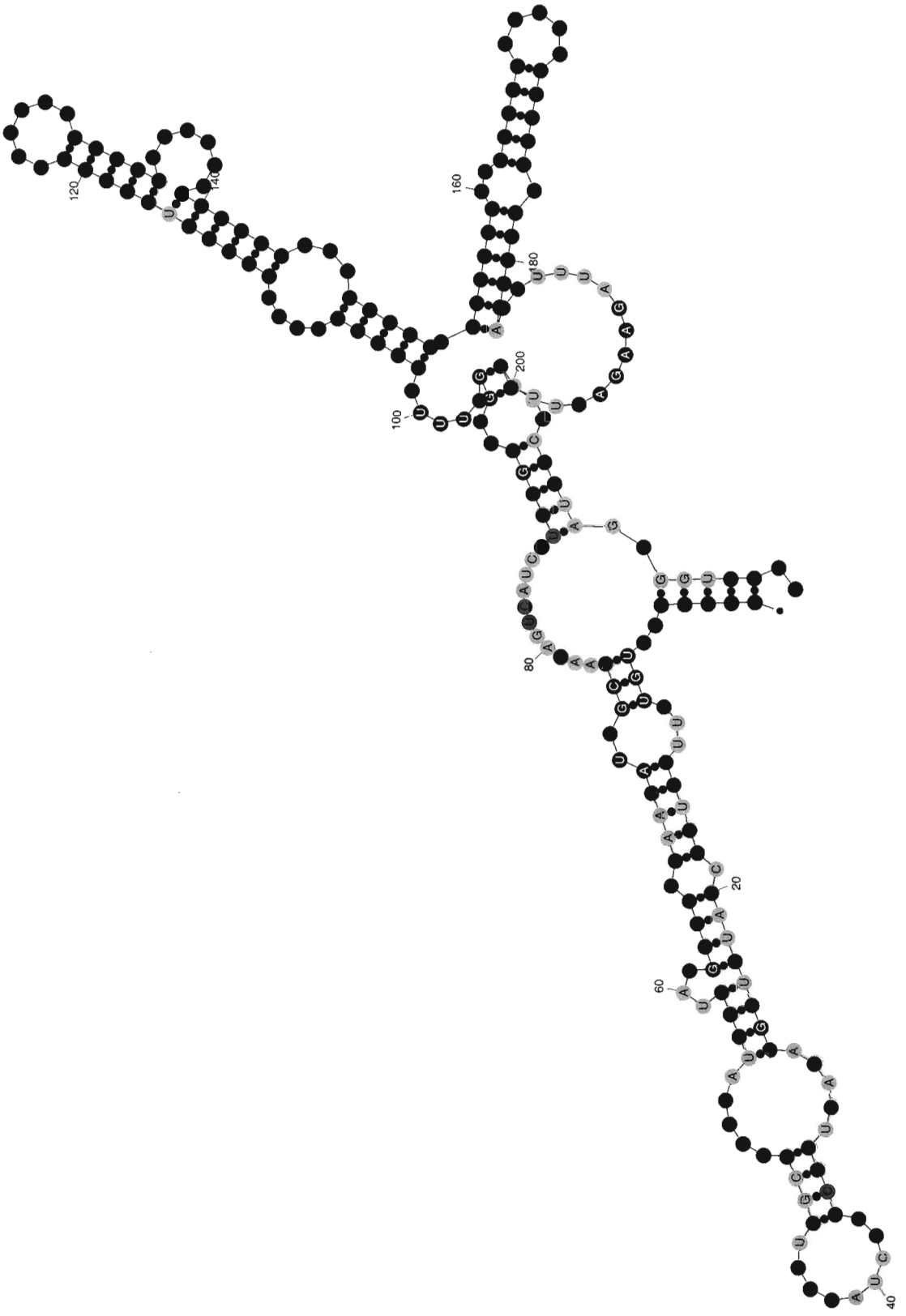

FIGURE 6.18 The predicted, color-annotated structure of the *D. Sucinea* R2 element using the *mfold* server. See Worked Example for details.

coarse-grained structures that can provide an overall sense of the structure, although atomic-scale interactions can be incorrect. Many new experimental and computational results promise to provide insight into the forces that drive tertiary structure formation which should translate to more accurate tertiary structure models.

WORKED EXAMPLE

Two worked examples are presented. The first is the prediction of an RNA secondary structure with color-annotation of "well-definedness" using the *mfold* server. The second example presents the simultaneous prediction of secondary structure and sequence alignment for two sequences using Dynalign.

Mfold Server and Color Annotation

To demonstrate the utility of color annotation on the *mfold* server, predict the secondary structure for the *Drosophila sucinea* R2 3′ UTR as shown in Figure 6.2. R2 elements are a class of retrotransposons that are found in most arthropods (Eickbush, 2002). During retrotransposition, the 3′ UTR of the message RNA is specifically recognized by the reverse transcriptase during target-primed reverse transcription (Luan & Eickbush, 1995; Luan et al., 1993). The secondary structure of the 3′ UTR was predicted for *Drosophila* with comparative sequence analysis of 10 sequences (Mathews et al., 1997). The sequence of the R2 element from *D. sucinea*, which can adopt the comparative analysis structure, was later determined (Lathe & Eickbush, 1997). This sequence has been chosen for this example because it has a known secondary structure and the prediction of this secondary structure by free energy minimization is less accurate than average, so that the usefulness of color annotation is demonstrated (Zuker & Jacobson, 1995; Zuker & Jacobson, 1998).

Download the *D. sucinea* R2 3′ UTR sequence from the Book's Web site. Access the *mfold* Web server and paste the *sucinea* R2 element sequence into the large field on the server Web site for the input sequence. Scroll to the bottom of the Web page, to the section marked `Choose color annotation`. Select the button after `p-num` to choose a color annotation that reflects how well determined base pairs are. Keep the default settings for all other fields. Note, however, that there are links to a help page with an explanation of each user definable setting.

Click the *Fold RNA* button at the bottom of the form. This sequence is short enough that the default immediate job can be performed, so the Web browser will move quickly to the results page. The results remain available on the server for 24 hours. Note that the energy dot plot can be viewed by following a hyperlink at the top of the page. Furthermore, a zip or tar file can be downloaded that contains all the predicted structures. On the results page, view the first individual structure by clicking jpg under `Structure 1`. The jpeg format can be displayed by every graphical Web browser.

Figure 6.18 shows the predicted structure for the *D. sucinea* R2 element, including the p-num color annotation. Five of the predicted helices are identical to helices in the structure based on comparative sequence analysis (Figure 6.2). A sixth helix is predicted that is consistent with, but not included in, the comparative sequence analysis structure. These helices are all between nucleotides U88 and A207. These correctly predicted helices are largely composed of base pairs in which most nucleotides are annotated in red, indicating that there are few competing suboptimal pairs to these base pairs (Zuker & Jacobson, 1998). Most of the remainder of the paired nucleotides, which are not correctly predicted in the lowest free energy structure, are annotated in green, purple, and blue. These colors indicate that there are competing base pairs to these pairs within a small energy increment. The color annotation expresses a measure of confidence in the base pairs where, in this case, 92.3% of base pairs in which both nucleotides are annotated in red are correct. In total, only 54.2% of the predicted pairs are correct.

Dynalign

To demonstrate the usefulness of Dynalign, predict the secondary structures common to the two tRNA sequences RD0260 and RD1140. Download and install RNAstructure on a personal computer using Microsoft Windows. (Alternatively, a text interface version of Dynalign can be used by downloading and compiling onto any Unix or Linux machine with a C++ compiler.) Download the sequence files from the Book's Web site. The sequence file format used by RNAstructure is illustrated by these files. There must be at least one line beginning with a semicolon for comments. The next line must contain a title for the sequence. The following lines contain the sequence, ignoring white space and terminated with a "1." Lower-case nucleotides are forced single stranded.

Start RNAstructure and choose `File|Dynalign` from the menu. Figure 6.19 shows a screen shot of the Dynalign program. Click the `Sequence File 1` button and select the RD0260.seq file with the open file dialog box. Then Click the `Sequence File 2` button and select the RD1140.seq file. The remainder of the fields will fill with default values as shown in Figure 6.19. The output will be saved in three files, `CT File 1`, `CT File 2`, and `Alignment File`. The ct files save the base pairing information and the alignment file is a plain text file with the sequence alignment.

Click the start button to begin the calculation, which will take approximately 6 minutes on a 3.06-GHz Pentium 4 computer. The program then displays the common structure for each sequence in its own window. Click on the window with RD0260 drawn, as illustrated in Figure 6.20. This structure contains all of the correct pairs, as determined by comparative sequence analysis, as does the RD1140 structure. Without the constraints of a second sequence, RD0260 is a tRNA sequence with a poorly

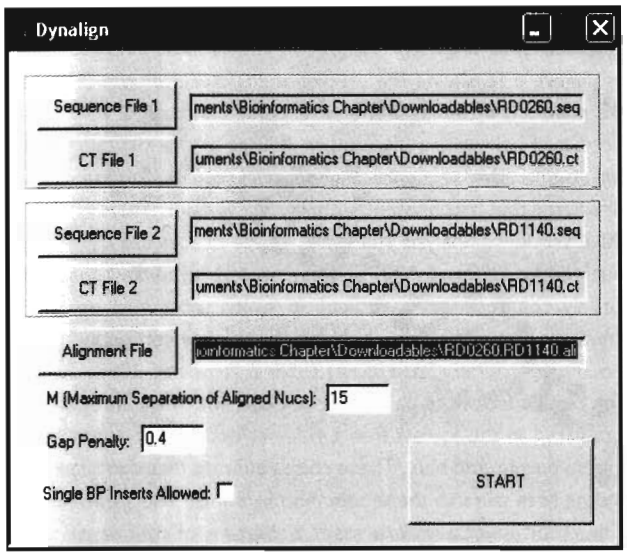

FIGURE 6.19 A screen shot of the Dynalign input form as seen using RNAstructure on Microsoft Windows. See Worked Example for details.

predicted structure. Figure 6.21 shows a screen shot from RNAstructure with the predicted minimum free energy structure for RD0260 when it is predicted alone. This example demonstrates that Dynalign can provide improved accuracy of secondary structure prediction when a second sequence is included to constrain the possible structures.

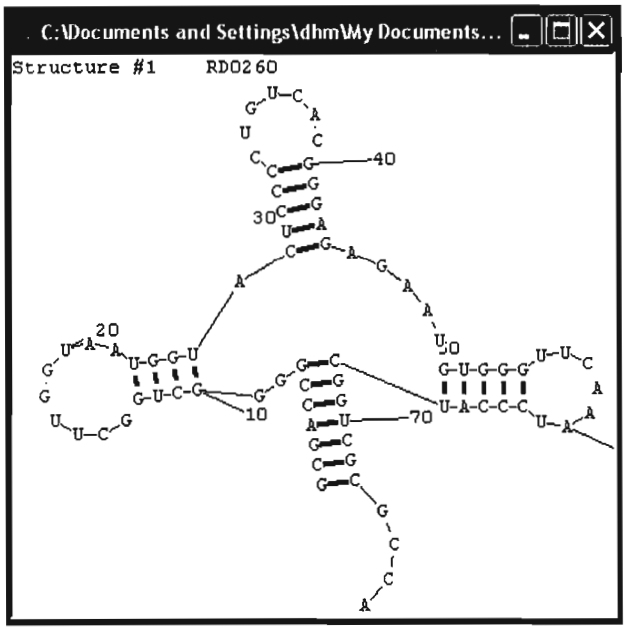

FIGURE 6.20 A screen shot of the RD0260 structure as predicted by Dynalign, in RNAstructure version 3.71. RD1140 was used as the second sequence in the calculation. See Worked Example for details.

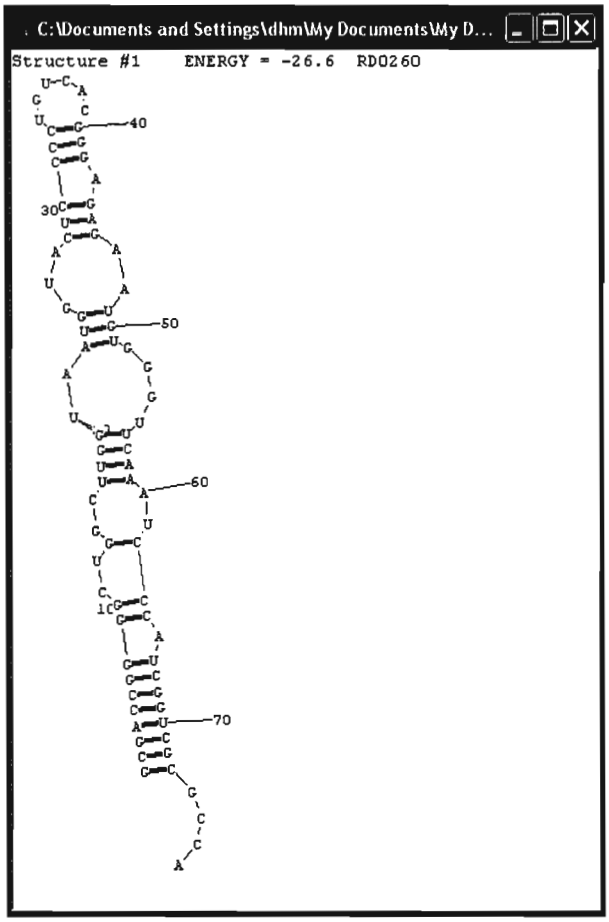

FIGURE 6.21 A screen shot of the RD0260 structure as predicted by free energy minimization of the single sequence in RNAstructure. This tRNA sequence was chosen as an example because its secondary structure is poorly predicted without the constraints provided by a second sequence. See Worked Example for details.

PROBLEM SET

The Book's Web site has a set of four homologous sequences. Predict the secondary structure for each of the sequences using the same program. Determine the consensus secondary structure for these four sequences.

INTERNET RESOURCES

Secondary Structure Drawing Programs

Sir_graph	http://www.bioinfo.rpi.edu/applications/mfold/export/
RnaViz	http://rrna.uia.ac.be/rnaviz/
XRNA	http://rna.ucsc.edu/rnacenter/xrna/xrna_intro.html

Secondary Structure Prediction Programs for a Single Sequence

mfold server	http://www.bioinfo.rpi.edu/applications/mfold
PKNOTS	http://www.genetics.wustl.edu/eddy/software/
RNAstructure	http://rna.chem.rochester.edu/RNAstructure.html
SFold Server	http://www.bioinfo.rpi.edu/applications/sfold/
STAR	http://wwwbio.leidenuniv.nl/~Batenburg/STROrder.html
Vienna RNA Package	http://www.tbi.univie.ac.at/~ivo/RNA/

Secondary Structure Prediction Programs for a Multiple Sequence

Construct	http://www.biophys.uni-duesseldorf.de/local/ConStruct/ConStruct.html
FOLDALIGN server	http://www.bioinf.au.dk/FOLDALIGN/
FOLDALIGN server mirror	http://bifrost.wustl.edu/FOLDALIGN/
Genetic Algorithm	ftp://ftp.ncifcrf.gov
PFold server	http://www.daimi.au.dk/~compbio/rnafold/

Tertiary Structure Prediction Software

MANIP	http://www-ibmc.u-strasbg.fr/upr9002/westhof/
MC-SYM	http://www-lbit.iro.umontreal.ca/mcsym/
Yammp	http://rumour.biology.gatech.edu/Programs/YammpWeb/default.html

FURTHER READING

DURBIN, R., EDDY, S., KROGH, A., AND MITCHISON, G. (1998). *Biological Sequence Analysis. Probablistic Models of Proteins and Nucleic Acids* (Cambridge University Press, New York). An excellent primer on probabilistic models for sequence analysis, including hidden Markov models and stochastic context-free grammars.

TURNER, D. H. (2000). Conformational changes. In *Nucleic Acids* (Bloomfield, Crothers, and Tinoco, eds.), (University Science Books, Sausalito, CA), p. 259–334. A thorough review of the free energy nearest neighbor parameters for RNA secondary structure.

REFERENCES

AUFFINGER, P., BEIELECKI, L., AND WESTHOF, E. (2003). The Mg^{2+} binding sites of the 5S rRNA loop E motif as investigated by molecular dynamics simulations. *Chem. Biol.* 10, 551–561.

AUFFINGER, P., AND WESTHOF, E. (2000). Water and ions binding around RNA and DNA (C,G) oligomers. *J. Mol. Biol.* 300, 1113–1131.

BAN, N., NISSEN, P., HANSEN, J., MOORE, P. B., AND STEITZ, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* 289, 905–920.

BITTKER, J., PHILLIPS, K., AND LIU, D. (2002). Recent advances in the *in vitro* evolution of nucleic acids. *Curr. Opin. Chem. Biol.* 6, 367–374.

BROWN, J. W. (1999). The ribonuclease P database. *Nucl. Acids Res.* 27, 314.

CANNONE, J. J., SUBRAMANIAN, S., SCHNARE, M. N., COLLETT, J. R., D'SOUZA, L. M., DU, Y., FENG, B., LIN, N., MADABUSI, L.V., MULLER, K.M., et al. (2002). The comparative RNA Web (CRW) site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* 3, Epub 2002 Jan 17.

CATE, J. H., GOODING, A. R., PODELL, E., ZHOU, K., GOLDEN, B. L., KUNDROT, C. E., CECH, T. R., AND DOUDNA, J. A. (1996). Crystal structure of a Group I ribozyme domain: principles of RNA packing. *Science* 273, 1678–1685.

CHEN, J., LE, S., AND MAIZEL, J. V. (2000). Prediction of common secondary structures of RNAs: a genetic algorithm approach. *Nucl. Acids Res.* 28, 991–999.

CHEN, J., NOLAN, J. M., HARRIS, M. E., AND PACE, N. R. (1998). Comparative photocross-linking analysis of the tertiary structures of *Escherichia coli* and *Bacillus subtilis* RNase P RNAs. *EMBO J.* 17, 1515–1525.

CULLEN, B. R. (2002). RNA interference: Antiviral defense and genetic tool. *Nat. Immunol.* 3, 597–599.

DE RIJK, P., WUYTS, J., AND DE WACHTER, R. (2003). RnaViz2: an improved representation of RNA secondary structure. *Bioinformatics* 19, 299–300.

DIAS, N., AND STEIN, C. A. (2002). Antisense oligonucleotides: basic concepts and mechanisms. *Mol. Cancer Ther.* 1, 347–355.

DING, Y., AND LAWRENCE, C. (2001). Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. *Nucl. Acids Res.* 29, 1034–1046.

DING, Y., AND LAWRENCE, C. E. (1999). A Bayesian statistical algorithm for RNA secondary structure prediction. *Comput. Chem.* 23, 387–400.

DING, Y., AND LAWRENCE, C. E. (2003). A statistical sampling algorithm for RNA secondary structure prediction. *Nucl. Acids Res.* 31, 7280–301.

DOUDNA, J., AND CECH, T. (2002). The chemical repertoire of natural ribozymes. *Nature* 418, 222–228.

EICKBUSH, T. H. (2002). R2 and related site-specific non-long terminal repeat retrotransposons. In *Mobile DNA II* (Craig, N. L., Craigie, R., Gellart, M., and Lambowitz, A. M., eds.), (ASM Press, Washinton, DC) p. 813–835.

FERRE-D'AMARE, A., AND DOUDNA, J. (1999). RNA folds: insights from recent crystal structures. *Annu. Rev. Biophys. Biomol. Struct.* 28, 57–73.

GABASHVILI, I. S., AGRAWAL, R. K., SPAHN, C. M., GRASSUCCI, R. A., SVERGUN, D. I., FRANK, J., AND PENCZEK, P. (2000). Solution structure of the *E. coli* 70S ribsome at 11.5 A resolution. *Cell* 100, 537–549.