Secondary structure models of the 3′ untranslated regions of diverse R2 RNAs

AMY M. RUSCHAK,[1] DAVID H. MATHEWS,[3,4] ARKADIUSZ BIBILLO,[2] SHERRY L. SPINELLI,[1,3]
JESSICA L. CHILDS,[1] THOMAS H. EICKBUSH,[2] and DOUGLAS H. TURNER[1,3]

[1]Department of Chemistry and [2]Department of Biology, University of Rochester, Rochester, New York 14627-0216, USA
[3]Center for Human Genetics and Molecular Pediatric Disease, Aab Institute of Biomedical Sciences, and [4]Department of Biochemistry and Biophysics, University of Rochester School of Medicine and Dentistry, Rochester, New York 14642, USA

ABSTRACT

The RNA structure of the 3′ untranslated region (UTR) of the R2 retrotransposable element is recognized by the R2-encoded reverse transcriptase in a reaction called target primed reverse transcription (TPRT). To provide insight into structure–function relationships important for TPRT, we have created alignments that reveal the secondary structure for 22 *Drosophila* and five silkmoth 3′ UTR R2 sequences. In addition, free energy minimization has been used to predict the secondary structure for the 3′ UTR R2 RNA of *Forficula auricularia*. The predicted structures for *Bombyx mori* and *F. auricularia* are consistent with chemical modification data obtained with β-ethoxy-α-ketobutyraldehyde (kethoxal), dimethyl sulfate, and 1-cyclohexyl-3-(2-morpholinoethyl)carbodiimide metho-p-toluene sulfonate. The structures appear to have common helices that are likely important for function.

Keywords: secondary structure; 3′ UTR; R2 RNA; retrotransposon

INTRODUCTION

RNA secondary structure provides a foundation for discovery of structure–function relationships in RNA and for identifying recurring elements of structure. To date, firmly established secondary structures are mainly available for various classes of structural and enzymatic RNAs (Michel and Westhof 1990; Gutell et al. 1993; Gutell 1994; Brown 1998; Larsen et al. 1998; Sprinzl et al. 1998; Szymanski et al. 1998; Chen et al. 2000a; Li et al. 2002). It is possible that expanding the existing database to other RNAs will reveal new motifs or expand the functional significance of known motifs.

Non-long terminal repeat (non-LTR) retrotransposons are among the most abundant components of eukaryotic genomes. For example, over 1 million copies of the L1 and L2 elements constitute ~20% of the human genome, and the retrotransposition machinery of these elements is believed responsible for the insertion of short interspersed nucleotide elements and processed pseudogenes constituting another 15% of human DNA (Lander et al. 2001). The mechanism of non-LTR retrotransposition is best understood for R2, elements that insert into a unique location in the 28S rRNA genes of insects (Eickbush 2002). As shown in Figure 1A, the critical step in R2 retrotransposition is sequence-specific cleavage of one strand of the 28S gene target site, and the use of the released 3′ hydroxyl to serve as primer for the reverse transcription of the R2 transcript. This step, termed target primed reverse transcription (TPRT; Luan et al. 1993), is believed to be used by all other non-LTR retrotransposons (Chaboissier et al. 2000; Cost et al. 2002). TPRT is also the mechanism used in retrohoming by group II introns (Belfort et al. 2002).

R2 elements have been identified in all investigated lineages of arthropods, and their sequence phylogeny suggests they have been vertically inherited throughout the 500-million-year history of arthropods (Burke et al. 1998; Malik et al. 1999). The R2 TPRT reaction in *Bombyx mori* requires the RNA sequence corresponding to the 250-nucleotide 3′ untranslated region (3′UTR) of the R2 element (Luan et al. 1993; Luan and Eickbush 1995). Previously, the R2-encoded protein from *B. mori* was also found to conduct TPRT by using RNA from the R2 element of *Drosophila melanogaster* (Mathews et al. 1997). Because the 3′ UTR sequences of *B. mori* and *D. melanogaster* have little primary sequence identity, functional 3′ UTR sequences must have common secondary and tertiary structural features. The secondary

Reprint requests to: Douglas H. Turner, Department of Chemistry, University of Rochester, Rochester, NY 14627–0216, USA; e-mail: turner@chem.rochester.edu; fax: (585) 506-0205.

Article and publication are at http://www.rnajournal.org/cgi/doi/10.1261/rna.5216204.

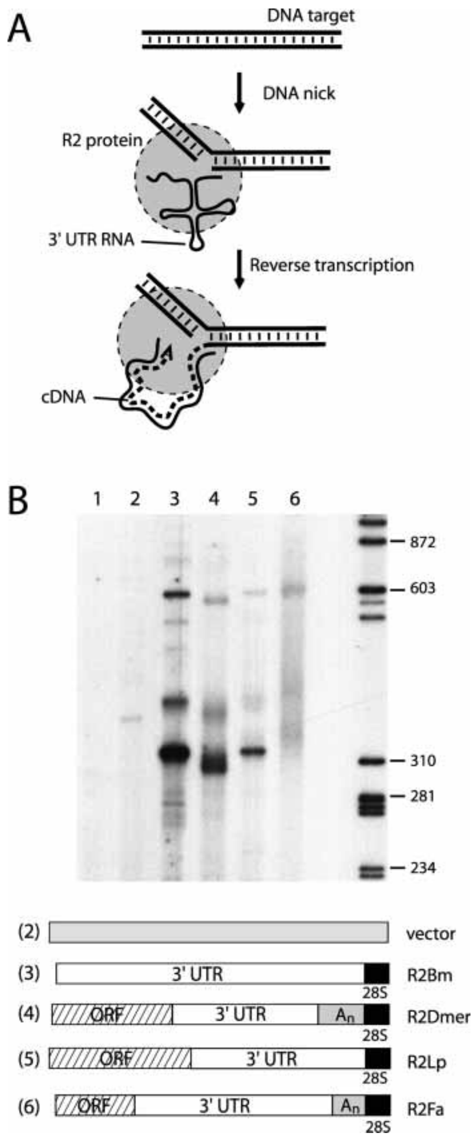

FIGURE 1. The R2Bm TPRT reaction and assay. (*A*) Diagram of the target primed reverse transcription (TPRT) reaction. The R2-encoded protein (gray circle) binds the 28S rRNA gene target site and the endonuclease domain cleaves the bottom strand. The reverse transcriptase domain of the R2 protein binds RNA corresponding to the 3′ UTR of the element by means of its secondary structure. The RNA is positioned opposite the DNA cleavage such that the reverse transcriptase can use the nick to initiate the formation of cDNA (dotted line). (*B*) Autoradiograph of the reaction products of a TPRT reaction using 3′ UTR sequences from four different R2 elements. The DNA target in this reaction was 164 bp in length and ^{32}P-labeled at the 5′ end of the *bottom* strand. The cleavage site is located 60 bp from that end. The various RNAs used in the reaction are diagrammed at the *bottom*. The open box corresponds to the 3′UTR; a gray shaded area at the end of this region corresponds to a poly(A) (A_n) tail. Some of the 3′ UTRs are <250 nt in length; thus the end of the ORF is included in the template (diagonally shaded areas). At the 3′ end of all R2 RNAs are 20 nt of 28S rRNA sequences. Reverse transcription starts at the junction between the R2 sequences and this 28S rRNA sequence. Numbers to the *left* of each diagram correspond to the numbered lanes in the autoradiogram. (R2Bm) R2 element from *B. mori*; (R2Dmer) R2 element from *D. mercatorum*; (R2Lp) R2 element from *L. polyphemus*; (R2Fa) R2 element from *F. auricularia*. Lane *1* is a control without RNA.

structures of the 3′ UTRs for R2 elements from 10 closely related sequences from the genus *Drosophila* and for the sequence from *B. mori* were deduced by sequence comparison, free energy minimization, and chemical modification (Mathews et al. 1997).

The 3′ UTR sequences of R2 elements from another 28 arthropod species are now available (Lathe and Eickbush 1997; Burke et al. 1999). Here, we demonstrate that the R2 protein from *B. mori* can use in the TPRT reaction R2 RNA from even the most evolutionarily divergent arthropod. Comparative sequence analyses in conjunction with free energy minimization identify conserved structural elements. Four additional 3′ UTR sequences from silkmoth R2 elements reveal covariation to refine the secondary structure of the *B. mori* R2 sequence itself. Finally, these structural predictions were tested experimentally by chemical modification of the R2 RNAs from *B. mori* and *Forficula auricularia* (earwig).

RESULTS

3′ UTR RNAs with no primary sequence similarity are recognized by the same R2 protein

We have previously described sensitive in vitro TPRT assays for the R2 element found in *B. mori* (Yang and Eickbush 1998; Bibillo and Eickbush 2002). In these assays, short labeled DNA substrates containing the 28S gene target site are incubated with purified *B. mori* R2 protein (hereafter referred to as R2Bm) and template RNA. Sequence-specific cleavage of the first DNA strand, use of the free DNA end to prime cDNA synthesis, and second-strand cleavage readily occur in the presence of 25 μM dNTPs, 0.2 M NaCl, 10 mM $MgCl_2$ (pH 7.5; Fig. 1A). After incubation, TPRT products are detected on denaturing polyacrylamide gels as labeled single-stranded DNA molecules that correspond to the combined length of the cleaved DNA strand and RNA template. Figure 1B shows the results of such a reaction with 3′ UTR sequences from four R2 elements that represent the wide range of arthropod R2 diversity. Each RNA template contains ~250 nt from the 3′ end of the R2 element as well as 20 nt from the flanking downstream 28S gene sequences. Reverse transcription is initiated at the junction between the R2 and 28S gene sequences, not at the 3′ end of the RNA (Luan and Eickbush 1996; Bibillo and Eickbush 2002).

Lane 3 in Figure 1B corresponds to the TPRT products obtained with the 3′ UTR sequence from the same R2 element that supplied the protein (R2Bm). A band at ~315 nt was observed corresponding to the labeled DNA downstream of the cleavage site (60 bp) and the R2 sequences within the RNA template (255 nt). Two additional bands were also detected in this assay. The band at ~365 nt was a result of the ability of the R2 protein to use the 3′ end of the full-length upper DNA strand to prime reverse transcription (110 nt DNA target + 255 nt R2 sequences). The nature of this product was confirmed by PCR reactions, using primers specific to either orientation of the DNA substrate and to the RNA sequence (data not shown). The band at ~590 nt resulted from the R2 protein synthesizing to the end of a first RNA molecule and then jumping to a second RNA molecule, where it continued reverse transcription (60 nt DNA substrate + 255 nt first RNA template + 275 nt of the second RNA template; see Bibillo and Eickbush 2002 for more detailed studies of these template jumps).

Lanes 4–6 in Figure 1B correspond to TPRT products generated with RNA templates containing the 3′ UTR of sequences from the *Drosophila mercatorum* R2 element (R2Dmer, Fig. 1B, lane 4), the *Limulus polyphemus* (horseshoe crab) R2 element (R2Lp, Fig. 1B, lane 5), and the *F. auricularia* (earwig) R2 element (R2Fa, Fig. 1B, lane 6). The 3′ UTRs of these other R2 elements are shorter than that of R2Bm, and the R2Dmer and R2Fa elements end in a poly(A) tail (Lathe and Eickbush 1997; Burke et al. 1999). The three distant R2 RNAs supported TPRT from the cleavage site (bands at ~310 nt), TPRT from the 3′ end of the upper strand (bands at ~360 nt), and template jumping to a second RNA (bands at ~590 nt). The other R2 templates supported TPRT at relative levels of 63% (R2Dmer), 18% (R2Lp), and 36% (R2Fa) compared with the R2Bm template. As a control, lane 2 represents the use of a non-R2 template, a 285-nt sequence from pBluescript vector. Only low levels of TPRT products were obtained (2% of the R2Bm level) and did not involve specific recognition of an RNA structure, because reverse transcription began at the 3′ end of the RNA (60 nt target + 285 nt), not from an internal site as observed with the R2 templates.

The products obtained with both R2Dmer and R2Fa are not of a precise length because these R2 elements end in poly(A) tails ~30 nt in length. Labeled cDNA products of variable lengths were generated because TPRT initiates at variable positions within such poly(A) tails, and the reverse transcriptase is able to add additional residues to the DNA target by slippage before firmly engaging the RNA (Luan and Eickbush 1996). For example, most of the R2 Dmer TPRT products resulted from initiation within the poly(A) tail, whereas the R2Fa products were highly variable in length with most containing longer tails, generated by slippage. These results indicate that the R2Bm protein is able to recognize the RNA derived from the 3′ UTR of diverse R2 elements and use it in the TPRT reaction. Because these various 3′ UTRs are of variable length and exhibit no readily observed primary sequence identities, we conclude that all R2 RNAs assume secondary or tertiary structures with common features that can be recognized by the R2Bm protein.

Secondary structure predictions of R2 3′ UTR

The most extensive collection of R2 3′ UTR sequences has been obtained from the *Drosophila* genus (22 species). We previously compared R2 sequences from 10 closely related species within the *Sophophora* subgenus (Mathews et al. 1997). The 3′ UTRs of those elements were ~250 nt in length and can be aligned and folded into a similar structure. Four additional R2 sequences from more divergent species within this subgenus (*D. bipectinata*, *D. willistoni*, *D. sucinea*, and *D. persimilis*) can be readily aligned with these previous sequences (Lathe and Eickbush 1997; data not shown).

The 3′ UTR of R2 elements from species within the *Drosophila* subgenus are on average only ~150 nt in length. They can be aligned with each other (Fig. 2), but exhibit little sequence similarity to the R2 elements from the *Sophophora* subgenus. The *D. mercatorum* R2 element within this group (R2Dmer) was used for the TPRT reactions in Figure 1. Figure 3 shows the proposed secondary structure for an example from the *Sophophora* group, *D. maritiana* (Panel A) and from the *Drosophila* group, *D. nasuta* (Panel B). Covariational analysis supports each labeled helix in the *Drosophila* group; that is, each labeled helix exhibits covariation, providing a minimum of two compensating changes (Pace et al. 1989). Most labeled helices in the *Sophophora* subgroup are also supported by compensating base changes and those helices that are not (helices B, C, and D) have highly conserved sequences. Helix A, adjacent

```
Drosophila subgenus
                                                                            A                 A                    B
  nasuta       (nas) UAGCA----CCGAGGUUUACUUA-CCUCAUU----------GAAA--GUCAACUAC-ACUGGC-CAACAUCUAUU-GU-GUACAAAG-UAAAGGGAAA
  mercatorum   (mer) CAGCU------CUAUUUUACGUAUCCUAC------------AAAA-CGCCUGCUAACUU-GGCACAACAUCUUAAAAAGUUAUAAAG--AAAGCAC-A
  buzzatii     (buz) CA-CUAUAUUA----------------CAUUUCCUUUC---AAAA--GCUUGCUACCUU-GGCACAACAUCUUAAGAAGGUACAAAG-AAAU-CGG-A
  sordidula    (sor) UAGGAAACUAUG--UUUUACAAG------------UUCCUCUAAUUUGCUUGCUACCUU-GGCAUAACAUCUAAAAAAGGUACAAAGU-AAUCCGGGA
  robusta      (rob) UAGGAAACUAUG---UUUACGAG------------UUCCUUUAACUUGCUUGCUACCUU-GGCAUAACAUCUAAU-AAGGUACAAAGU-AAUCCGGGA
  euronotus    (eur) UA-GAA-CUAUU---UUUGCAG-------------UUCCUAAAACUUGCUUGCUACCUU-AGCAUAAGAGCCAUAAAA-GUACAAAGU-AAUCCGGA-
  falleni      (fal) UGAAUGU--AUUGGCUAUCAAUAUUGCAUUUCCUUUUCCUAUAAA--GCCAGCUACCUUUGGCAUAACAUCUAAA-AAGGUAUAAAA--UAACGCAU-
  recens       (rec) U-UACAAUAUUACGUUU---------------CUUUUCUAUAAAC--GCCUACUACCUU-GGCAUAACAUC-AAA-AAGGUAUAAA--GUAACGCAUA

                               H         I            I       H                                       B
Nas   A---------------UAGUCA-UAUGAC--GUGCCAU-CCAUAU-AUGGCAUUGUUGUG-UGGGUAAAUUCA--------------------UUUCCU----ACAAA
Mer   CGGGU----------------GUCUUAGC-U-UAUCAGAUAUU-CUCUGGUAUAGC-GAGUCAUAUA----------------------------GUGC---AACAAA
Buz   CCGGUCCGGACGCAUGCCGGUCGUCUUUA-UCUAUCAGAAAUU-C-CUGAUAGGUGAGCAGCGAGCCAGACAAUUGGACCGU-------------CCG--U----AAA
Sor   UCGGUCCGGACGCAGACAGU--CGCUCUAACCUAUCAGAAAUCAC-CUGGUAGGUGGAGCGUAAGGCUUCUGGACGGACCAGUCAAUGCGACACUCCGG-UU---AAA
Rob   UCGGUCCGGACGCAGACAGU--CGCUCUAACCUAUCAGAAACUAC-CUGGCAGGUGGAGCGUAAGGCUUCUGGACGGACCAGACAAUGCGACACUCCGG-UU---AAA
Eur   CC-GUC--GACGCAGACGGU--CACUCUAA-CUAUCAGAUCUCUC-CUGGUAGGUGGAGUGUAAGA-GCUGGACCAGAC-----AAUGCGACACUCCGG-UUU--AAA
Fal   ---------------UCAGUGGCGCUACAA--UCGAAA---UCA--UUUCGUUAUGUGGUG--AACUA--------------------------AUGUGCUA--CAAA
Rec   ---------------U-AGUUGCUCCACAA---UCGGACUCUCAU-UCCGAC--UGUGGG--U-GCUA-------------------------AUGUGCUACAGAAA
```

FIGURE 2. RNA sequence alignment and secondary structures of *Drosophila* 3′ UTR R2 elements from eight species of the *Drosophila* subgenus. Underlined nucleotides are in helices as labeled by letters *above*. Boldfaced nucleotides are in positions of compensating base changes.

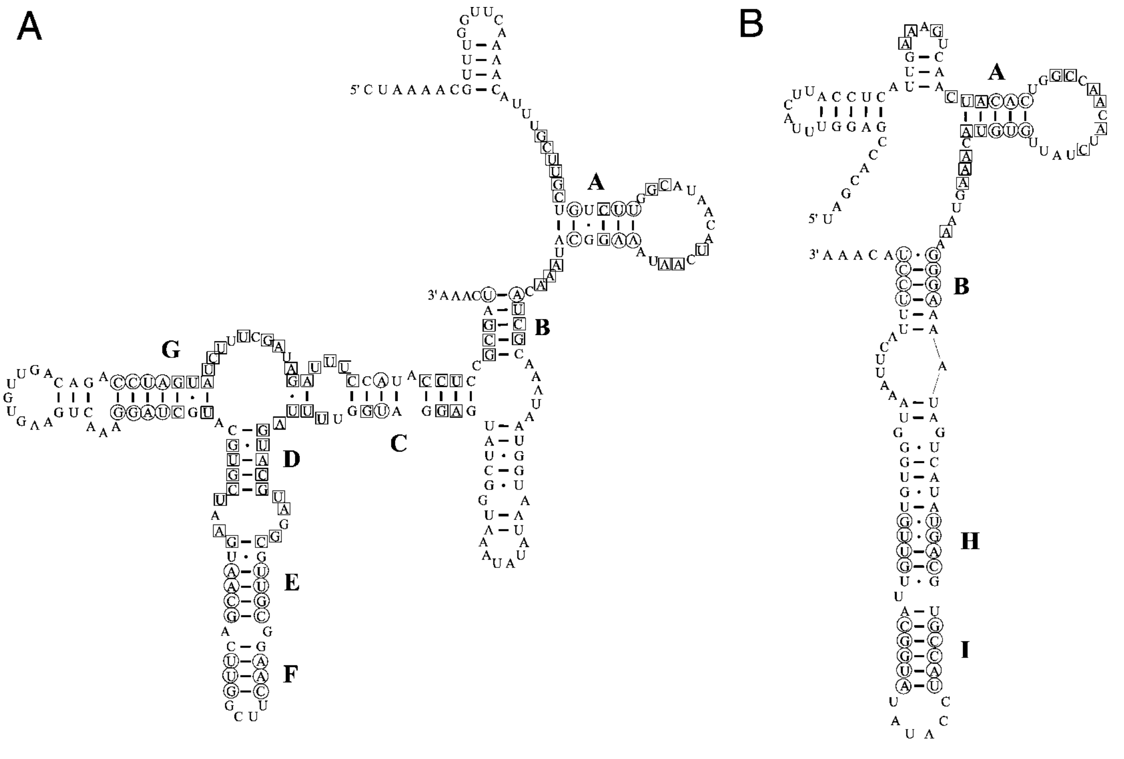

FIGURE 3. Secondary structures of the 3′ UTR R2 elements from *D. maritiana* (*A*) and *D. nasuta* (*B*). Nucleotides at positions of compensating base changes are circled and nucleotides conserved throughout an alignment are boxed. Conserved helices in *B* are lettered as shown in the alignment in Figure 2.

to the exterior loop and closing a large hairpin loop (14–16 nt) and helix B, the closest helix to the 3′ end, are apparently conserved in both subgenera. The *Sophophora* RNA structures have two multibranch loops, as demonstrated previously (Mathews et al. 1997), but the *Drosophila* group has no multibranch loops in its secondary structure. With the available data, it is unclear as to which set of helices of the *Sophophora* group, C, D, E, F, and G, are homologous to helices H and I of the *Drosophila* group.

Because the R2-encoded protein from *B. mori* is used in most studies of the TPRT reaction, the structure of the 3′ UTR of *B. mori* is the most important to understand in order to provide a foundation for structure–function studies. In our previous report (Mathews et al. 1997), however, we were not able to use covariational analysis to test the secondary structure prediction for the 3′ UTR of the R2Bm element because no other R2 element had primary sequence similarity to that of *B. mori*. We therefore cloned and sequenced the 3′ UTR of R2 elements obtained from four other species of silkmoths: *Samia cynthia*, *Callosamia promethea*, *Coscinocera hercules*, and *Saturnia pyri* (see Friedlander et al. 1998 for the relationship of these species to *B. mori*). The sequence alignments are shown in Figure 4 and the structures of *C. promethea* and *B. mori* are shown in

```
                                   1                   2           3
Bombyx mori          (BM)   GCCUUGCACAGUAGUCCAGCGGUAAGGG-UGUAGAUCAGGCCC------GUCUGUUUCUCCCCCGGAGCUCGCUCC----------CU
Coscinocera hercules (CH)   UCCGGUCAUCCU-GUAGGUUCUUUCGGGCUGC-CGUCGGGCCC-UA---GGUUGGUUUUGAG--CGCAGUCAU-UGA--GUACUGGUU
Saturnia  pyri       (SP)   UCCGAAGAUCCU-GU--GU------GGGCUGC-CGCCGGGCCCGUUCUCGUGGCUUUCGAA---CGCAGUCA---GAUUGUACUGG-U
Samia cynthia        (SC)   UCCGAUCAUCCU-GUUGGUU-UCUCGGGCUGC-CGUCGAGCCC-UA---GGCUGGUUUCGAG--CGCAGCCAUGGGAU-GUACUGAUU
Callosamia promethea (CP)   UCCGAAGAUCCU-GUGAGU------GGGUUGC-CGCCGAGCCCUUAA--CGGUGGUUUCGAG--CGCAGUCAUGGGAUUAUACUGAUU

            3                             2                      1          4                                5
BM      U-GGCUUCCCUUAUAUAUUUUAACAUC--AGAAACAGACA-UUAAACAUCU--ACUGAUCCAAUUUCGCCGGCGUACGGCCAC-----------GA-UCG-GGAGGGUG
CH      UAGCUGCGAAUGGUGCCGUUUAACAUCUUAGA-CC-GCCG-UUUACCAUC--AACCGGCGCAAU-UAGUGACUGC---CCUA-GGUACUGGUUGUGG-CACCGGG-----
SP      GAGCUGCGGAGUGUGCCAUCCUACAUC--UGAAGGCCACAACUUAACAUCUUAACCGGCGCAAU--GUUGACU-----CCCG--CAGAUGGU---GA-UAG-UGGC----
SC      UAGCUGUGAAAGGUGCCGUUUUUUAUCU-GAAACCGGCCG-UUCAACAUC--AAUCGGCGCAAU-UAGUGACUGC---CCUUAGGUAC-------GG-GUG-GAGG----
CP      UAACUGCGAACGGUGCCAUUUAACAUCU-GGAACCGCCG--CCGACCAUCU--AUCGGCGUAAUGUAGUGACUGC---CCUUCUA----------GGUACG-GGUAGG--

                              5                                                                    4
BM      GGAAUCUCGGGGGUCUUC-CGAUCCUAAUCCAUGAUGAUUACGACC---------UGA-GUCACUAAAG-AC--GA-------UGGCAUGAUGAUCCGGCGAUGAAAAA
CH      UUCAUCUG-------UUC-GUGUC---------------------------------GUUGGUAAAGCACU-GAC------GAGGC--GUG--GUCGCUAU-AACAA
SP      -----ACAGG----GUUAUCUAUCCCAUCUGAGUUAAUCUGUAAUGAUUAUAGCGAAGCCUCGUGAAAUCACUAGAUGGGGCAUGGGA----A--GUCGGC-UUAACAA
SC      ----CACCGGG-UUUCUC-CAUUC----------------------------AUGUCGUCGGU-AUGCACU-GAC------GGGGC--GUG--GUCGCU-UGAAAAA
CP      --GGCGCCGGG-UUUAUC-CGUUC----------------------------AUGACGUCGGUGAAGCACU-GAC------GGGGC--GUG--GUCGCUACUAACAA
```

FIGURE 4. RNA sequence alignment and secondary structures of five silkmoth R2 3′ UTR RNA sequences. Helices are underlined and labeled *above* with numbers. Boldfaced nucleotides are in positions of compensating base changes.

Figures 5 and 6, respectively. Each of the five conserved helices found in the silkmoth sequences is supported by compensating base changes (Fig. 4). The structure for *B. mori* based on this comparative analysis differs significantly from the Mathews et al. (1997) model. The previous model, based only on free energy minimization and chemical modification, included only helices 2 and 4, with all other helical regions involving different sequences. The model shown in Figure 6 and the previous model are, however, consistent with the previously published chemical modification data. Because the four new silkmoth species could not be folded into the structure proposed by Mathews et al. (1997), our findings confirm the strength of combining covariational analysis with free energy minimizations (Chen et al. 2000b; Mathews and Turner 2002).

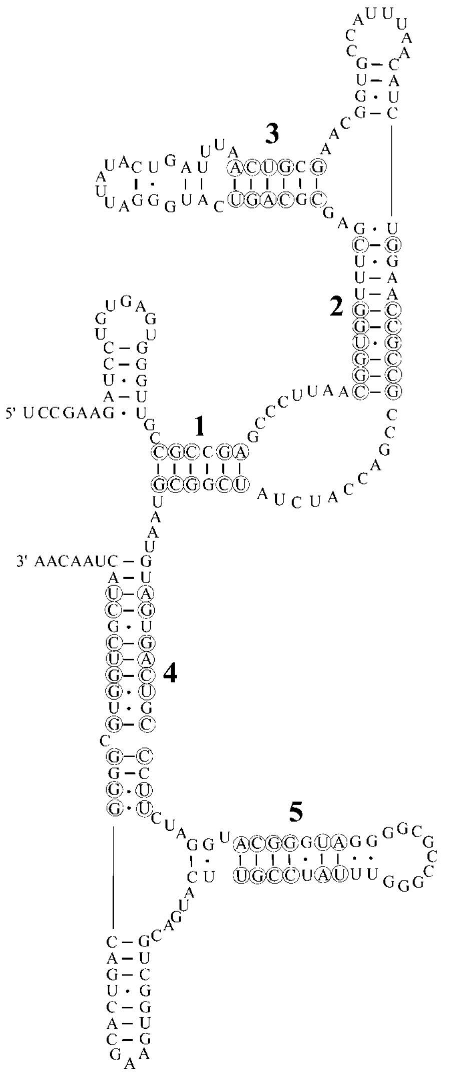

FIGURE 5. The secondary structure of the *C. promethea* R2 3′ UTR RNA. Helices are numbered according to the sequence alignment shown in Figure 4. Circled nucleotides are at positions of compensating base changes.

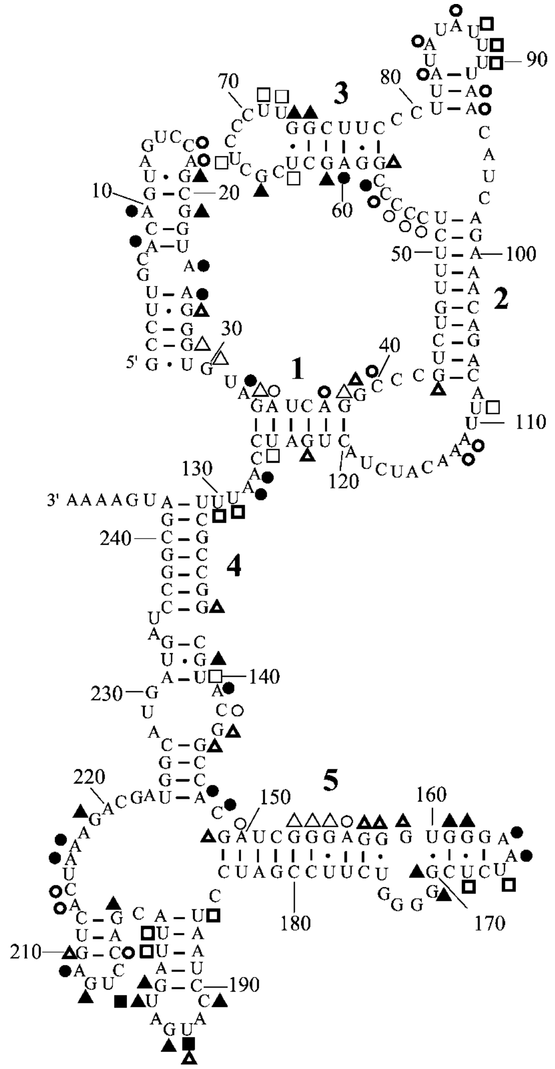

FIGURE 6. The secondary structure of the *B. mori* R2 3′ UTR RNA. Conserved helices are numbered as shown in the alignment in Figure 4. Note that helix 4 extends from nucleotides 131–147 and 224–242, but helix 5 only extends from nucleotides 149–158 and 175–184. Helices without numbers are predicted by free energy minimization. Chemical modification data are superimposed: kethoxal with triangles, CMCT with squares, and DMS with circles. Solid symbols indicate strong modification, darkly outlined symbols indicate moderate modification, and weakly outlined symbols indicate weak modification.

Chemical modification of the 3′ UTR of R2 RNA from *B. mori* and *F. auricularia*

The R2 RNA from *B. mori* used in our previous study (Mathews et al. 1997) contained an additional 50-nt vector sequence at its 5′ end. To ensure that this vector sequence did not cause the RNA to misfold, we made a new construct that eliminated these sequences from the RNA used for chemical modification. Also differing from the previous set of experiments, the RNA was chemically modified in the same buffer used for the TPRT reactions, and native gel electrophoresis was conducted after RNA folding to ensure that the RNA was in a single conformation before modification.

An identical approach was used to map the R2 sequence from *F. auricularia*, an element that contains one of the shortest 3′ UTR sequences (165 nt excluding the poly(A) tail). On the basis of the sequence of its encoded protein, the R2 element from *F. auricularia* is highly divergent from R2Bm (Burke et al. 1998), yet its 3′ UTR RNA sequence supports the TPRT reaction catalyzed by the R2Bm protein (Fig. 1). The secondary structure of this R2 3′ UTR (Fig. 7) was predicted only on the basis of free energy minimization (Zuker 1989; Mathews et al. 1999) because this sequence could not be reliably aligned with any other R2 sequences. Chemical mapping was therefore important as a means of testing the proposed structure because free energy minimization is, on average, only 73% accurate at predicting canonical base pairs (Mathews et al. 1999).

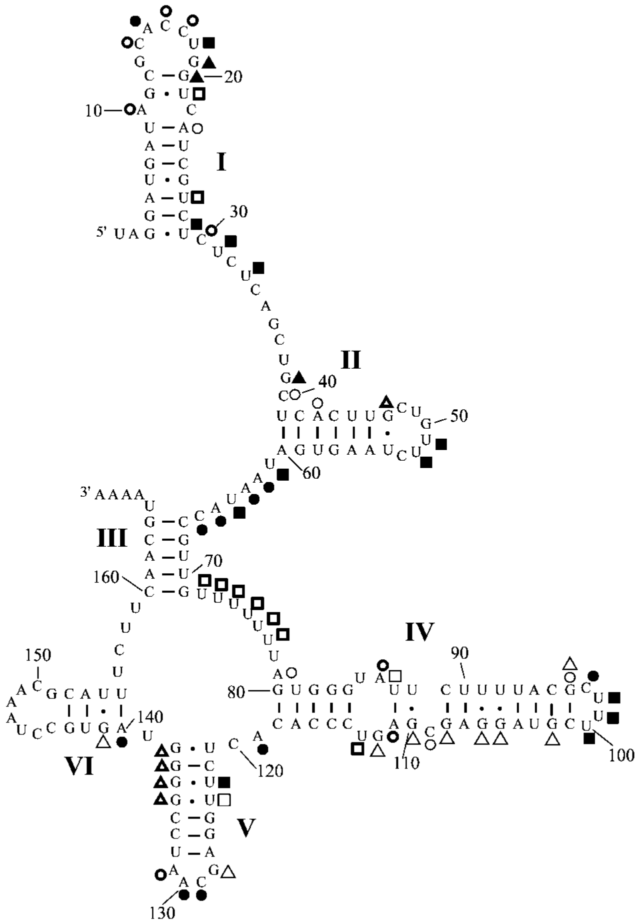

FIGURE 7. The R2 3′ UTR RNA secondary structure predicted for *F. auricularia* (predicted $\Delta G°_{37} = -44.2$ kcal/mole). Chemical modification data are superimposed: kethoxal with triangles, CMCT with squares, and DMS with circles. Solid symbols indicate strong modification, darkly outlined symbols indicate moderate modification, and weakly outlined symbols indicate weak modification. Helices are labeled with Roman numerals. An alternative folding consistent with the chemical modification data and predicted to be less stable by 3.5 kcal/mole pairs nucleotides 1–13 with 67–79 and leaves nucleotides 157–165 unpaired.

The 3′ UTRs of R2 RNA from *B. mori* and *F. auricularia* were modified with β-ethoxy-α-ketobutyraldehyde (kethoxal), 1-cyclohexyl-3-(2-morpholinoethyl)carbodiimide metho-p-toluene sulfonate (CMCT), and dimethyl sulfate (DMS), and the results are shown in Figures 6 and 7. In evaluating whether chemical modification data support a proposed structure, only the strong and medium hits were considered because weak hits may result from a small portion of the RNAs being in an alternative, minor conformation. Modified bases are considered to be consistent with the proposed structure if they are not Watson–Crick paired or are adjacent to a nucleotide not in a Watson–Crick pair (Moazed et al. 1986).

The chemical modification data for *B. mori* R2 RNA are consistent with helices 2, 4, and 5 (Fig. 6). Helix 3 is inconsistent with the strong kethoxal modification at G61. This modification, however, is consistent with a pairing of helix 3 that is slipped by one nucleotide to give 5′GGAG<u>C</u>UCG/3′CCUU<u>C</u>GGU, a helix with seven canonical base pairs and a CC mismatch (underlined). This slippage is not possible for the other moth sequences in Figure 4; therefore, it is probably not functionally important. Helix 1 is inconsistent with the moderate kethoxal modification at G122 and the moderate DMS modification at A37. Helix 1, however, is supported by compensating base pairs (Fig. 4). The moderate chemical modifications may reflect unexpected dynamics for this helix in vitro. The stability of this helix is not reliably predicted because highly asymmetric internal loops of the type closed by helix 1 have not been studied. The chemical modification data for the *F. auricularia* R2 RNA are highly consistent with the predicted secondary structure shown in Figure 7, thus providing support for this proposed secondary structure.

The chemical modification data obtained in this study of *B. mori* R2 RNA are similar to that obtained by Mathews et al. (1997). The data differed mostly in the intensity of the modifications. For example, C53–55 are strongly modified by DMS according to Mathews et al. (1997), but in this study they are only weakly modified. G170–174 were strongly modified by kethoxal according to Mathews et al. (1997), but in this study only G170 and 171 are strongly modified. Some of the differences may reflect changes in tertiary structure because the buffer used in the studies presented here had higher Na^+ and Mg^{2+} concentrations in order to more closely match the TPRT reaction conditions.

The structure for *B. mori* R2 RNA proposed by Mathews et al. (1997) and the one proposed in this study are consistent with both sets of chemical modification data, assuming that a base can be modified if it is not in a Watson–Crick pair or is adjacent to a nucleotide not in a Watson–Crick pair. The four new silkmoth sequences cannot be folded into the structure proposed by Mathews et al. (1997), however. On the basis of primary sequence identity between

B. mori and the other silkmoth sequences and the compensating changes supporting the new structure, we suggest the structure in Figure 6 is the best model for the secondary structure of this RNA. Only helices 2 and 4 were present in the earlier model based on chemical modification data and the predicted free energies of folding (Mathews et al. 1997). Evidently, the promiscuity of chemical modification reagents and of base pairing make it difficult to determine a secondary structure on the basis of a single sequence.

DISCUSSION

Molecular recognition involving the R2 element is particularly interesting because R2 inserts at unique sites in a genome. This function is potentially useful for applications such as gene therapy (Wickelgren 2003). One goal of studying the structure of the 3′ UTR of R2 RNA is to determine which structural elements are recognized by the R2-encoded reverse transcriptase. Four secondary-structure models from divergent classes of R2 RNA 3′ UTR sequences are proposed in this report. The *Drosophila* genus has two subgenera with different structures, as proposed on the basis of comparative sequence analysis. The structural model for the sequences from the *Sophophora* subgenus (Fig. 3A) is largely unchanged by the addition of four new sequences. This group contains the *D. melanogaster* sequence, which was previously tested by chemical modification (Mathews et al. 1997). The secondary structure for sequences from the *Drosophila* subgenus (Fig. 2) is illustrated in Figure 3B and is based only on comparative sequence analysis and free energy minimization. The high levels of sequence diversity within this group and the resultant extensive covariation detected for all labeled helical regions provide considerable support for this structure. The newly proposed silkmoth R2 secondary structure model is based on comparative sequence analysis and free energy minimization, and was tested by chemical modification (Fig. 6). Finally, a structure for the divergent sequence from *F. auricularia* was predicted by free energy minimization and is supported by chemical modification (Fig. 7).

All four sets of RNA structures are used by the reverse transcriptase encoded by the R2 element of *B. mori* (Fig. 1; and Fig. 2 in Mathews et al. 1997). The ability of the R2 reverse transcriptase to recognize the structures in Figures 3, A and B, and 7, even though it evolved to recognize the structure in Figure 6, lays testament to our poor understanding of how proteins are able to recognize RNA structures.

What might be the structural features of RNA recognized by the R2 reverse transcriptase? The ability of the R2 reverse transcriptase to use an RNA template for the TPRT reaction involves two steps: the ability of the R2 protein to recognize and bind the RNA, and the ability of the protein to position the 3′ end of that RNA (or the junction of the 3′ UTR and the downstream 28S gene sequences) opposite the nicked DNA strand for the initiation of cDNA synthesis (see Fig. 1A). The ability of the RNA to be positioned opposite the nicked site does not involve annealing of the RNA with the downstream DNA sequences (Luan and Eickbush 1996). It is not known whether RNA recognition and RNA positioning involve the same or different motifs.

The different classes of R2 3′ UTR secondary structures do have several common features. Invariably, the exterior loop (which contains the 5′ and 3′ ends of the sequence) has at least two conserved helices. It is easy to speculate that the helix closest to the 3′ end of the sequence (helix B in Fig. 3; helix 4 in Figs. 5, 6; and helix III in Fig. 7) serves a conserved role in each structure. The importance of this helix at the 3′ end is consistent with deletion experiments showing that the 3′ end of the sequence of *B. mori* is important for efficient TPRT (Luan and Eickbush 1995). A second invariably conserved feature of all four structures is a helix (helix A in Fig. 3; helix 1 in Figs. 5, 6; and helix II in Fig. 7) separated from the first by a short sequence highly enriched in A and U nucleotides. Again, deletion of the first 50 nt from the 5′ end of the R2Bm RNA, which deletes this helix, completely eliminates the TPRT reaction (Luan and Eickbush 1995). Because these are the only two helices clearly present in all of the RNA structures used by the R2Bm protein, they likely provide two recognition elements. As shown in Figure 1, however, different RNAs provide different specificities in the TPRT reaction catalyzed by the R2Bm protein. Thus, it is likely that species-specific recognition elements also exist. Because it is not currently possible to match up definitively the recognition elements between different classes of R2 secondary structures, helices in each class have been labeled differently.

Each RNA structure also contains an extensive region enclosed by the conserved helix close to the 3′ end (helices C–G in Fig. 3A; helices H and I in Fig. 3B; helix 5 in Figs. 5, 6; and helices IV–VI in Fig. 7). Often, this is most of the sequence outside the two conserved helices. When sequence comparison is possible, this region appears conserved within a subgenus. Thus, this region may be the species-specific recognition element for the reverse transcriptase. Future mutagenesis experiments using both the R2Bm and the R2Dmer structures and tests for both binding and the ability to function as a template for TPRT should help resolve the significance of the various RNA structures.

There are other potential recognition elements not indicated in the alignments and secondary structures of Figures 2–7. For example, in each class there is a potential pseudoknot that can form between the unpaired region enclosed by the 5’ invariably conserved helix feature (helix A in Fig. 3; helix 1 in Figs. 5, 6; and helix II in Fig. 7) and the sequence 5’ of this conserved helix. There is not enough sequence variation within each class, however, to prove such a pseudoknot, and in many cases the potential upstream pairing partners are able to base pair with nucleotides even further upstream without forming a pseudoknot.

It will be interesting to compare the structure and function of the R2 RNA and protein with other reverse transcriptase complexes. Although some other non-LTR retrotransposons specifically recognize the 3′ UTR of the elements (Takahashi and Fujiwara 2002), many non-LTR retrotransposons apparently do not. For example, if a strong SV40 polyadenylation signal is added well downstream of the weak polyadenylation site of the L1 element, the L1 reverse transcriptase will efficiently initiate TPRT at the poly(A) tail of the new 3′ end (Moran et al. 1996). Perhaps more interesting comparisons will involve the similarities of the R2 RNA and reverse transcriptase with those of mobile group II introns and of telomerase. Similarities with telomerase include the association of protein with a specific RNA template and the ability to position one segment of this RNA template opposite a free DNA end for reverse transcription. In the case of telomerase, however, the RNA template is held rigidly enough so that only a short region of the RNA is reverse transcribed (Lingner et al. 1997). In the case of R2, all of the RNA template is eventually stripped away from the protein by the formation of the RNA:cDNA heteroduplex. In this regard, the R2 enzyme is more similar to that of group II introns, in that after binding a specific RNA template, the RNA template is completely reverse transcribed (Belfort et al. 2002). It will be interesting to compare the structures and properties of these three very ancient types of reverse transcriptase complexes, as well as what enables some non-LTR elements to specifically bind RNA templates and others to use many RNA templates.

MATERIALS AND METHODS

Sequence alignment

Alignments were created for the *Drosophila* and silkmoth RNA sequences by arranging them phylogenetically and aligning them by primary sequence similarity, using free energy minimization (Mathews et al. 1999) as a guide. Highly conserved anchor regions were also found with the help of Dynalign (Mathews and Turner 2002). The following criteria were used for choosing the sequence alignments shown in Figures 2 and 4: (1) columns represent homologous positions throughout all sequences; (2) canonical base pairs in conserved helices are maximized; (3) compensating changes in conserved helices are minimized; (4) sequence conservation outside of helices is maximized; (5) gaps are minimized; (6) helices are considered proven when they contain at least three compensating base changes.

TPRT assays

Synthesis of labeled DNA target sites and purification of the R2Bm protein and the TPRT assays were conducted as described in Bibillo and Eickbush (2002).

Synthesis of RNA

RNA for the TPRT assays was made from PCR templates as described in Bibillo and Eickbush (2002). Primers ~250 bp from the 3′ end of each element were as follows: *L. polyphemus*, CTGCAGTAATACGACTCACTATAGGTCAAAATATTTTGAAGAGTTCTTG; *F. auricularia,* CTGCAGTAATACGACTCACTATAGGATGTTTAATTCAACAACCTCAG; *D. mercatorum*, CTGCAGTAATACGACTCACTATAGGTGACAGCAATGTTATCAGTTC; and *B. mori*, CTGCAGTAATACGACTCACTATAGGTTGAGCCTTGCACAGTAG. Each of these primers was used in combination with the primer for the 28S gene sequence immediately downstream of the R2 insertion site, GATGACGAGGCATTTGGCTA.

The *B. mori* and *F. auricularia* R2 element 3′ UTR regions were also cloned into pUC19. For *B. mori*, the oligonucleotide primer, CGGGATCCTAATACGACTCACTATAGGCCTTGCACAGTAGTCCAGCGG, which contains the T7 promoter, and CGGGCTGCAGGAATTCGA were used to PCR amplify the clone micro-R2–28S (Eickbush et al. 2000). For *F. auricularia,* the primer CGGGGATCCTAATACGACTCACTATAGGATGATAGCGCACCTGGTC, again containing the T7 promoter, and GCTGCAGAATTTTTTCGTTGAAGAAATGCG were used to PCR amplify one of the M13 clones used in the sequencing of the 3′ end of the *F. auricularia* R2 element (Burke et al. 1999). The amplified DNA was digested with BamH1 and PstI and individually cloned into these same sites of pUC19. T7 transcription of these plasmids digested with XmnI allowed the synthesis of RNA corresponding to the complete 3′ UTR of each element, starting within the termination codon of the R2 ORF and ending with a run of four A residues.

The plasmids were linearized via digestion with XmnI (Promega), followed by a phenol:chloroform extraction and ethanol precipitation. The DNA was then transcribed with T7 RNA polymerase by using an Ambion MEGAScript in vitro transcription kit, followed again by a phenol:chloroform extraction and ethanol precipitation. The manufacturer's protocol was used with a plasmid concentration of 200 ng/μL instead of 50 ng/μL and an incubation time of 7.5 h instead of 2 h. Next, the RNA was gel purified on a denaturing, 8% polyacrylamide, 8 M urea gel. To elute the RNA, we excised the bands containing full-length RNA, found by UV shadow, and crushed them and soaked them in water. The stock was desalted by running the RNA through Sephadex G-25 columns (Amersham Biosciences).

Native gel electrophoresis

Before conducting mapping experiments, native gel electrophoresis was used to establish conditions under which the RNA is in a single conformation. The RNA at 0.42 μM was renatured in TPRT buffer (40 mM Hepes, 40 mM sodium Hepes, 110 mM NaCl, 10 mM $MgCl_2$) by (1) incubating for 2 min at 90°C and slow cooling to 37°C; (2) incubating for 30 min at 45°C and slow cooling to room temperature; (3) incubating for 45 min at 37°C; or (4) only thawing at room temperature. After renaturation, glycerol was added to give a 10% solution, and the samples were run on 4% Aquapor High Resolution Agarose gels (National Diagnostics) for 2 h at 50 V at room temperature. The RNA was viewed under UV light after staining the gel with ethidium bromide. Both *B. mori* and *F. auricularia* RNAs ran as a single band under all renaturation conditions tested, including only thawing.

Chemical modification

RNA was chemically modified with CMCT, DMS, and kethoxal. CMCT functions better at pH 8.2; therefore, native gel electrophoresis was used to verify that increasing the pH of the buffer did not change the structure of the RNAs. Stock solutions for CMCT consisted of 10.4 mg CMCT (Aldrich) in 280 µL sterile water; for DMS, 2 µL neat DMS (Aldrich) in 12 µL 99% ethanol; and for kethoxal, 1 µL kethoxal (ICN Biomedicals) in 10 µL 99% ethanol and 30 µL sterile water. Samples of 3 pmole RNA in 7.2 µL TPRT buffer were incubated for 45 min at 37°C prior to chemical modification. Modifications were initiated by adding either 3.75 µL, 0.15 µL, or 0.75 µL of the CMCT, DMS, or kethoxal stock solutions, respectively. Because it is desirable to have one modification per molecule, the reaction time for each modification was determined experimentally. Optimal incubation times were 10–25 min for CMCT, 15–25 min for DMS, and 4–15 min for kethoxal. Control reactions were also run, where chemical modification reagents were not added. All reactions were carried out at 37°C (the same temperature used for TPRT reactions). Reactions were quenched by adding 8.36 µg bulk RNA and 2.2 µL of 3 M sodium acetate, and they were subsequently ethanol precipitated.

Modifications were detected by reverse transcription. Primers were synthesized by standard methods on an Applied Biosystems 392 synthesizer. Bases on the primers were deprotected by incubation in 1 mL NH_4OH at 55°C overnight. Ammonia was removed by placing the samples under a vacuum, and the primers were desalted through Sep-Pak cartridges. Primers for *B. mori* were CGCCGGATCATCATCATGC, CATGGATTAGGATCGG, CCCCCGAGATTCCCACCC, CGAAATTGGATCAGTAG, and GGGAAGCCAAGGGAGC. For *F. auricularia,* primers were CGT TGAAGAAATGCG, CCCCGGATTGCTCC, CAACGGTATTAT CAC, and CGGTATTATCACTTAG.

For reverse transcription reactions, the primers were labeled with T4 polynucleotide kinase (Invitrogen) and γ-^{32}P ATP (Perkin-Elmer Life Sciences). Excess ATP was removed with Chroma Spin +STE-10 columns. Next, 2.5 pmole of radiolabeled primer was annealed to 1 pmole RNA in 0.435 M sodium acetate in a total volume of 8 µL by incubating the reaction mixture for 10 min at 70°C, 10 min at room temperature, and then 10 min on ice. After spinning down the samples, primer extension was carried out in 1 mM each dNTP, 1× reaction buffer, and 1 U AMV reverse transcriptase (Life Sciences) in a total volume of 25 µL. For sequencing lanes, 0.75 mM each ddNTP was also added. The reaction mixture was incubated for 1 h at 42°C. Reactions were quenched by adding 25 µL of stop buffer (12 M urea, 12 mM EDTA).

Samples were run on 8% polyacrylamide, 8 M urea gels for 2–3 h at 60–70 watts. The gel was dried and visualized by autoradiography. The strengths of the chemical modification hits were visually estimated.

ACKNOWLEDGMENTS

We thank Dr. Matthew Disney for help with gel electrophoresis. This work was supported by the University of Rochester Chemistry REU Program (NSF 9987991), the GEBS Summer Scholars Program, and NIH grants GM22939 (D.H.T.) and GM42790 (T.H.E.). D.H.M. was a trainee in the University of Rochester Medical Scientist Training Program (NIH grant 5T32 GM07356). J.L.C. was partially supported by National Institutes of Health grant T32 DE07202.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

Received October 22, 2003; accepted March 10, 2004.

REFERENCES

Belfort, M., Derbyshire, V., Parker, M.M., Cousineau, B., and Lambowitz, A.M. 2002. Mobil introns: Pathways and proteins. In *Mobile DNA II* (eds N.L. Craig et al.), pp. 761–783. American Society for Microbiology, Washington, DC.

Bibillo, A. and Eickbush, T.H. 2002. The reverse transcriptase of the R2 non-LTR retrotransposon: Continuous synthesis of cDNA on non-continuous RNA templates. *J. Mol. Biol.* **316:** 459–473.

Brown, J.W. 1998. The ribonuclease P database. *Nucleic Acids Res.* **26:** 351–352.

Burke, W.D., Malik, H.S., Lathe, III, W.C., and Eickbush, T.H. 1998. Are retrotransposons long-term hitchhikers? *Nature* **392:** 141–142.

Burke, W.D., Malik, H.S., Jones, J.P., and Eickbush, T.H. 1999. The domain and structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. *Mol. Biol. Evol.* **16:** 502–511.

Chaboissier, M.C., Finnegan, D., and Bucheton, A. 2000. Retrotransposition of the I factor, a non-long terminal repeat retrotransposon of *Drosophila*, generates tandem repeats at the 3′ end. *Nucleic Acids Res.* **28:** 2467–2472.

Chen, J.L., Blasco, M.A., and Greider, C.W. 2000a. Secondary structure of vertebrate telomerase RNA. *Cell* **100:** 503–514.

Chen, J., Le, S., and Maizel, J.V. 2000b. Prediction of common secondary structures of RNAs: A genetic algorithm approach. *Nucleic Acids Res.* **28:** 991–999.

Cost, G.J., Feng, Q., Jacquier, A., and Boeke, J.D. 2002. Human L1 element target-primed reverse transcription in vitro. *EMBO J.* **21:** 5899–5910.

Eickbush, T.H. 2002. R2 and related site-specific non-long terminal repeat retrotransposons. In *Mobile DNA II* (eds. N.L. Craig et al.), pp. 813–835. American Society for Microbiology, Washington, DC.

Eickbush, D.G., Luan, D.D., and Eickbush, T.H. 2000. Integration of *Bombyx mori* R2 sequences into the 28S ribosomal RNA genes of *Drosophila melanogaster*. *Mol. Cell. Biol.* **20:** 213–223.

Friedlander, T.P., Horst, K.R., Regier, J.C., Mitter, C., Peigler, R.S., and Fang, Q.Q. 1998. Two nuclear genes yield concordant relationships within Attacini (Lepidoptera: Sayurniidae). *Mol. Phylogenet. Evol.* **9:** 131–140.

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

Gutell, R.R, Gray, M.W., and Schnare, M.N. 1993. A compilation of large subunit (23 S- and 23 S-like) ribosomal RNA structures. *Nucleic Acids Res.* **21:** 3055–3074.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Larsen, N., Samuelsson, T., and Zwieb, C. 1998. The signal recognition particle database (SRPDB). *Nucleic Acids Res.* **26:** 177–178.

Lathe, III, W.C. and Eickbush, T.H. 1997. A single lineage of R2 retrotransposable elements is an active, evolutionarily stabile component of the *Drosophila* rDNA locus. *Mol. Biol. Evol.* **14:** 1232–1241.

Li, X., Frank, D.N., Pace, N., Zengel, J.M., and Lindahl, L. 2002. Phylogenetic analysis of the structure of RNase MRP RNA in yeasts. *RNA* **8:** 740–751.

Lingner, J., Hughes, T.R., Shevchenko, A., Mann, M., Lundblad, V., and Cech, T.R. 1997. Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science* **276:** 561–567.

Luan, D.D. and Eickbush, T.H. 1995. RNA template requirements for

DNA-primed reverse transcription by the R2 retrotransposable element. *Mol. Cell. Biol.* **15:** 3882–3891.

———. 1996. Downstream 28S gene sequences on the RNA template affect choice of primer and the accuracy of initiation by the R2 reverse transcriptase. *Mol. Cell. Biol.* **16:** 4726–4734.

Luan, D.D., Korman, M.H., Jakubczak, J.L., and Eickbush, T.H. 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.

Malik, H.S., Burke, W.D., and Eickbush, T.H. 1999. The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* **16:** 793–805.

Mathews, D.H. and Turner, D.H. 2002. Dynalign: An algorithm for finding the secondary structure common to two sequences. *J. Mol. Biol.* **317:** 191–203.

Mathews, D.H., Banerjee, A.R., Luan, D.D., Eickbush, T.H., and Turner, D.H. 1997. Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. *RNA* **3:** 1–16.

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

Michel, F. and 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., and Noller, H.F. 1986. Rapid chemical probing of conformation in 16S ribosomal RNA and 20S ribosomal subunits using primer extension. *J. Mol. Biol.* **187:** 399–416.

Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D., and Kazazian, H.H. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* **87:** 917–927.

Pace, N.R., Smith, D.K., Olsen, G.J., and James, B.D. 1989. Phylogenetic comparative analysis and the secondary structure of ribonuclease P RNA—A review. *Gene* **82:** 65–75.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., and Steinberg, S. 1998. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **26:** 148–153.

Szymanski, M., Specht, T., Barciszewska, M.Z., Barciszewski, J., and Erdmann, V.A. 1998. 5S rRNA data bank. *Nucleic Acids Res.* **26:** 156–159.

Takahashi, H. and Fujiwara, H. 2002. Transplantation of target specificity by swapping the endonuclease domains of two LINEs. *EMBO J.* **21:** 408–417.

Wickelgren, I. 2003. Spinning junk into gold. *Science* **300:** 1646–1649.

Yang, J. and Eickbush, T.H. 1998. RNA-induced changes in the activity of the endonuclease encoded by the R2 retrotransposable element. *Mol. Cell. Biol.* **18:** 3455–3465.

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