RNA (1997), 3:1–16. Cambridge University Press. Printed in the USA.

Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element

DAVID H. MATHEWS,[1] ALOKE R. BANERJEE,[1,3] DONGMEI D. LUAN,[2]
THOMAS H. EICKBUSH,[2] and DOUGLAS H. TURNER[1]

[1] Department of Chemistry, University of Rochester, Rochester, New York 14627-0216, USA
[2] Department of Biology, University of Rochester, Rochester, New York 14627, USA

ABSTRACT

RNA transcripts corresponding to the 250-nt 3 untranslated region of the R2 non-LTR retrotransposable element are recognized by the R2 reverse transcriptase and are sufficient to serve as templates in the target DNA-primed reverse transcription (TPRT) reaction. The R2 protein encoded by the *Bombyx mori* R2 can recognize this region from both the *B. mori* and *Drosophila melanogaster* R2 elements even though these regions show little nucleotide sequence identity. A model for the RNA secondary structure of the 3 untranslated region of the *D. melanogaster* R2 retrotransposon was developed by sequence comparison of 10 species aided by free energy minimization. Chemical modification experiments are consistent with this prediction. A secondary structure model for the 3 untranslated region of R2 RNA from the R2 element from *B. mori* was obtained by a combination of chemical modification data and free energy minimization. These two secondary structure models, found independently, share several common sites. This study shows the utility of combining free energy minimization, sequence comparison, and chemical modification to model an RNA secondary structure.

Keywords: methylation; retrotransposon; reverse transcriptase; RNA folding

INTRODUCTION

A thorough understanding of RNA structure and the general rules that determine RNA folding will be invaluable for understanding the mechanisms of RNA catalysis and protein recognition and for the design of therapeutics targeting RNAs. There are several methods available for the determination of RNA secondary and tertiary structure (Jaeger et al., 1993a). X-ray diffraction (Kim et al., 1974; Robertus et al., 1974; Westhof & Sundaralingam, 1986; Dock-Bregeon et al., 1989; Holbrook et al., 1991; Cate et al., 1996) and NMR (Heus & Pardi, 1991; Varani et al., 1991; Varani & Tinoco, 1991; SantaLucia & Turner, 1993; Borer et al., 1995; Greenbaum et al., 1995; Puglisi et al., 1995; Szewczak & Moore, 1995; Ye et al., 1995; Dieckmann et al., 1996; Jiang et al., 1996; Yang et al., 1996) are useful techniques for determining RNA structures, although solving structures larger than tRNA is difficult. Sequence comparison, or phylogeny, is the standard technique for determining secondary structure when genetically related RNAs are available (James et al., 1989) and has even been used to model the three-dimensional structure of the group I self-splicing intron (Michel & Westhof, 1990). Thermodynamic prediction of RNA secondary structure by free energy minimization is an alternative technique when a number of genetically related RNAs are not available (Turner et al., 1988; Zuker, 1989; Jaeger et al., 1989, 1990a; Zuker et al., 1991). In a recent report, the Zuker algorithm for secondary structure prediction based on free energy minimization was found to be 74% accurate, on average, at predicting helices of self-splicing introns and of domains of small subunit rRNAs (Walter et al., 1994). It is less accurate when tested against complete rRNAs, where the average length is 1,500 nt (Konings & Gutell, 1995). Free energy minimization has been used in conjunction with phylogeny to deduce structures for RNAs with a limited number of related sequences (Konings & Hogeweg, 1989; Lück et al., 1996). Finally, chemical modification, a technique in which small chemicals probe for ex-

Reprint requests to: Douglas H. Turner, Department of Chemistry, University of Rochester, Rochester, New York 14627-0216, USA; e-mail: turner@chem.chem.rochester.edu.

[3] Present address: Department of Microbiology and Molecular Genetics, University of Vermont, Burlington, Vermont 05405-0068, USA.

posed bases in an RNA structure, can be used to test and refine models of RNA structure (Inoue & Cech, 1985; Moazed et al., 1986; Ehresmann et al., 1987).

This paper is an examination of the ability to model the secondary structure of a novel functional RNA. Our test case was the 3′ untranslated sequence from R2, a non-LTR retrotransposable element of insects (Burke et al., 1987). The R2 element inserts into a specific sequence of the 28S ribosomal gene of its host and has been found in a few percent to more than half the rDNA units of most insects (Jakubczak et al., 1991). The protein encoded by the R2 element from *Bombyx mori* (R2Bm) has been expressed in *Escherichia coli* (Xiong & Eickbush, 1988) and an in vitro DNA cleavage/reverse transcription system has been developed for studying the retrotransposition mechanism (Luan et al., 1993). The ~250-nt 3′ untranslated region of the R2Bm transcript is recognized specifically by the R2-encoded reverse transcriptase and is required for the RNA to serve as a template in a reaction termed target-primed reverse transcription (TPRT) (Luan et al., 1993; Luan & Eickbush, 1995). In this reaction, the 3′ hydroxyl group generated by cleavage of the 28S gene target site in DNA is used as the primer for reverse transcription.

We show here that the R2Bm reverse transcriptase can also recognize RNA transcripts corresponding to the 250-nt 3′ untranslated region of the *Drosophila melanogaster* R2 element (R2Dm). The 3′ untranslated region of the R2Dm RNA was modeled by sequence comparison of 10 *Drosophila* species aided by free energy minimization. This model was then tested by chemical modification studies. The R2Bm 3′ RNA structure was also modeled. The R2Bm sequence could not be aligned with the *Drosophila* sequences and a diverse set of *Bombyx* species is not available. Thus, it was necessary to deduce its structure independently using chemical modification to sort through a set of structures suggested by free energy minimization.

RESULTS

R2 RNA from *D. melanogaster* is recognized by the R2 reverse transcriptase from *B. mori*

We have previously described an in vitro assay for the combined DNA cleavage and reverse transcription reactions catalyzed by the R2 protein from *B. mori* (Luan et al., 1993; Luan & Eickbush, 1995). A plasmid DNA substrate (pB109) containing a 1.1-kb segment of the 28S gene is predigested with *Eco*R I and *Hin*d III to release the 28S gene fragment with the R2 insertion site from the 2.7-kb pUC vector. These DNA fragments are incubated at 37 °C with the purified R2Bm protein and R2Bm RNA transcripts generated in vitro with T7 RNA polymerase. Three steps in the retrotransposition process occur efficiently in vitro in the presence of 25 μM dNTPs, 0.2 M NaCl, and 10 mM $MgCl_2$, pH 8 (Fig. 1). First, an endonuclease activity of the R2 protein nicks the target site (first strand cleavage). This nick defines the precise location of the eventual R2 insertion. Second, the DNA polymerase activity of the R2 protein reverse transcribes the R2 RNA template starting at or near the 3′ end of the RNA template. The primer for this synthesis is the 3′ OH generated on the target DNA by the first strand cleavage. This reverse transcription converts the 1.1-kb substrate DNA to a branched molecule containing an RNA:DNA hybrid. The third step in R2 retrotransposition occurs after reverse transcription, and involves cleavage of the second DNA strand two base pairs upstream of the nicked site.

Typical examples of such TPRT reactions are shown in Figure 2A. In the absence of any RNA template (lane 1), the R2 protein was only able to cleave the first strand of the target DNA, because both reverse transcription and second strand cleavage are dependent upon the presence of RNA. In the presence of non-R2 RNA sequences (in this case, an 825-nt vector RNA transcript from pBluescript) (lane 2), both first and second strand cleavage of the target DNA occurred.

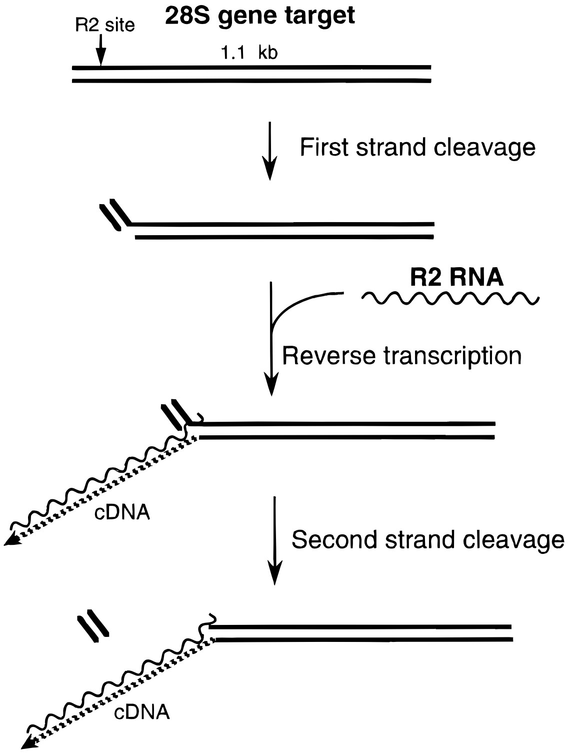

FIGURE 1. Diagram of the initial steps involved in the target-primed reverse transcription assay of the R2 protein. The 28S gene insertion site is located 60 bp from one end of the cloned restriction fragment. Endonuclease activity of the R2 protein first cleaves the lower DNA strand of the DNA target. The reverse transcriptase activity of the R2 protein uses the 3′ hydroxyl group exposed by this cleavage to prime cDNA synthesis using an R2 RNA molecule as template. After reverse transcription, the second strand of the target DNA is cleaved. For further details of this reaction, see Luan et al. (1993).

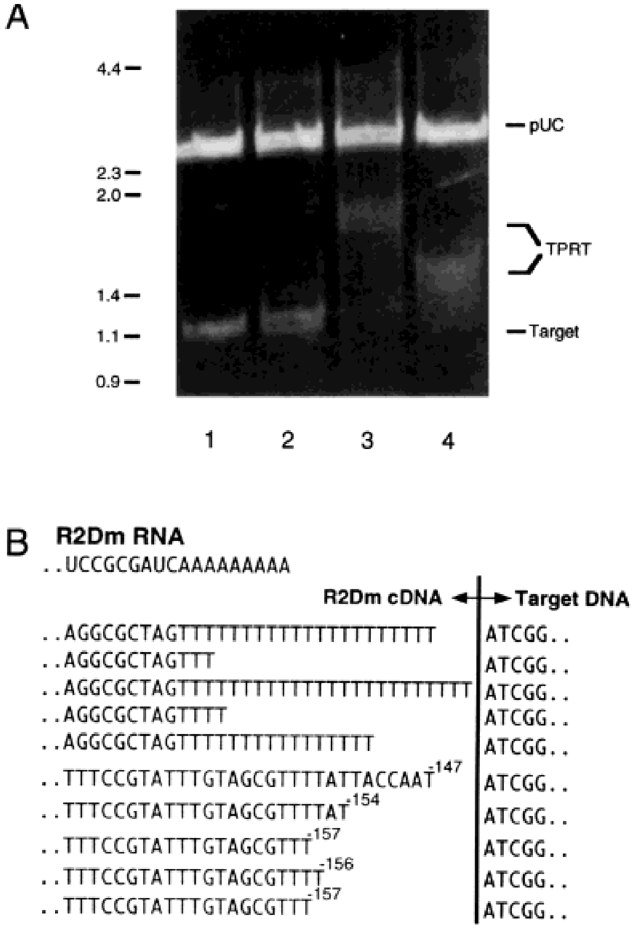
A
4.4
2.3
2.0
1.4
1.1
0.9
pUC
TPRT
Target
1
2
3
4
B
R2Dm RNA
..UCCGCGAUCAAAAAAAAA
R2Dm cDNA
Target DNA
..AGGCGCTAGTTTTTTTTTTTTTTTTTTTTT ATCGG..
..AGGCGCTAGTTT ATCGG..
..AGGCGCTAGTTTTTTTTTTTTTTTTTTTTTTTT ATCGG..
..AGGCGCTAGTTTT ATCGG..
..AGGCGCTAGTTTTTTTTTTTTTTTT ATCGG..
..TTTCCGTATTTGTAGCGTTTTATTACCAAT -147 ATCGG..
..TTTCCGTATTTGTAGCGTTTTAT -154 ATCGG..
..TTTCCGTATTTGTAGCGTTT -157 ATCGG..
..TTTCCGTATTTGTAGCGTTTT -156 ATCGG..
..TTTCCGTATTTGTAGCGTTT -157 ATCGG..

FIGURE 2. RNA from the *D. melanogaster* R2 element can be used as template by the R2 protein encoded by the *B. mori* element. **A:** Ethidium-bromide stained agarose gel of the TPRT reaction products. The 2.7-kb DNA fragment represents the pUC plasmid, whereas the 28S gene target site is located on the 1.1-kb fragment. Reaction conditions are defined in the Materials and methods. Lane 1, no added RNA; lane 2, addition of an 825-nt pBSKS vector RNA; lane 3, addition of an 802-nt HR4 R2 RNA transcript corresponding to the 3′ end of the R2Bm element (similar activity has been observed with RNA transcribed from plasmid pBmR2-249 [Luan & Eickbush, 1995]); lane 4, addition of a 300-nt RNA transcript corresponding to the 3′ end of the R2Dm element. The positions of DNA length markers are shown on the left. The TPRT product with the ~800-nt R2Bm RNA is approximately 1.9 kb, whereas the TPRT product with the 300-nt R2Dm RNA is approximately 1.4 kb. **B:** Sequence of the R2/28S gene junctions generated by the TPRT reaction with R2Dm RNA. The sequence at the 3′ end of the R2Dm RNA template is shown at the top. Below this RNA sequence are the sequences of the cDNA-28S gene junctions obtained by PCR amplification and cloning of the reaction products. Some of the products resulted from reverse transcription that initiated at internal positions within the R2 template. The distances of these internal sites from the 3′ end of the sequence shown in Figure 4 are indicated.

Because the R2 target site was only 60 bp from one end of the 28S gene fragment, second strand cleavage of about one-half of the target molecules results in the 1.1-kb target fragment becoming a doublet. TPRT is dependent upon the presence of the R2 sequences on the RNA template, thus no reverse transcription occurred with the vector RNA sequences. In the presence of an ~800-nt RNA fragment corresponding to the 3′ end of the R2Bm element (lane 3), the 1.1-kb fragment was converted to a diffuse ~1.9-kb fragment representing the Y-shaped and linear TPRT products. We have shown that the size of this TPRT product is directly proportional to the length of the input R2 template as long as the 250-nt 3′ untranslated region of the R2Bm element is located near the 3′ end of the RNA template (Luan & Eickbush, 1995). Deletion of several segments within this 3′ untranslated region eliminated its ability to be used as template in the TPRT reaction, suggesting that the structure recognized by the R2 protein involves sequences throughout this 250 nt.

The 3′ untranslated region of the R2 element from *D. melanogaster* is also approximately 250-nt in length (Jakubczak et al., 1990). The exact length is variable due to a poly(A) tail present at the 3′ junction with the 28S gene. Sequence identity within the 3′ untranslated region of the R2 elements from *B. mori* and *D. melanogaster* is minimal, possibly limited to only a few short regions (Eickbush et al., 1995). To determine if the R2 protein from *B. mori* could also recognize the RNA transcript from the *D. melanogaster* R2 element (R2Dm), an ~300-nt RNA fragment corresponding to the 3′ untranslated region of the R2Dm element was used in the TPRT reaction (lane 4). The target DNA was converted to a diffuse band extending up to 1.4 kb in length, the expected size of a TPRT product with a 300-nt RNA template. To confirm that this product band was the result of a TPRT reaction, the products were PCR amplified with one primer complementary to the cDNA strand generated by reverse transcription of the *D. melanogaster* RNA transcript and the second primer complementary to the 28S gene downstream of the insertion site. The PCR products were ligated into a sequencing vector (Burke et al., 1995) and individual clones were sequenced (Fig. 2B). Five of the sequenced clones had derived from reverse transcription initiating within the short poly(A) tail defining the 3′ end of the R2Dm transcript. The other five clones had initiated reverse transcription at internal sites within the R2Dm sequence (Fig. 2B). These internal deletions explain the diffuse nature of the TPRT product in lane 4 (Fig. 2A). Initiation of reverse transcription at internal sites within the R2 3′ untranslated region are also detected at low levels with R2Bm templates, becoming more frequent if small deletions or substitutions are made at the 3′ end of the template (Luan & Eickbush, 1995). These results demonstrate that the *D. melanogaster* R2 RNA can be used by the R2 protein from *B. mori* in the TPRT reaction. The efficiency of the reaction is comparable to that of the R2Bm template. The R2Bm protein, however, appears less able to position correctly the R2Dm RNA to enable the initiation of reverse transcription at the precise 3′ end of the template.

Modeling the structure of the *Drosophila* R2 3 untranslated region

The secondary structure of the *Drosophila* R2 3′ sequences was modeled by sequence comparison of 10 species: *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. mauritiana*, *D. yakuba*, and *D. teissieri* from the *melanogaster* species subgroup; *D. ananassae* and *D. takahashii* individual species from two other subgroups of the *melanogaster* species group; and D. *pseudoobscura* and *D. ambigua* from the *ambigua* species group (Eickbush & Eickbush, 1995; Eickbush et al., 1995; W.D. Lathe & T.H. Eickbush, in prep.). A sequence alignment was conducted initially by simply matching bases. The alignment for the five species representing most of the sequence diversity is shown in Figure 3A. This simple alignment was examined for putative conserved helices, allowing for shifts in the register of the original sequence alignment. This process was aided by comparing the thermodynamically predicted structures of each species, generated by the Zuker (1989) algorithm, with the parameters of Walter et al. (1994), as described in the Materials and methods. Based on these thermodynamic predictions, four regions of the RNA sequence that contained low levels of sequence identity (boldface sequences in Fig. 3A) required shifts in the register of the sequence alignment. The sequence alignment was also examined for possible pseudoknots, but none were found that were consistent with phylogeny.

The modeled structure has seven helices of at least three base pairs that are conserved in all species. The final sequence alignment of all 10 species is shown in Figure 3B, with the seven proposed helices underlined (labeled A–G). There are 13 compensating changes, defined as a change in the helix of one species from an A-U to a G-C pair or the reverse in the orientation of a pair, e.g., 5′-G-C-3′ replaced by a 5′-C-G-3′ (boldface positions in Fig. 3B). In addition to the compensating changes, there are six pairs in which at least one species has a change in either a G-C or an A-U to a G-U pair. These nucleotide changes support the validity of six of the seven helical regions (the seventh helix, D, showed no sequence variation).

The final model for the secondary structure of the R2Dm 3′ untranslated region is shown in Figure 4. Except for helix A, the structure predicted solely on thermodynamic considerations was identical to the final structure based on phylogenetic information. The structure in Figure 4 shows three additional helices (labeled 1–3) that were inferred entirely from thermodynamic considerations. The other *Drosophila* species have a similar overall structure to that of *D. melanogaster*, with the major variation being in the length and sequence of the regions containing helices 1–3. In Figure 4A, nucleotides in the *melanogaster* structure that are invariant through the 10 species are boxed, and those with compensating changes are circled. Also in Figure 4A is an alternative structure for helix C that is consistent with the alignment.

Chemical modification of R2Dm

To test the secondary structure model of R2Dm derived from the above phylogenetic approach, chemical modification of an RNA containing the 250-nt R2Dm RNA was conducted with 1-cyclohexyl-3-(2-morpholinoethyl)carbodiimide metho-*p*-toluenesulfonate (CMCT), dimethyl sulfate (DMS), and β-ethoxy-α-ketobutyraldehyde (kethoxal). These reagents react at nucleotides not involved in Watson–Crick pairing and at nucleotides at the ends of helices. Thermal melting curves were used to choose conditions for the chemical modification experiments. Formation of tertiary structure often buries non-base paired nucleotides, leaving them nonreactive to modification reagents (Inoue & Cech, 1985; Banerjee et al., 1993). Tertiary structure, however, often melts at lower temperatures than most secondary structure (Crothers et al., 1974; Hilbers et al., 1976; Banerjee et al., 1993; Jaeger et al., 1993b; Laing & Draper, 1994) and is also destabilized at low Mg^{2+} concentrations (Inoue & Cech, 1985; Jaeger et al., 1990b; Laing & Draper, 1994). Thus, better definition of secondary structure may be achieved from chemical modification at elevated temperatures and reduced Mg^{2+} concentration. Figure 5 shows melting curves obtained at 0.1, 1.0, and 5 mM free Mg^{2+}. The best separation of transitions, i.e., separation in the maxima of the derivative, is achieved with 1 mM Mg^{2+}. This gives transitions centered around 10, 55, and 70 °C. Chemical modification was done at 20 °C to map the structure between the first and second transitions. Another map was done at 42 °C to examine the structure in the second transition region.

The modification data at 20 and 42 °C are shown on the proposed secondary structure for R2Dm RNA in Figure 4B and C, respectively. The map at 20 °C is consistent with the proposed phylogenetic structure when allowing for modifications at G-U pairs, as has been observed in other studies (Moazed et al., 1986; Banerjee et al., 1993). The modification pattern is also consistent with the alternative helix in Figure 4A. Helix 1 (a helix unique to *melanogaster*) shows a modification that can be explained by the rearrangement shown in Figure 4B. If both helices occur in solution, then A17 and A19 could both be accessible to modification as observed. It is also possible that the base pairing in helix 1 is perturbed by vector-derived bases at the 5′ end of the construct, forming a separate structure in equilibrium with the two shown.

At 42 °C (Fig. 4C), more modification hits are expected due to unfolding of the tertiary structure. For example, for two group I introns, the lowest temperature transition is the unfolding of tertiary structure (Banerjee et al., 1993; Jaeger et al., 1993b). For tRNA,

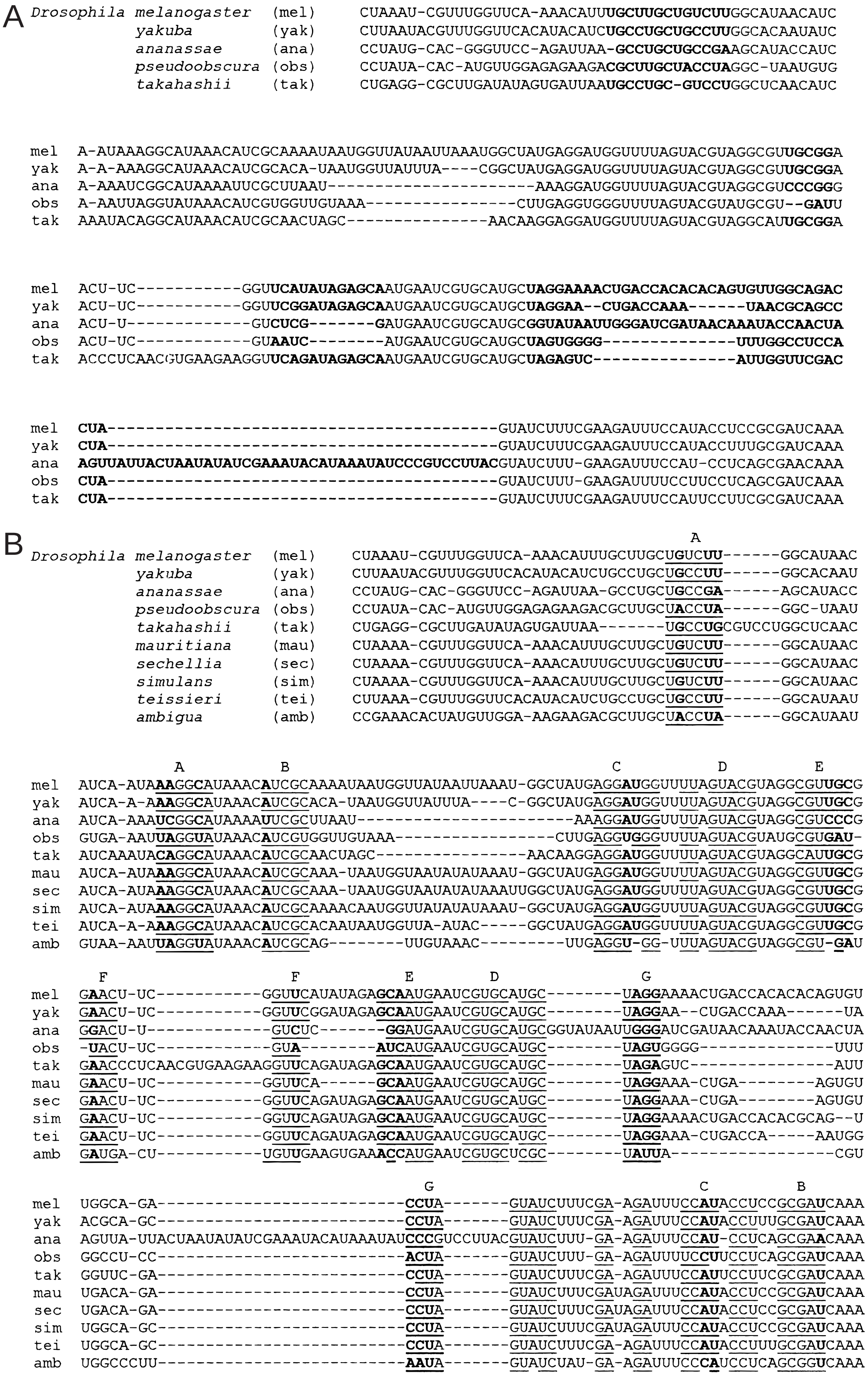
A
Drosophila melanogaster (mel) CUAAAU-CGUUUGGUUCA-AAACAUUUGCUUGCUGUCUUGGCAUAACAUC
yakuba (yak) CUUAAUACGUUUGGUUCACAUACAUCUGCCUGCUGCCUUGGCACAAUAUC
ananassae (ana) CCUAUG-CAC-GGGUUCC-AGAUUAA-GCCUGCUGCCGAAGCAUACCAUC
pseudoobscura (obs) CCUAUA-CAC-AUGUUGGAGAGAAGACGCUUGCUACCUAGGC-UAAUGUG
takahashii (tak) CUGAGG-CGCUUGAUAUAGUGAUUAAUGCCUGC-GUCCUGGCUCAACAUC
mel A-AUAAAGGCAUAAACAUCGCAAAAUAAUGGUUAUAAUUAAAUGGCUAUGAGGAUGGUUUUAGUACGUAGGCGUUGCGGA
yak A-A-AAAGGCAUAAACAUCGCACA-UAAUGGUUAUUUA----CGGCUAUGAGGAUGGUUUUAGUACGUAGGCGUUGCGGA
ana A-AAAUCGGCAUAAAAUUCGCUUAAU----------------------AAAGGAUGGUUUUAGUACGUAGGCGUCCCGGG
obs A-AAUUAGGUAUAAACAUCGUGGUUGUAAA----------------CUUGAGGUGGGUUUUAGUACGUAUGCGU--GAUU
tak AAAUACAGGCAUAAACAUCGCAACUAGC---------------AACAAGGAGGAUGGUUUUAGUACGUAGGCAUUGCGGA
mel ACU-UC-----------GGUUCAUAUAGAGCAAUGAAUCGUGCAUGCUAGGAAAACUGACCACACACAGUGUUGGCAGAC
yak ACU-UC-----------GGUUCGGAUAGAGCAAUGAAUCGUGCAUGCUAGGAA--CUGACCAAA------UAACGCAGCC
ana ACU-U-------------GUCUCG-------GAUGAAUCGUGCAUGCGGUAUAAUUGGGAUCGAUAACAAAUACCAACUA
obs ACU-UC------------GUAAUC--------AUGAAUCGUGCAUGCUAGUGGGG--------------UUUGGCCUCCA
tak ACCCUCAACGUGAAGAAGGUUCAGAUAGAGCAAUGAAUCGUGCAUGCUAGAGUC---------------AUUGGUUCGAC
mel CUA---GUAUCUUUCGAAGAUUUCCAUACCUCCGCGAUCAAA
yak CUA---GUAUCUUUCGAAGAUUUCCAUACCUUUGCGAUCAAA
ana AGUUAUUACUAAUAUAUCGAAAUACAUAAAUAUCCCGUCCUUACGUAUCUUU-GAAGAUUUCCAU-CCUCAGCGAACAAA
obs CUA---GUAUCUUU-GAAGAUUUUCCUUCCUCAGCGAUCAAA
tak CUA---GUAUCUUUCGAAGAUUUCCAUUCCUUCGCGAUCAAA
B
A
Drosophila melanogaster (mel) CUAAAU-CGUUUGGUUCA-AAACAUUUGCUUGCUGUCUU------GGCAUAAC
yakuba (yak) CUUAAUACGUUUGGUUCACAUACAUCUGCCUGCUGCCUU------GGCACAAU
ananassae (ana) CCUAUG-CAC-GGGUUCC-AGAUUAA-GCCUGCUGCCGA------AGCAUACC
pseudoobscura (obs) CCUAUA-CAC-AUGUUGGAGAGAAGACGCUUGCUACCUA------GGC-UAAU
takahashii (tak) CUGAGG-CGCUUGAUAUAGUGAUUAA-------UGCCUGCGUCCUGGCUCAAC
mauritiana (mau) CUAAAA-CGUUUGGUUCA-AAACAUUUGCUUGCUGUCUU------GGCAUAAC
sechellia (sec) CUAAAA-CGUUUGGUUCA-AAACAUUUGCUUGCUGUCUU------GGCAUAAC
simulans (sim) CUAAAA-CGUUUGGUUCA-AAACAUUUGCUUGCUGUCUU------GGCAUAAC
teissieri (tei) CUUAAA-CGUUUGGUUCACAUACAUCUGCCUGCUGCCUU------GGCAUAAU
ambigua (amb) CCGAAACACUAUGUUGGA-AAGAAGACGCUUGCUACCUA------GGCAUAAU
A B C D E
mel AUCA-AUAAAGGCAUAAACAUCGCAAAAUAAUGGUUAUAAUUAAAU-GGCUAUGAGGAUGGUUUUAGUACGUAGGCGUUGCG
yak AUCA-A-AAAGGCAUAAACAUCGCACA-UAAUGGUUAUUUA----C-GGCUAUGAGGAUGGUUUUAGUACGUAGGCGUUGCG
ana AUCA-AAAUCGGCAUAAAAUUCGCUUAAU-----------------------AAAGGAUGGUUUUAGUACGUAGGCGUCCCG
obs GUGA-AAUUAGGUAUAAACAUCGUGGUUGUAAA-----------------CUUGAGGUGGGUUUUAGUACGUAUGCGUGAU-
tak AUCAAAUACAGGCAUAAACAUCGCAACUAGC----------------AACAAGGAGGAUGGUUUUAGUACGUAGGCAUUGCG
mau AUCA-AUAAAGGCAUAAACAUCGCAAA-UAAUGGUAAUAUAUAAAU-GGCUAUGAGGAUGGUUUUAGUACGUAGGCGUUGCG
sec AUCA-AUAAAGGCAUAAACAUCGCAAA-UAAUGGUAAUAUAUAAAUUGGCUAUGAGGAUGGUUUUAGUACGUAGGCGUUGCG
sim AUCA-AUAAAGGCAUAAACAUCGCAAAACAAUGGUUAUAUAUAAAU-GGCUAUGAGGAUGGUUUUAGUACGUAGGCGUUGCG
tei AUCA-A-AAAGGCAUAAACAUCGCACAAUAAUGGUUA-AUAC-----GGCUAUGAGGAUGGUUUUAGUACGUAGGCGUUGCG
amb GUAA-AAUUAGGUAUAAACAUCGCAG--------UUGUAAAC---------UUGAGGU-GG-UUUAGUACGUAGGCGU-GAU
F F E D G
mel GAACU-UC-----------GGUUCAUAUAGAGCAAUGAAUCGUGCAUGC--------UAGGAAAACUGACCACACACAGUGU
yak GAACU-UC-----------GGUUCGGAUAGAGCAAUGAAUCGUGCAUGC--------UAGGAA--CUGACCAAA------UA
ana GGACU-U-------------GUCUC-------GGAUGAAUCGUGCAUGCGGUAUAAUUGGGAUCGAUAACAAAUACCAACUA
obs -UACU-UC------------GUA--------AUCAUGAAUCGUGCAUGC--------UAGUGGGG--------------UUU
tak GAACCCUCAACGUGAAGAAGGUUCAGAUAGAGCAAUGAAUCGUGCAUGC--------UAGAGUC---------------AUU
mau GAACU-UC-----------GGUUCA------GCAAUGAAUCGUGCAUGC--------UAGGAAA-CUGA--------AGUGU
sec GAACU-UC-----------GGUUCAGAUAGAGCAAUGAAUCGUGCAUGC--------UAGGAAA-CUGA--------AGUGU
sim GAACU-UC-----------GGUUCAGAUAGAGCAAUGAAUCGUGCAUGC--------UAGGAAAACUGACCACACGCAG--U
tei GAACU-UC-----------GGUUCAGAUAGAGCAAUGAAUCGUGCAUGC--------UAGGAAA-CUGACCA-----AAUGG
amb GAUGA-CU-----------UGUUGAAGUGAAACCAUGAAUCGUGCUCGC--------UAUUA-----------------CGU
G C B
mel UGGCA-GA--------------------------CCUA-------GUAUCUUUCGA-AGAUUUCCAUACCUCCGCGAUCAAA
yak ACGCA-GC--------------------------CCUA-------GUAUCUUUCGA-AGAUUUCCAUACCUUUGCGAUCAAA
ana AGUUA-UUACUAAUAUAUCGAAAUACAUAAAUAUCCCGUCCUUACGUAUCUUU-GA-AGAUUUCCAU-CCUCAGCGAACAAA
obs GGCCU-CC--------------------------ACUA-------GUAUCUUU-GA-AGAUUUUCCUUCCUCAGCGAUCAAA
tak GGUUC-GA--------------------------CCUA-------GUAUCUUUCGA-AGAUUUCCAUUCCUUCGCGAUCAAA
mau UGACA-GA--------------------------CCUA-------GUAUCUUUCGAUAGAUUUCCAUACCUCCGCGAUCAAA
sec UGACA-GA--------------------------CCUA-------GUAUCUUUCGAUAGAUUUCCAUACCUCCGCGAUCAAA
sim UGGCA-GC--------------------------CCUA-------GUAUCUUUCGAUAGAUUUCCAUACCUCCGCGAUCAAA
tei UGGCA-GC--------------------------CCUA-------GUAUCUUUCGA-AGAUUUCCAUACCUUUGCGAUCAAA
amb UGGCCCUU--------------------------AAUA-------GUAUCUAU-GA-AGAUUUCCCAUCCUCAGCGGUCAAA

FIGURE 3. Alignment of *Drosophila* sequences. **A:** Initial alignment of five sequences determined only by matching bases. Boldfaced regions are those that were changed in the final alignment. **B:** Final alignment of all 10 sequences demonstrating the modeled secondary structure. Underlined regions represent bases in phylogenetically determined helices, which are labeled above the alignment by letters. Columns in boldface are positions in which compensating changes were found. The base pairing of a hairpin loop such as that closed by helix F in *ananassae* has been determined by NMR (Jucker & Pardi, 1995).

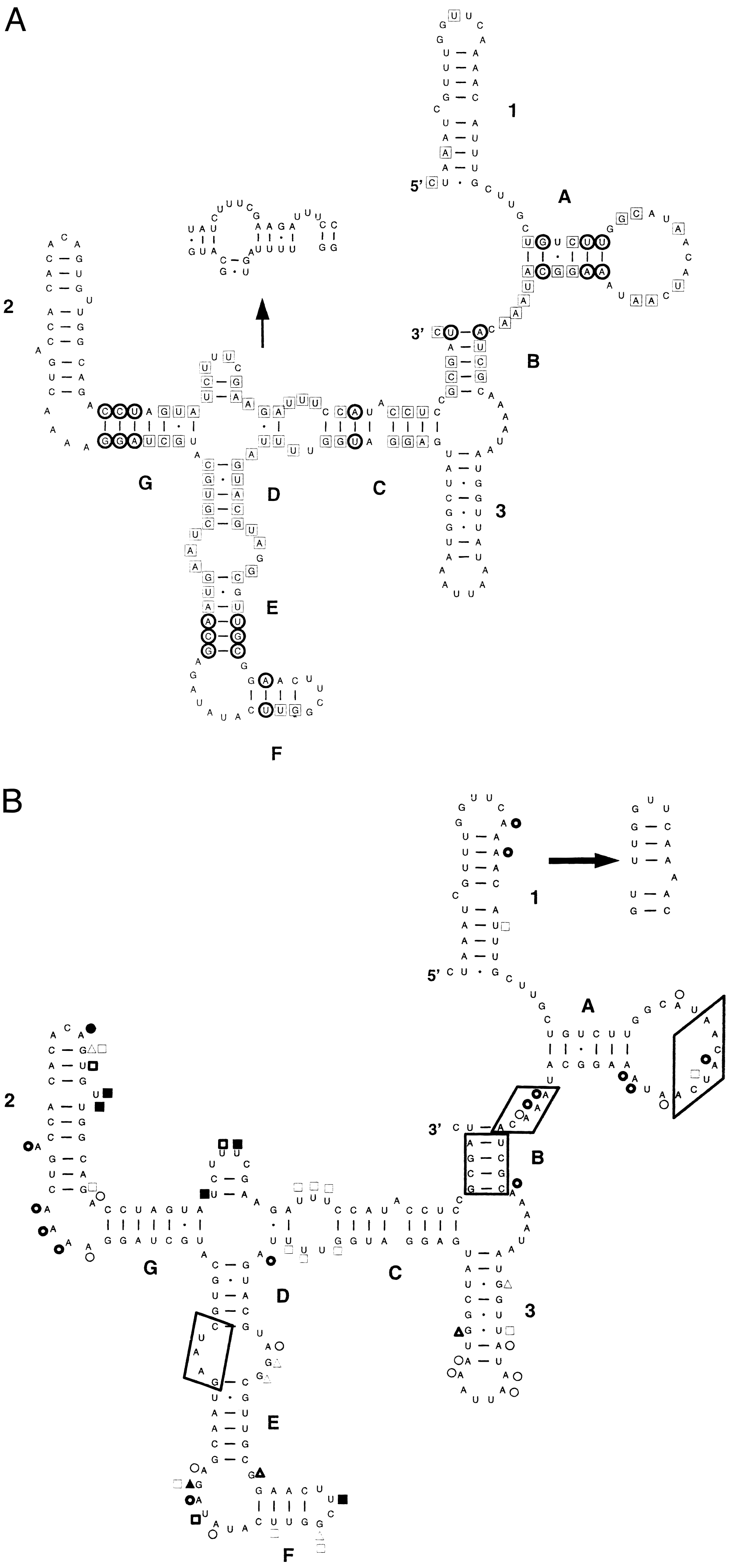

FIGURE 4. Secondary structure model for the R2Dm 3′ untranslated RNA. **A:** Nucleotides invariant throughout the alignment (Fig. 3) are boxed and nucleotides in positions of compensating changes are circled. The arrow points to an alternative helix that is consistent with phylogeny. Lettered helices correspond to phylogenetically determined helices labeled in the alignment (Fig. 3). Numbered helices are regions unique to *melanogaster*. **B:** Chemical modifications at 20 °C. The arrow points to a possible rearrangement consistent with the modification data. Boxed regions are structural elements in common with those in the R2Bm 3′ untranslated RNA structure model (Fig. 6). **C:** Chemical modification at 42 °C. Triangles represent modifications by kethoxal, circles represent DMS, and squares represent CMCT. Solid symbols represent strong modification, darkly outlined symbols represent moderate modification, and lightly outlined symbols represent weak modification. The chemically mapped RNA contained the additional vector sequences $G_3CGA_2U_2G_3U AC_2G_3C_7UCGAG_2UC$ on the 5′ end and $AGC U_2GAUAUCGA_2U_2$ at the 3′ end. (*Figure continued on facing page.*)

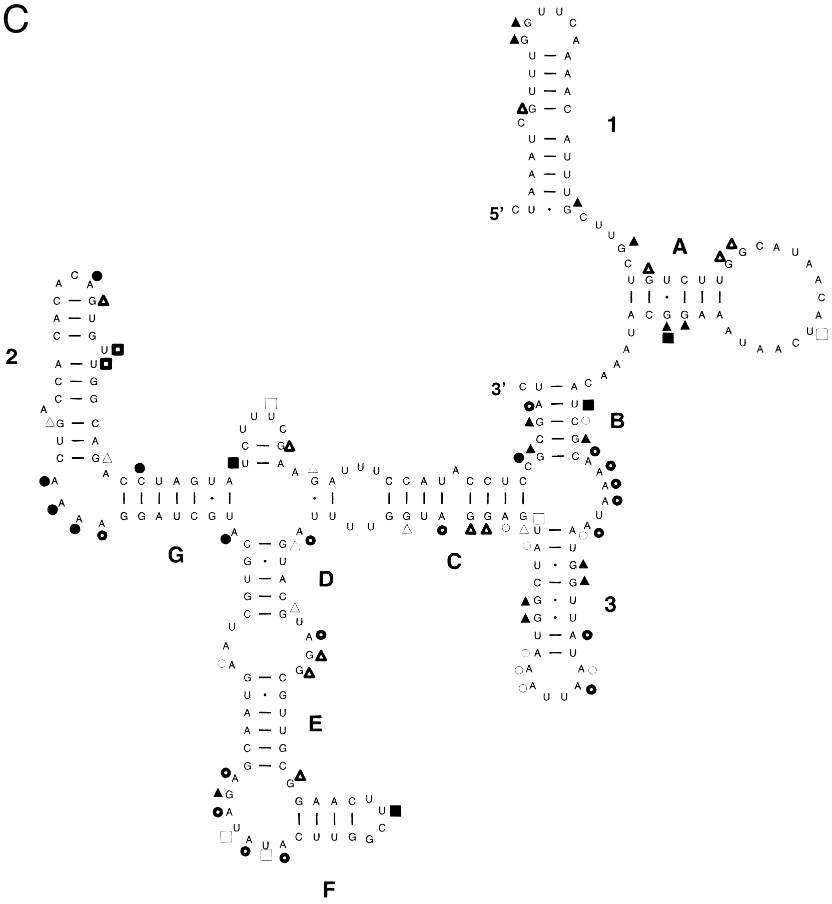

FIGURE 4. *Continued.*

the lowest transition is the unfolding of tertiary structure and a helix in the cloverleaf (Crothers et al., 1974; Hilbers et al., 1976). For R2Dm RNA, the hits at 42 °C are consistent with the unfolding of both tertiary structure and some secondary structure, mostly in helix B, which has a predicted melting temperature of about 55 °C (Freier et al., 1986; Williams et al., 1989; Serra & Turner, 1995).

Modeling the structure of R2Bm 3 RNA

The R2Bm 3′ untranslated RNA could not be aligned successfully to the *Drosophila* phylogenetic structure. Therefore, to model the R2Bm structure, a computer program, called mix&match, was written to use thermodynamic structure prediction in conjunction with chemical modification data (see the Materials and methods). Mix&match combines the most thermodynamically favorable domains that are also compatible with the modification data into an overall structure. This was necessary because strong modification can occur at terminal base pairs at either end of helices and the recursive algorithm for free energy minimization cannot take this into account. The condition that a base pair involving a modified base can occur only if the next base is not paired violates recursion. It is also not sufficient to search through suboptimal structures generated by the Zuker algorithm to find a structure that is compatible with the data. This is because the Zuker algorithm does not produce every possible structure within the specified increment of the lowest free energy. Rather, it produces representative structures, the total of which show each possible base pair (Zuker, 1989; Jaeger et al., 1990a).

The secondary structure model of the R2Bm 3′ RNA deduced with mix&match is shown in Figure 6 with all chemical modification hits mapped. pBSK(-K) vector sequences present at the 5′ and 3′ ends of the R2Bm RNA made in vitro by T7 RNA polymerase were included in this structure prediction. The predicted structure shows these extra nucleotides forming domains apart from the main body of the structure.

A structure predicted by free energy minimization, without taking into account the modification data, matched the R2Bm model except for two locations. It did not predict any of the base pairs in helix 6 and, in helix 9, it predicted a base pair between A210 and U229, in disagreement with the strong modification observed for G209.

The proposed secondary structure model is consistent with all 46 strong hits by design. Of 77 moderate and light hits, only 7 moderate and 3 light hits mapped to nonterminal Watson–Crick base pairs not adjacent to G-U pairs. This is consistent with the proposed struc-

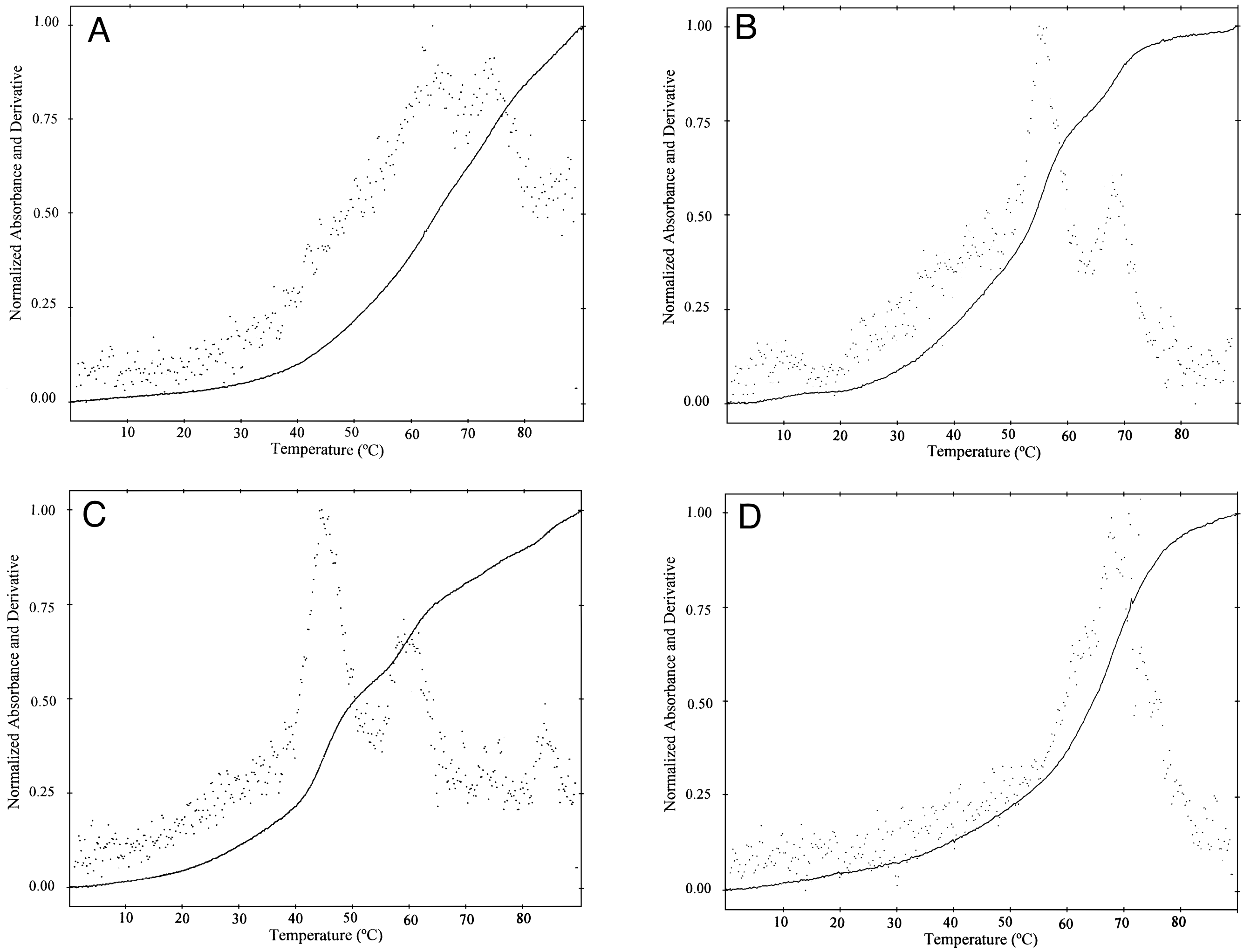

FIGURE 5. Melting curves and their derivatives. Optical melting curves for R2Dm RNA in 5.5 (**A**), 1.5 (**B**), and 0.6 mM $MgCl_2$ (**C**) are shown. Curve of R2Bm RNA at 1.5 mM $MgCl_2$ is also shown (**D**). The continuous line is the absorbance and the dots are the derivative of absorbance. All melts were in 80 mM Hepes, pH 7.5, 10 mM NaCl, and 0.5 mM EDTA.

ture model if a small population of other structures occur in solution or if a small portion of kinetically trapped structures exist in the region of those helices. This is possible because the recognition assay (above) does not require that all the strands be in functional form.

DISCUSSION

The 120-kDa protein encoded by the R2 retrotransposable element of *B. mori* contains an endonuclease activity that specifically requires the 35 bp surrounding the 28S gene insertion site (Xiong & Eickbush, 1988) and a reverse transcriptase activity that is highly specific for the RNA sequence corresponding to the 250-nt 3′ untranslated region of the element (Luan & Eickbush, 1995). In this report, we have shown that this protein from *B. mori* will also recognize the RNA corresponding to the 3′ untranslated region of the R2 element from *D. melanogaster*. Because the nucleotide sequences of these two 3′ untranslated regions show very little sequence identity, we determined separately the secondary structure of each RNA sequence. The R2Dm 3′ untranslated RNA structure was modeled by phylogenetic comparison aided by free energy minimization and supported with chemical modification. The R2Bm 3′ untranslated RNA structure was modeled by free energy minimization aided by chemical modification, because no closely related sequences are available for this species.

Approximately 50% (120 nt) of the *Drosophila* 3′ untranslated region is composed of a conserved core (Fig. 4). Six helical regions (A–E, G) are the basis of this core. Except for a few nucleotides at helix termini or in loops, all positions of these helices are either conserved in sequence or there are compensatory changes that confirm the importance of these paired regions. The remaining half of the R2Dm RNA structure (120 nt) contains four regions that are highly variable in both sequence and length. Comparison of the R2Bm structure to the conserved core of the R2Dm structure reveals sev-

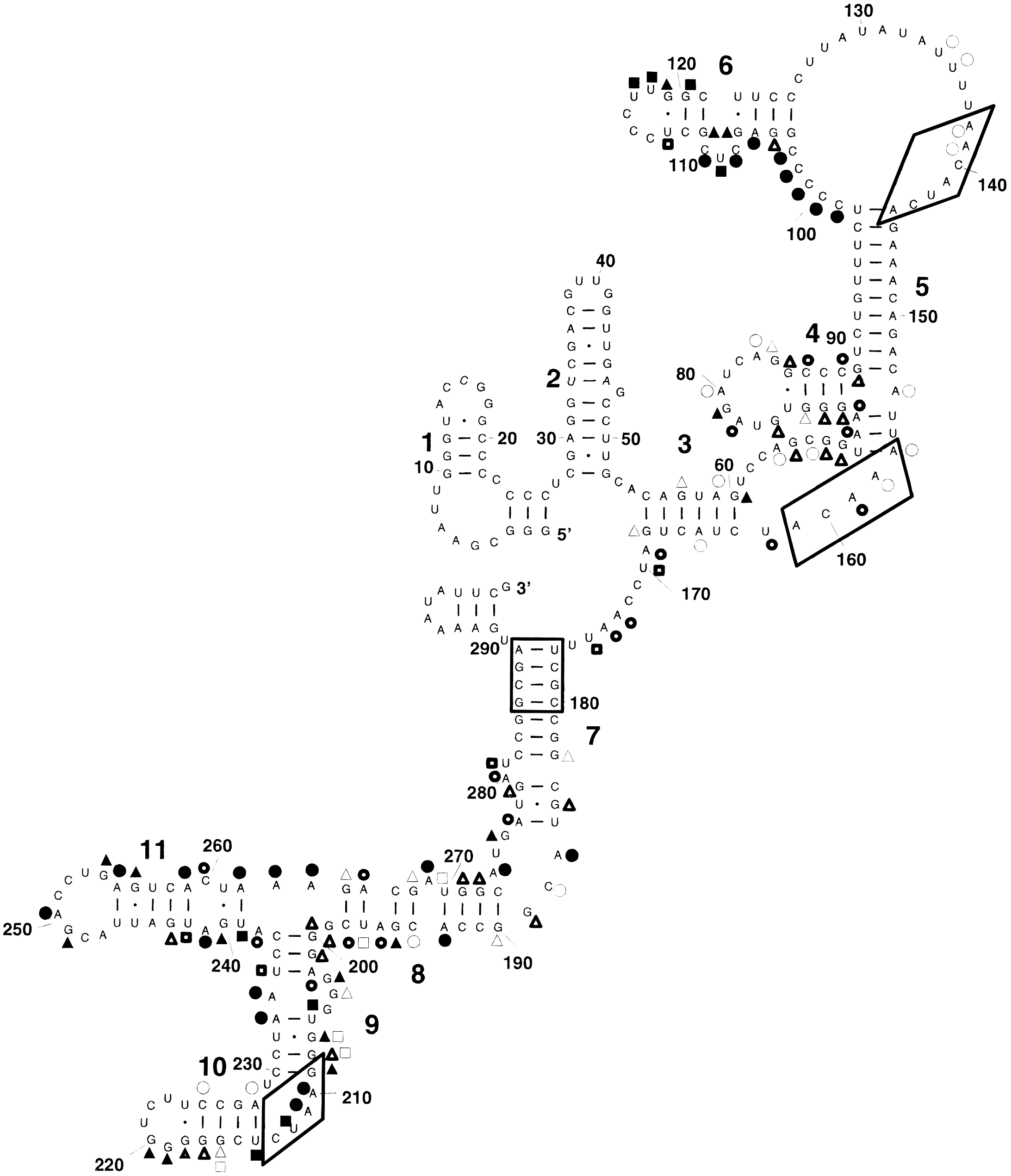

FIGURE 6. Secondary structure model of the R2Bm 3′ RNA with a chemical modification map at 20 °C. Helices are numbered. Triangles represent modifications by kethoxal, circles represent DMS, and squares represent CMCT. Solid symbols represent strong modification, darkly outlined symbols represent moderate modification, and lightly outlined symbols represent weak modification. Boxed regions are structural elements in common with those in the R2Dm 3′ untranslated RNA structure model (Fig. 4B). The base numbering is of all bases in the pBmR2-249 construct (the R2Bm untranslated sequence begins at base 47 and ends at 294).

eral similarities. Both structures contain a helical region that pairs a region within a few nucleotides of the 3′ end of the element to a sequence well within the 3′ untranslated region (helix B in Fig. 4 and helix 7 in Fig. 6). Four base pairs of this helical stem are even conserved in sequence between *B. mori* and the *Drosophila* elements. This conserved helix is preceded by a short (5–8 nt) single-stranded region connecting to another helix (A in Fig. 4, 3 in Fig. 6), which is annealed to a region very near (7–30 nt) to the termination codon of the R2 elements. The loop formed by helix A is usually about 15 nt in the *Drosophila* species. The equivalent region in *B. mori* is much larger (101 nt), with several helical regions. However, a short single-stranded part of this region, AACAUCA, is identical in sequence with the central portion of the *Drosophila* loop (boxed region). Another possible similarity between the structures of the *D. melanogaster* and *B. mori* structures is the sequence GAAUC, forming one side of an internal loop between helical stems D and E in the *Drosophila* structures, and nt 209–213 of the *B. mori* structure. Finally, both secondary structures have a GU pair as the penultimate pair of a helix flanking a multibranch loop (see helix G in Fig. 4A and 240 G•U 261 in Fig. 6). Conserved GU pairs are thought to be important for protein recognition and tertiary interactions (Hou & Schimmel, 1988; McClain & Foss, 1988; Gautheret et al., 1995; Strobel & Cech, 1995; Gabriel et al., 1996).

Previously conducted experiments (Luan & Eickbush, 1995) support the secondary structure model proposed here for the R2Bm 3′ untranslated region and the importance of the regions discussed above. We have shown that RNA templates containing only the 200-nt 3′ end of the R2Bm element are not used as a template in the TPRT reaction. This 50-nt deletion would remove the region forming part of the helical domain 3 in Figure 6, which is equivalent to A in Figure 4A. Second, we have found that deletions or substitutions of the 6 nt at the extreme 3′ end of the R2Bm RNA (288–294 in Fig. 6) can affect the efficiency of its utilization in the TPRT reaction and the tendency to add nontemplate nucleotides, but the vast majority of the initiations still occur at the 3′ end of the template. If, however, we make substitutions that extend to the GC-rich helix 7 of the R2Bm structure, then 90% of the TPRT reactions do not initiate reverse transcription at the 3′ end of the RNA. Finally, another indication supporting the secondary structure for R2Dm is that the internal start sites observed for half of the reverse transcription products (see Fig. 2B) are near helix B, which is paired to the 3′ end of the RNA (see Fig. 4).

This study demonstrates the utility of free energy minimization when used with phylogeny and chemical mapping to model RNA secondary structure. Although free energy minimization and phylogeny were combined manually for this study, an algorithm has been reported recently that automates the process (Lück et al., 1996). The free energy rules are derived largely from thermodynamic studies of oligonucleotides and from comparison of predicted and phylogenetically determined structures of ribosomal and self-splicing RNAs (Turner et al., 1988; Walter et al., 1994). The results here suggest that these rules work well for an RNA with a completely different function. Tests of the 1989 version (Jaeger et al., 1989; Zuker, 1989) of the folding algorithm against known secondary structures of entire rRNAs indicated that predictions for eukaryotic sequences were much worse than for archæa and prokaryotic sequences (Konings & Gutell, 1995). This trend is not observed in shorter sequences (Jaeger et al., 1989; Walter et al., 1994). The results presented here also indicate that the folding algorithm can give reasonable predictions for eukaryotic sequences of about 250 nt. These secondary structure models provide a framework for conducting site-directed mutagenesis and deletion experiments to further define the structure and protein binding regions.

MATERIALS AND METHODS

Protein purification and the TPRT reaction

R2 protein was purified from *E. coli* JM109/pR260 as described previously (Luan et al., 1993; Luan & Eickbush, 1995). All TPRT reactions were conducted in 20-μL volumes containing 0.4 μg of pB109 plasmid DNA predigested with *Hin*d III and *Eco*R I, 0.2 μg RNA templates with 15 U of RNasin (Pharmacia), and 0.1 μg of R2 protein peak isolated from the DNA–cellulose column. Reaction conditions were 200 mM NaCl, 10 mM $MgCl_2$, 50 mM Tris-HCl, pH 8.0, 5 mM dithiothreitol, and 25 μM deoxynucleoside triphosphates (dNTPs). Incubations were for 2 h at 37 °C. The products were mixed with 5× loading buffer (0.02% bromophenol blue, 5% Sarkosyl, 100 mM EDTA, pH 8.0, and 50% glycerol).

Production of RNA

The 3′ untranslated region RNAs of *B. mori* and *D. melanogaster* were made by T7 runoff transcription of cleaved plasmids. *B. mori* RNA was prepared from plasmid HR4 (Luan et al., 1993) and pBmR2-249 (Luan & Eickbush, 1995). Construct pDmR2-238, used to produce the *D. melanogaster* RNA, was generated by PCR amplification of R2Dm sequence in clone p303 (Jakubczak et al., 1990). To generate a R2Dm transcript, which ended with a short poly(A) tail, the downstream primer, 5′-GTTGACAAGCTTTTTTTTGATCGCGGAGGT ATG-3′, was used in combination with a sequencing primer within the R2Dm element, approximately 400 bp from its 3′ end. A 280-bp *Hin*d III–*Alu* I fragment from the PCR amplification was cloned into the *Hin*d III + *Hin*c II digested pBSKS(−) plasmid. For the experiment in Figure 2, the HR4 construct was digested with *Xmn* I to generate an RNA ending at the precise 3′ end of the R2Bm element, and the pDmR2-238 plasmid was digested with *Hin*d III to generate an RNA ending with eight A nucleotides.

To generate RNA for the chemical modification experiments, clones pBmR2-249 and pDmR2-238 were digested with *Eco*R I. Each RNA was purified by PAGE on a 4% gel of 7 M urea. Bands were visualized by UV shadowing and excised. The RNA was eluted into 0.1% SDS, 0.1 mM Na_2EDTA, and 40 mM ammonium acetate by the crush and soak method (Barfod & Cech, 1988). It was then desalted on a Sephadex G-50 column, precipitated in ethanol, and stored at −20 °C. The final *B. mori* RNA carried 36 bases of vector sequence and 10 bases of R2 sequence before the untranslated region. The *D. melanogaster* RNA had 34 nt of vector sequence on the 5′ end and 16 nt on the 3′ end.

Computer prediction of secondary structures by free energy minimization

Structures were predicted on the basis of free energy minimization as follows (Jaeger et al., 1990a). The dynamic algorithm of Zuker (1989) was used to produce a set of suboptimal structures within 20% of the lowest free energy with a window size of zero. The thermodynamic parameters were those used by Walter et al. (1994) (Freier et al., 1986; Jaeger et al., 1989; He et al., 1991). A second program, efn2, then recalculated free energies using a model for coaxial stacking (Walter et al., 1994) and a Jacobson–Stockmayer function for stabilities of multibranch loops larger than 6 nt (Jacobson & Stockmayer, 1950). The structure with the lowest free energy after this recalculation was chosen as the thermodynamic prediction. This work was done on a Silicon Graphics work station. The FORTRAN program (Walter et al., 1994) is available from Michael Zuker on his World Wide Web homepage at http://www.ibc.wustl.edu/~zuker/.cgi. A C++ version of the program, written for a PC with Windows 95 or Windows NT, is available on the Turner group homepage at http://rna.chem.rochester.edu.

Because only one sequence was available for the R2 *Bombyx* 3′ untranslated RNA, chemical modification data were used as constraints in modeling the secondary structure. The program mix&match, written for this purpose, uses the following approach. First, a set of suboptimal structures is generated by the Zuker algorithm as described above. Then, for each structure, mix&match uses the efn2 algorithm to calculate the free energy of each possible subdomain that is compatible with the modification data. A possible subdomain is defined as a portion of the structure that contains at least one base pair and does not contain any base pairs to regions outside that portion. A subdomain is considered compatible with the data if there are no strong hits on paired nucleotides that are not at the end of a helix, involved in a G-U pair, or adjacent to a G-U pair. The number of the suboptimal structure with the lowest free energy for a modification compatible subdomain, from nucleotides i to j, is stored in a matrix, $S(i, j)$ with $i < j$, and the calculated energy is stored in a matrix, $E(i, j)$. The program then reconstructs a secondary structure by combining these subdomains in such a way to produce the lowest free energy structure. The first step in this process (Fig. 7) is to search through $E(i, j)$ to determine the combination of subdomains that give the lowest free energy for each fragment i, j. $S(i, j)$ is reset to zero if the lowest free energy structure from i to j is a combination of two or more smaller fragments. The next step is to output the resulting structure using the algorithm shown in Figure 8. The computer code for this program, written in C++, is available on the World Wide Web at the Turner group homepage, http://rna.chem.rochester.edu. This method makes two assumptions. First, it neglects pseudoknots because the free energy minimization al-

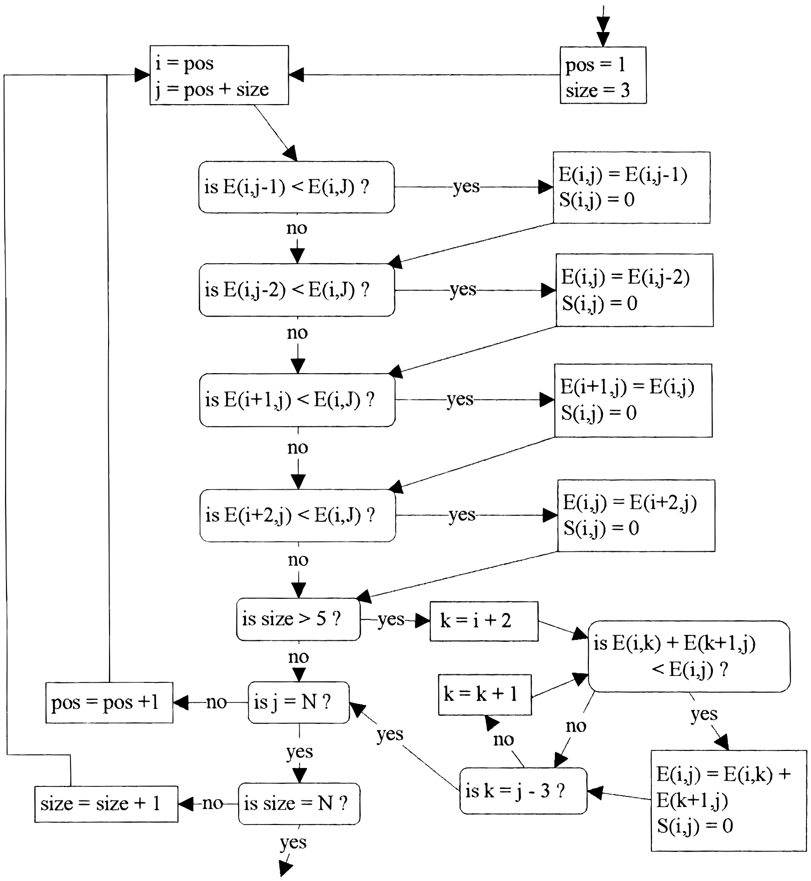

FIGURE 7. Sorting routine from mix&match. This flow chart illustrates the algorithm for finding the combination of modification-compatible domains from suboptimal structures that minimizes total free energy. N is the number of bases in the structure. $E(i,j)$ is the matrix containing the lowest free energy found on a compatible domain, i–j, in a suboptimal structure. $S(i,j)$ is the number of the suboptimal structure in which the lowest free energy domain occurs. $S(i,j)$ is changed to zero in this algorithm if a more favorable combination of subdomains is found.

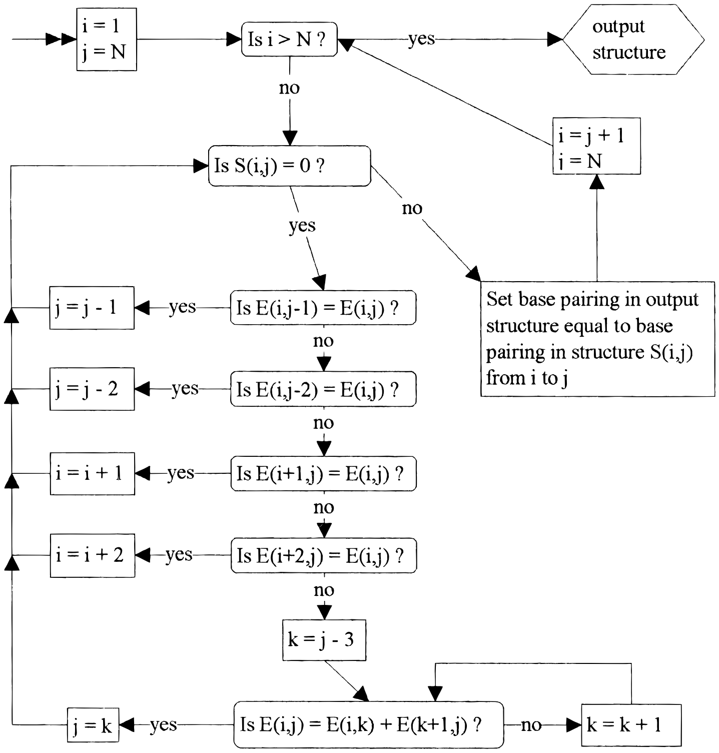

FIGURE 8. Recombination routine from mix&match. To produce the final structure, the scheme of Figure 7 is applied. All variables are as defined in Figure 7.

gorithm does not allow pseudoknots. This is a weakness, but it is important to note that the algorithm correctly predicts a majority of the helices in Group I introns without being able to predict the pseudoknot (Walter et al., 1994). Second, mix&match assumes that the free energy of two domains combined is the sum of the energy of each domain; thus, it does not treat internal loops and multibranch loops rigorously. This is probably not much worse than other assumptions because all the factors that determine free energies of internal and multibranch loops have not been determined (Serra & Turner, 1995). The advantage of including experimental constraints from modification data seems to more than compensate for the reduced rigor in calculation of free energy.

Optical melting curves

Absorbance versus temperature curves for *B. mori* and *D. melanogaster* R2 RNA were measured at 260 and 280 nm with a Gilford spectrophotometer interfaced to a Zenith 250 computer. The buffer was 80 mM Hepes, pH 7.5 (made by adjusting the pH of the free acid with NaOH), 10 mM NaCl (for a total Na^+ concentration of 50 mM), 0.5 mM EDTA, and variable concentrations of $MgCl_2$ from 0.6 to 5.5 mM. Other melts were performed replacing sodium with potassium to simulate intracellular salt conditions. The samples were heated from 0 to 90 °C at a rate of 0.5 °C/min. Derivatives of the melting curves were found by a Savitsky–Golay smoothing routine set to a quadratic and using seven points in a window (Press et al., 1992). The temperatures of transition were identical with 260-nm and 280-nm melts and with sodium buffer and potassium buffer.

Chemical modification

D. melanogaster R2 RNA secondary structure was mapped at 20 and 42 °C and the *B. mori* R2 RNA was mapped at 20 °C using chemical modification as described by Banerjee et al. (1993). Modification reactions were performed with DMS, CMCT, and kethoxal. Reverse transcription was performed with ^{32}P end-labeled primers and the DNA was visualized on a 10% polyacrylamide sequencing gel. Reverse transcriptase dideoxy nucleotide sequencing reactions were run on the same gel to identify the modified bases.

Stock solutions of reagents were prepared immediately before the reaction and kept on ice. The CMCT stock solution was 10.4 mg of CMCT (from Sigma) in 280 μL sterile water. DMS was prepared by adding 2 μL of neat DMS (from Aldrich) to 12 μL 99% ethanol. Kethoxal stock was 300 μL water, 100 μL 99% ethanol, and 1 μL kethoxal (from Upjohn or US Biochemical).

For each reaction, 12.6 pmol of RNA in 24 μL melting buffer was pre-incubated for at least half an hour at the temperature of modification. This handling of the RNA is similar to that used in assays measuring the activity of the enzyme with the RNA (Luan et al., 1993). Temperatures were controlled by immersing the reactions in a circulating water bath. For CMCT and DMS reactions, respectively, 12.5 and 1 μL of stock solution were added. For kethoxal reactions, R2Dm RNA was modified with 2 μL and R2B with 4 μL of

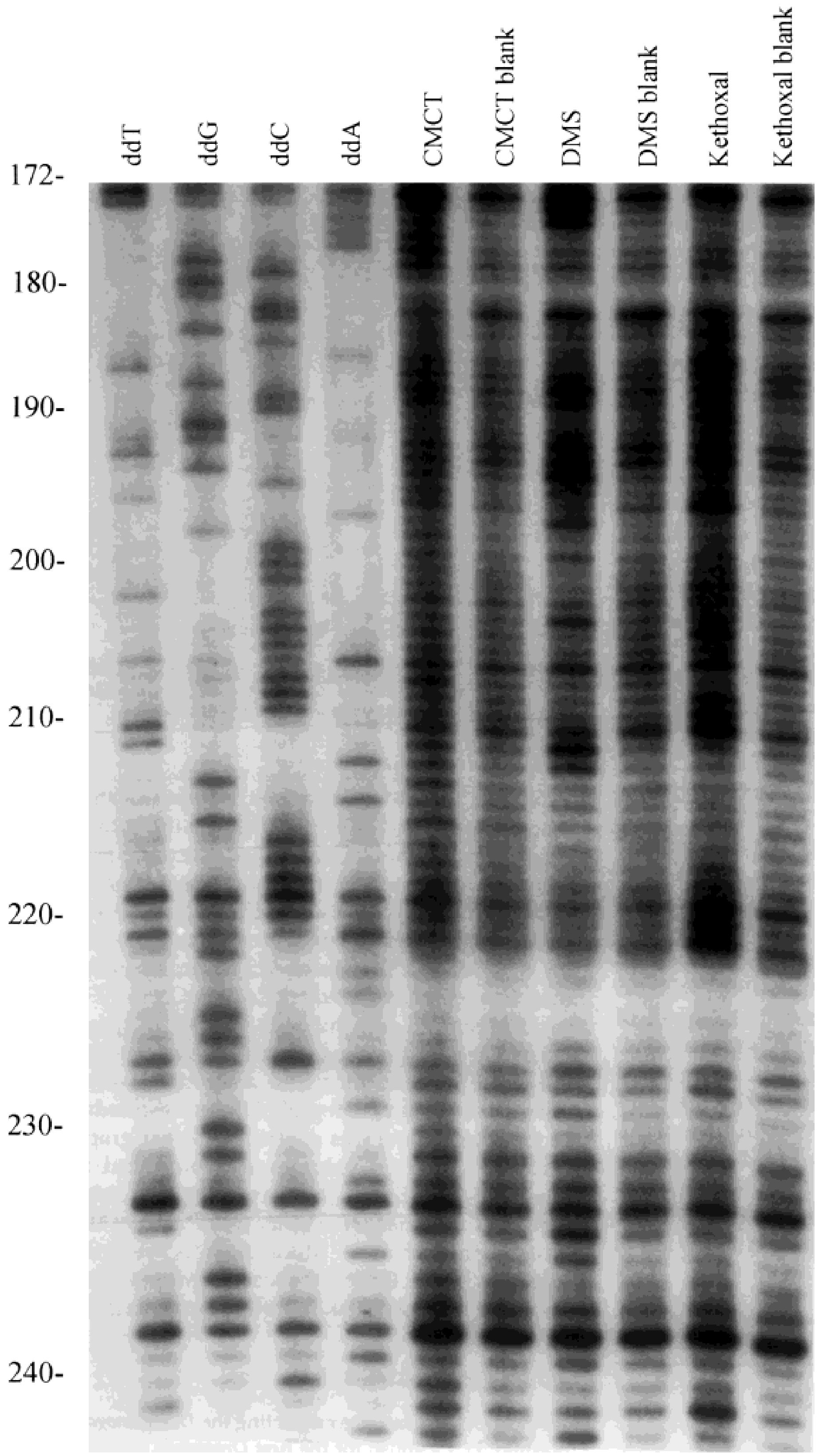

FIGURE 9. Representative gel from chemical modification. It shows the modification experiments for R2Bm from nt 172 to 242. Nucleotides are numbered as in Figure 6. The first four lanes are dideoxy sequencing reactions. Then, from left to right, are CMCT reaction, CMCT control, DMS reaction, DMS control, kethoxal reaction, and kethoxal control. The polymerase stops at the position before a modification, so bands represent modification at the next base in the sequence.

stock solution. Reaction times, summarized in Table 1, were chosen to provide a full range of modifications (weak, moderate, and strong) and still allow full extension of the transcribed DNA. Optimal times of modification for each reagent at each temperature were found by a series of time titrations. The middle point of these titrations was selected on the basis of the reaction kinetics reported by Banerjee et al. (1993), using the Arrhenius equation, $k = A\exp(-E_a/RT)$, to extrapolate a time of reaction.

Controls, identical except lacking the modifying reagent, were run for each reaction. Modification reactions were stopped by adding 4.4 μL of 1.9 μg/μL tRNA in 1.5 M aqueous Na acetate. For kethoxal reactions, 8.2 μL of 0.5 M K borate, pH 7.5, was also added to stabilize the kethoxal adduct (Litt, 1969). The RNA was then precipitated with three volumes of ethanol. After freezing on dry ice for 30 min and 15 min of centrifugation, the pellets were washed with 70% ethanol. Then, 10 μL of sterile water was added, giving an RNA concentration of about 1.2 pmol/μL.

DNA primers for reverse transcription were synthesized on an Applied Biosystems 392 synthesizer using standard methods. These were desalted by *n*-butanol precipitation (Sawadogo & Van Dyke, 1991) and purified by PAGE. Primers were eluted from gel by crush and soak and ethanol precipitated. Primers were chosen to be complementary to three evenly spaced single-stranded regions in the structure. The sequences for *D. melanogaster* were: 5′-AATTCGATATCAAGC, 5′-GGTCAGTTTTCCTAG, and 5′-GCCATTTAATTATAA. For *B. mori*, primers were: 5′-CGAATATTTTCATCG, 5′-CCCACCCTCCCGATC, and 5′-GTTAAAATATATAAG. Primers were labeled with T7 polynucleotide kinase from Gibco-BRL and γ-^{32}P ATP from New England Nuclear.

Modifications were detected by reverse transcription (Inoue & Cech, 1985; Moazed et al., 1986) with 0.5 pmol of 5′ ^{32}P-labeled primer added to 1.2 pmol of RNA and annealed by the procedure of Moazed et al. (1986). Reverse transcription was at 45 °C using AMV enzyme from New England Bio Labs. Reverse transcription for sequencing was at 55 °C with 1 pmol unmodified RNA in the presence of 0.2 mM ddNTP.

Results were visualized on 8% or 10% polyacrylamide (30:1 bis:acrylamide) sequencing gels run for 2–4 h at 60 W. A typical gel is presented in Figure 9.

TABLE 1. Modification times for each reagent at each temperature for *D. melanogaster* and *B. mori* R2 3′ untranslated RNA.

Reagent	*Drosophila* (20 °C)	*Drosophila* (42 °C)	*Bombyx* (20 °C)
Kethoxal	3 h	6 min	2.5 h
CMCT	3 min	20 s	60 min
DMS	3 min	25 s	8 min

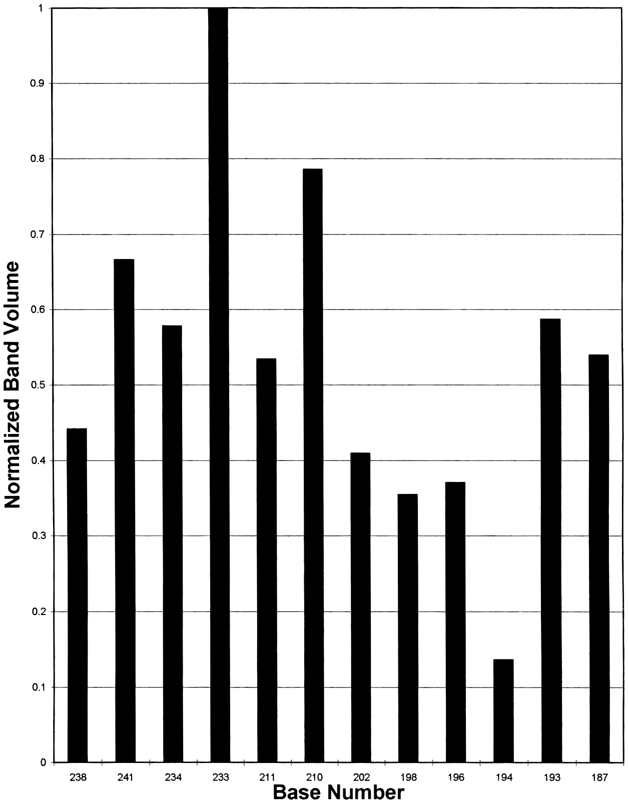

FIGURE 10. Bar graph for quantifying chemical modification by DMS from nt 187 to 238 of R2Bm. Normalized band volumes are calculated as described in Materials and Methods.

Valid hits were those bands in the modification lanes that were visibly darker than those in the control lanes.

For *D. melanogaster*, weak, moderate, and strong hits were approximated from autoradiograms on the basis of how much darker the modification lane was compared to the control. For *B. mori*, hits were quantified on a Molecular Dynamics 425 PhosphorImager running ImageQuant software (Zaug & Cech, 1995). Each band identified as a hit was volume integrated and normalized by dividing by the volume integration of the entire lane. To then find a corrected volume, the normalized volume integral of the corresponding band in the control lane was subtracted. A bar graph of the resulting corrected volumes was produced for each lane (Fig. 10). Strong hits were those bands whose corrected volumes were at least half that of the strongest hit in each lane. Moderate hits were those between 0.3 and 0.5 of the corrected volume of the strongest hit, and weak hits were between 0.1 and 0.3.

ACKNOWLEDGMENTS

We thank W. Burke for constructing the *D. melanogaster* R2 3′ plasmid, D. Eickbush for the first *Drosophila* sequence alignment, W.D. Lathe for unpublished R2 sequences, and S. Moran for his invaluable comments. This work was supported by National Institutes of Health grants GM22939 (D.H.T.)

and GM42790 (T.H.E.). D.H.M. received support from National Institutes of Health grant 5T32 GM07356.

Received July 23, 1996; returned for revision September 3, 1996; revised manuscript received October 23, 1996

REFERENCES

Banerjee AR, Jaeger JA, Turner DH. 1993. Thermal unfolding of a group I ribozyme: The low-temperature transition is primarily disruption of tertiary structure. *Biochemistry 32:153–163.*

Barfod ET, Cech TR. 1988. Deletion of nonconserved helixes near the 3′ end of the rRNA intron of *Tetrahymena thermophila* alters self-splicing but not core catalytic activity. *Genes & Dev 2*:652–663.

Borer PN, Lin Y, Wang S, Roggenbuck MW, Gott JM, Uhlenbeck OC, Pelczer I. 1995. Proton NMR and structural features of a 24-nucleotide RNA hairpin. *Biochemistry 34*:6488–6503.

Burke WD, Catalang CC, Eickbush TH. 1987. The site-specific ribosomal insertion element type II of *Bombyx mori* (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. *Mol Cell Biol 7*:2221–2230.

Burke WD, Müller F, Eickbush TH. 1995. R4, a non-LTR retrotransposon specific to the large subunit rRNA genes of nematodes. *Nucleic Acids Res 23*:4628–4634.

Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, Cech TR, Doudna JA. 1996. Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science 273*:1678–1685.

Crothers DM, Cole PE, Hilbers CW, Shulman RG. 1974. The molecular mechanism of thermal unfolding of *Escherichia coli* formylmethionine transfer RNA. *J Mol Biol 87*:63–88.

Dieckmann T, Suzuki E, Feigon J, Nakamura GK. 1996. Solution structure of an ATP-binding RNA aptamer reveals a novel fold. *RNA 2*:628–640.

Dock-Bregeon AC, Chevrier B, Podjarny A, Johnson J, de Bear JS, Gough GR, Gilham PT, Moras D. 1989. Crystallographic structure of an RNA helix: $[U(UA)_6A]_2$. *J Mol Biol 209*:459–474.

Eickbush DG, Eickbush TH. 1995. Vertical transmission of the retrotransposable elements *R1* and *R2* during the evolution of the *Drosophila melanogaster* species subgroup. *Genetics 139*:671–684.

Eickbush DG, Lathe WC III, Francino MP, Eickbush TH. 1995. R1 and R2 retrotransposable elements of *Drosophila* evolve at rates similar to those of nuclear genes. *Genetics 139*:685–695.

Ehresmann C, Baudin F, Mougel M, Romby P, Ebel J, Ehresmann B. 1987. Probing the structure of RNAs in solution. *Nucleic Acids Res 15*:9109–9128.

Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, Neilson T, Turner DH. 1986. Improved free energy parameters for predictions of RNA duplex stability. *Proc Natl Acad Sci USA 83*:9373–9377.

Gabriel K, Schneider J, McClain WH. 1996. Functional evidence for indirect recognition of G•U in $tRNA^{Ala}$ by alanyl-tRNA synthase. *Science 271*:195–197.

Gautheret D, Konings D, Gutell RR. 1995. G•U base pairing motifs in ribosomal RNA. *RNA 1*:807–814.

Greenbaum NL, Radhakrishnan I, Hirsh D, Patel DJ. 1995. Determination of the folding topology of the S11 RNA from *Caenorhabditis elegans* by multidimensional heteronuclear NMR. *J Mol Biol 252*:314–327.

He L, Kierzek R, SantaLucia J Jr, Walter AE, Turner DH. 1991. Nearest-neighbor parameters for G-U mismatches: $(GU)_2$ is destabilizing in the contexts $(CGUG)_2$, $(UGUA)_2$, and $(AGUU)_2$ but stabilizing in $(GGUC)_2$. *Biochemistry 30*:11124–11132.

Heus HA, Pardi A. 1991. Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science 253*:191–194.

Hilbers CW, Robillard GT, Shulman RG, Blake RD, Webb PK, Fresco R, Riesner D. 1976. Thermal unfolding of yeast glycine transfer RNA. *Biochemistry 15*:1874–1882.

Holbrook SR, Cheong C, Tinoco I Jr, Kim S. 1991. Crystal structure of an RNA double helix incorporating a track of non-Watson–Crick base pairs. *Nature 353*:579–581.

Hou YM, Schimmel P. 1988. A simple structural feature is a major determinant of the identity of transfer RNA. *Nature 333*:140–145.

Inoue T, Cech TR. 1985. Secondary structure of the circular form of the *Tetrahymena* rRNA intervening sequence: A technique for RNA structure analysis using chemical probes and reverse transcriptase. *Proc Natl Acad Sci USA 82*:648–652.

Jacobson H, Stockmayer WH. 1950. Intramolecular reaction in polycondensations. I. The theory of linear systems. *J Chem Phys 18*:1600–1606.

Jaeger JA, SantaLucia J Jr, Tinoco I Jr. 1993a. Determination of RNA structure and thermodynamics. *Annu Rev Biochem 62*:255–287.

Jaeger JA, Turner DH, Zuker M. 1989. Improved predictions of secondary structures for RNA. *Proc Natl Acad Sci USA 86*:7706–7710.

Jaeger JA, Turner DH, Zuker M. 1990a. Predicting optimal and suboptimal secondary structure for RNA. *Methods Enzymol 183*:281–306.

Jaeger JA, Zuker M, Turner DH. 1990b. Melting and chemical modification of a cyclized self-splicing group I intron: Similarity of structures in 1 M Na^+, in 10 mM Mg^{2+}, and in the presence of substrate. *Biochemistry 29*:10147–10158.

Jaeger L, Westhof E, Michel F. 1993b. Monitoring of cooperative unfolding of the sunY group I intron of bacteriophage T4. *J Mol Biol 234*:331–346.

Jakubczak JL, Burke WD, Eickbush TH. 1991. Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. *Proc Natl Acad Sci USA 88*:3295–3299.

Jakubczak JL, Xiong Y, Eickbush TH. 1990. Type I (R1) and type II (R2) ribosomal DNA insertions of *Drosophila melanogaster* are retrotransposable elements closely related to those of *Bombyx mori*. *J Mol Biol 212*:37–52.

James BD, Olsen GJ, Pace NR. 1989. Phylogenetic comparative analysis of RNA secondary structure. *Methods Enzymol 180*:227–239.

Jiang F, Kumar RA, Jones RA, Patel DJ. 1996. Structural basis or RNA folding and recognition in an AMP–RNA aptamer complex. *Nature 382*:183–186.

Jucker FM, Pardi A. 1995. Solution structure of the CUUG hairpin loop: A novel RNA tetraloop motif. *Biochemistry 34*:14416–14427.

Kim SH, Suddath FL, Quigley GJ, McPherson A, Sussman JL, Wang AHJ, Seeman NC, Rich A. 1974. Three dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science 185*:435–440.

Konings DAM, Gutell RR. 1995. A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. *RNA 1*:559–574.

Konings DAM, Hogeweg P. 1989. Pattern analysis of RNA secondary structure, similarity and consensus of minimal-energy folding. *J Mol Biol 207*:597–614.

Laing LG, Draper J. 1994. Thermodynamics of RNA folding in a conserved ribosomal RNA domain. *J Mol Biol 237*:560–576.

Litt M. 1969. Structural studies on transfer ribonucleic acid. I. Labeling of exposed guanine sites in yeast transfer ribonucleic acid with kethoxal. *Biochemistry 8*:3249–3253.

Luan DD, Eickbush TH. 1995. RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. *Mol Cell Biol 15*:3882–3891.

Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. *Cell 72*:595–605.

Lück R, Steger G, Riesner D. 1996. Thermodynamic prediction of conserved secondary structure: Application to the RRE element of HIV, the tRNA-like element of CMV and the mRNA of prion protein. *J Mol Biol 258*:813–826.

McClain WH, Foss K. 1988. Changing the identity of a tRNA by introducing a G•U wobble pair near the 3′ acceptor end. *Science 240*:793–796.

Michel F, Westhof E. 1990. Modeling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J Mol Biol 216*:585–610.

Moazed D, Stern S, Noller HF. 1986. Rapid chemical probing of conformation in 16S ribosomal RNA and 30S ribosomal subunits using primer extension. *J Mol Biol 187*:399–416.

Press WH, Teukolsk SA, Vetterling WT, Flannery BP. 1992. *Numerical recipes in fortran, the art of scientific computing, 2nd ed*. New York: Cambridge University Press.

Puglisi JD, Chen L, Blanchard S, Frankel AD. 1995. Solution structure of a bovine immunodeficiency virus TAT-Tar peptide–RNA complex. *Science 270*:1200–1203.

Robertus JD, Ladner JE, Finch JT, Rhodes D, Brown RS, Clark BFC, Klug A. 1974. Structure of yeast phenylalanine tRNA at 3 Å resolution. *Nature 250*:546–551.

SantaLucia J Jr, Turner DH. 1993. Structure of $(\text{rGGC}\underline{\text{GA}}\text{GCC})_2$ in solution from NMR and restrained molecular dynamics. *Biochemistry 32*:12612–12623.

Sawadogo M, Van Dyke MW. 1991. A rapid method for the purification of deprotected oligodeoxynucleotides. *Nucleic Acids Res 79*:674.

Serra MJ, Turner DH. 1995. Predicting thermodynamic properties of RNA. *Methods Enzymol 259*:242–261.

Strobel SA, Cech TR. 1995. Minor groove recognition of the conserved G•U pair at the *Tetrahymena* ribozyme reaction site. *Science 267*:675–679.

Szewczak AA, Moore PB. 1995. The sarcin/ricin loop, a modular RNA. *J Mol Biol 247*:81–98.

Turner DH, Sugimoto N, Freier SM. 1988. RNA structure prediction. *Annu Rev Biophys Chem 17*:167–192.

Varani G, Cheong C, Tinoco I Jr. 1991. Structure of an unusually stable RNA hairpin. *Biochemistry 30*:3280–3289.

Varani G, Tinoco I Jr. 1991. RNA structure and NMR spectroscopy. *Q Rev Biophys 24*:479–532.

Walter AE, Turner DH, Kim J, Lyttle MH, Müller P, Mathews DH, Zuker M. 1994. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc Natl Acad Sci USA 91*:9218–9222.

Westhof E, Sundaralingam M. 1986. Restrained refinement of the monoclinic form of yeast phenylalanine transfer RNA: Temperature factors and dynamics, coordinated waters, and base pair propeller twist angles. *Biochemistry 25*:4868–4878.

Williams AP, Longfellow CE, Freier SM, Kierzek R, Turner DH. 1989. Laser temperature-jump, spectroscopic, and thermodynamic study of salt effects on duplex formation by dGCATGC. *Biochemistry 28*:4283–4291.

Xiong Y, Eickbush TH. 1988. Functional expression of a sequence-specific endonuclease encoded by the retrotransposon R2Bm. *Cell 55*:235–246.

Yang Y, Kochogan M, Burgstaller P, Westhof E, Famulok M. 1996. Structural basis of ligand discrimination by two related RNA aptamers resolved by NMR spectroscopy. *Science 272*:1343–1347.

Ye XM, Kumar RA, Patel DJ. 1995. Molecular recognition in the bovine immunodeficiency virus TAT peptide Tar RNA complex. *Chem & Biol 2*:827–840.

Zaug AJ, Cech TR. 1995. Analysis of the structure of *Tetrahymena* nuclear RNAs in vivo: Telomerase RNA, the self-splicing rRNA intron, and U2 snRNA. *RNA 1*:363–374.

Zuker M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science 244*:48–52.

Zuker M, Jaeger JA, Turner DH. 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.