Conserved features of Y RNAs revealed by automated phylogenetic secondary structure analysis

A. Darise Farris, Gerald Koelsch, Ger J. M. Pruijn[1], Walther J. van Venrooij[1] and John B. Harley*

Oklahoma Medical Research Foundation, Oklahoma City, OK 73104, USA and [1]Department of Biochemistry, University of Nijmegen, Nijmegen, The Netherlands

Received October 8, 1998; Revised December 10, 1998; Accepted December 21, 1998

DDBJ/EMBL/GenBank accession nos[+]

ABSTRACT

Y RNAs are small 'cytoplasmic' RNAs which are components of the Ro ribonucleoprotein (RNP) complex. The core of this complex, which is found in the cell nuclei of higher eukaryotes as well as the cytoplasm, is composed of a complex between the 60 kDa Ro protein and Y RNAs. Human cells contain four distinct Y RNAs (Y1, Y3, Y4 and Y5), while other eukaryotes contain a variable number of Y RNA homologues. When detected in a particular species, the Ro RNP has been present in every cell type within that particular organism. This characteristic, along with its high conservation among vertebrates, suggests an important function for Ro RNP in cellular metabolism; however, this function has not yet been definitively elucidated. In order to identify conserved features of Y RNA sequences and structures which may be directly involved in Ro RNP function, a phylogenetic comparative analysis of Y RNAs has been performed. Sequences of Y RNA homologues from five vertebrate species have been obtained and, together with previously published Y RNA sequences, used to predict Y RNA secondary structures. A novel RNA secondary structure comparison algorithm, the suboptimal RNA analysis program, has been developed and used in conjunction with available algorithms to find phylogenetically conserved secondary structure models for YI, Y3 and Y4 RNAs. Short, conserved sequences within the Y RNAs have been identified and are invariant among vertebrates, consistent with a direct role for Y RNAs in Ro function. A subset of these are located wholly or partially in looped regions in the Y3 and Y4 RNA predicted model structures, in accord with the possibility that these Y RNAs base pair with other cellular nucleic acids or are sites of interaction between the Ro RNP and other macromolecules.

INTRODUCTION

The Ro ribonucleoprotein (RNP) complex consists of the 60 kDa Ro protein, which binds one of four human RNA molecules (Y RNAs), the 52 kDa Ro protein which does not directly bind Y RNAs, but appears to be associated with 60 kDa Ro by protein interactions, and the La protein which binds the poly(U) tails of Y RNAs in a subset of Ro RNPs (1–3).

Four distinct Y RNAs, each with unique sequences (termed hY1, hY3, hY4 and hY5) are immunoprecipitated from nucleated human cells with antibodies to the 60 kDa Ro protein. These RNAs range from 85 to 112 nucleotides (nt) in length, contain no modified bases, and are products of RNA polymerase III transcription (1,4–6).

The 60 kDa Ro protein has been characterized in mammals, amphibians and nematodes (7–10), and is detected with anti-60 kDa Ro antibodies in birds and fish (11). The La protein, which functions as a terminator of RNA polymerase III transcription, is also highly conserved, having been identified in amphibians, insects and yeast, as well as in mammals (12–15). The 52 kDa Ro protein appears to be less well conserved, being detected with antibody reagents only in human and simian tissues (16).

Like the Ro RNP protein components, Y RNAs have been detected in various vertebrates by immmunoprecipitation with anti-60 kDa Ro antibodies; however, the number of Y RNA homologue types present in a particular species varies (1,10,17–22). This may indicate overlapping functions for these RNAs in cells or some specialized function(s) for the non-conserved Y RNAs in particular species. The presence of the Y5 RNA appears to be least conserved, while Y3 is the most conserved Y RNA (10,21,22). Sequences of iguana, mouse, frog and *Caenorhabditis elegans* Y RNAs are known (9,10,22,23). The latter species has only one Y RNA molecule, which may be a Y3 homologue.

The 60 kDa Ro binding site on human Y RNAs has been identified by ribonuclease protection experiments to be at the 5′ and 3′ ends of the RNAs, which are homologous among the different Y RNAs (24). Complementarity between the ribonuclease protected sites suggests the presence of a homologous terminal stem in each of the Y RNA secondary structures. Though indirect evidence has suggested that one 60 kDa protein binds one Y RNA, biochemical purification studies have identified at least three subpopulations of Ro RNP particles based on structural heterogeneity (3,24). One particle contains hY5 RNA, one contains only hY4 RNA, and one includes hY1, hY3 and hY4 RNAs. Like the 60 kDa Ro protein, Y RNAs are found to reside in both the nucleus and cytoplasm, and not all Y RNAs in a given cell are complexed with 60 kDa Ro protein at a given time (25).

*To whom correspondence should be addressed. Tel: +1 405 271 7768; Fax: +1 405 271 4110; Email: harley@omrf.uokhsc.edu

[+]U82125, U82128, U82129, U84671 and U84678

A definitive function for the Ro RNP complex has not been identified, though some functions have been suggested. The finding that 60 kDa Ro protein unassociated with any Y RNA binds mutant ribosomal RNAs in *Xenopus laevis* oocytes has led to the suggestion that 60 kDa Ro facilitates the discard of mutant cellular ribosomal RNAs (26,27). Recently, antibody to 60 kDa Ro protein was found to immunoprecipitate telomerase activity in human cells, but neither the 60 kDa Ro protein nor Y RNAs were identified as essential components of telomerase activity *in vitro* (28).

Regardless of its true role in cellular metabolism, the primary functional component of the Ro RNP could be either Ro protein or the small, structural RNAs. Accordingly, the RNAs may be serving as purely architectural elements or alternatively may function in some catalytic capacity. In either case, Y RNAs may directly interact with other proteins or nucleic acids, in addition to their known interactions with the 60 kDa Ro and La proteins. Indeed, native Ro complexes have been observed to be high in molecular weight (3), suggesting the possible presence of additional macromolecules. Any sites on the Y RNA structures which interact with other cellular constituents would be predicted to be highly conserved in sequence, structure or both. One hypothesis is that Y RNAs directly participate in Ro function by base pairing with other cellular nucleic acids, as is the case with a number of other non-messenger RNAs, including the small nuclear U RNAs of the spliceosome complex (29–34). An additional prediction of this hypothesis is that any sequence-conserved interaction sites will be located, at least partially, in looped regions of the RNA secondary structures, available for the initiation of pairing interactions.

In order to test these predictions, we have obtained novel Y RNA sequences from cell lines of rabbit, duck, trout, guinea pig and cow origin. These sequences, along with the previously published human, mouse, iguana and frog Y RNA sequences have been used to predict the most likely Y1, Y3 and Y4 RNA secondary structures from a phylogenetic comparison analysis. This method of analysis is based upon the premise that molecular homologues will form highly similar structures, regardless of sequence differences. Although many RNA structures have been successfully predicted from phylogenetic comparison analyses, most notably, 16S rRNA, 5S rRNA and ribonuclease P RNA (35–37), an efficient and objective method of analysis has been lacking. This issue has been addressed in this study with an analysis which utilizes a novel secondary structure comparison algorithm to arrive at secondary structure models from computer predictions of optimal and suboptimal structures of homologous RNAs.

MATERIALS AND METHODS

GenBank accession numbers

The novel Y RNA sequences described herein may be accessed with the following GenBank codes: rabbit Y1, U82128; duck Y1, U82125; trout Y1, U82129; duck Y3, U82125; guinea pig Y1, U84678; and cow Y1, U84671.

RNA purifications, sequencing and reverse transcription (RT)–PCR

Rabbit, duck and trout Y RNAs were purified by immunoprecipitation as described (23,38) from SIRC, CCL141 and RTG-2 cell lines (ATCC, Rockville, MD), respectively, by A. D. Farris. Guinea pig and cow Y RNAs were obtained from Cav-12 and BBK cell lines by G. J. M. Pruijn.

RT–PCR of purified rabbit and duck Y RNAs was conducted using specific primers (17–20 nt long) constructed from human Y1 and Y3 RNA end sequences. Reverse transcriptions were carried out (Pharmacia First Strand cDNA Synthesis Kit, Alameda, CA), then cDNAs were amplified by standard PCR conditions (30 cycles of 30 s each at 94, 55 and 72°C in reactions containing 50 mM KCl, 10 mM Tris–HCl, pH 8.0, 1.5 mM $MgCl_2$, 0.1% Triton X-100, 0.2 mM each dNTP, 0.4 µM each primer and 2.5 U *Taq* DNA polymerase).

RT–PCR of guinea pig and cow Y RNAs was conducted using total cellular RNA as template with mixtures of Y RNA end sequence primer pairs (26–29 nt long). RT was carried out using Superscript II reverse transcriptase (Gibco BRL, Breda, The Netherlands), followed by PCR [five cycles of 1 min each at 95, 35 and 72°C, followed by 30 cycles of 1 min each at 95, 45 and 72°C in solutions containing 50 mM KCl, 20 mM Tris–HCl, pH 8.3, 2.5 mM $MgCl_2$, 0.05% W-1 (Gibco BRL), 0.1 mM each dNTP, 0.5 µM each primer and 2.5 U of *Taq* DNA polymerase].

To obtain sequence information from the trout Y1 RNA and the duck Y3 RNA, which did not share sufficient end sequence homology with human Y RNAs for successful RT–PCR, alternative strategies were employed. Purified trout Y RNA was radiolabeled at the 3′ end with [5′-^{32}P]pCp and RNA ligase, gel purified and sequenced enzymatically (Nuclease Method RNA Sequencing Kit, Amersham/USB, Arlington Heights, IL). Complementary DNA was then synthesized (First Strand cDNA Synthesis Kit, Pharmacia) with a trout Y1 specific reverse primer (tY1.rev, TAGTGAGCAGGTWGGGATCAC) constructed using the enzymatic RNA sequencing information. Following poly d(G) tailing with terminal deoxynucleotidyl transferase, the cDNA was amplified by PCR with tY1.rev and a poly d(C) adapter primer [LinC, GGCGAGCTCGAATTCGGTA$(C)_{14}$].

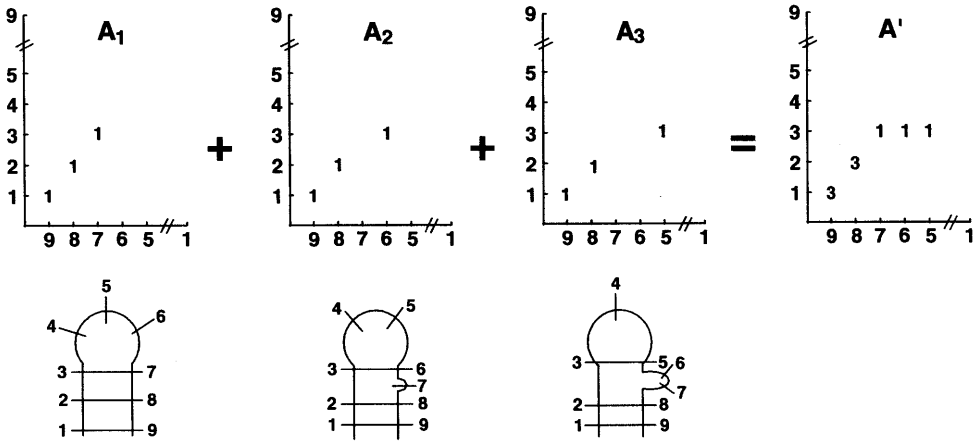
For the duck Y3 RNA, first strand cDNA synthesis was carried out with a human Y3 reverse primer followed by poly d(G) tailing and PCR with the Y3 specific primer and LinC.

Complementary DNA cloning and sequencing

The rabbit Y1 RT–PCR product was blunt-ended with T4 DNA polymerase, then ligated into *Hin*cII digested and purified pUC18. The duck and trout RT–PCR products were cloned into the pCRII TA cloning vector (Invitrogen, San Diego, CA) according to the manufacturer’s recommendations. Other RT–PCR products were digested with *Eco*RI and *Hin*dIII, gel purified and cloned into the *Eco*RI and *Hin*dIII sites of pGEM-3Zf. Double-stranded dideoxy DNA sequencing was performed with the Sequenase version 2.0 DNA Sequencing Kit (Amersham/USB).

Hybrid RNA sequences

For secondary structure modeling, hybrid sequences were constructed for the RNAs with one or more undetermined end sequences. The first 17 and last 19–20 nt of the rabbit, duck, trout and guinea pig Y1 sequences are from the human Y1 RNA sequence. Similarly, the first 15 and last 20 nt of the cow Y3 RNA are from the human Y3 sequence. It is already known that the bulk of nucleotides represented by human sequence in the hybrid RNAs base pair to form stems in Y RNA structures (24,39). Therefore, it was presumed that the analysis of hybrid sequences, as constructed,



$$\sum\left(\frac{a'_{ij}}{N}\right)^2 = \left(\frac{3}{3}\right)^2+\left(\frac{3}{3}\right)^2+\left(\frac{1}{3}\right)^2+\left(\frac{1}{3}\right)^2+\left(\frac{1}{3}\right)^2+0+0\ldots$$

$$\begin{aligned}\text{Score} &= BC + \sum\left(\tfrac{a'_{ij}}{N}\right)^2\\ &= (1)(0.45) + 2.\overline{33}\\ &= 2.78\overline{3}\end{aligned}$$

Figure 1. Hypothetical SORA score calculation. Matrices A_1, A_2 and A_3, representing hypothetical stem–loops 1, 2 and 3, are shown. A value of 1 is assigned to each base pair between nucleotides *i* and *j*, and is depicted as such in matrices A. The summation of matrices A, resulting in matrix A′, is shown. The elements of matrix A′, a′, are divided by the number of homologues, N, considered and squared. These values are then summed and added to the value BC, where B is the number of unanimously unpaired nucleotides and C is a constant used to modify the contribution of B to the score. Higher scores reflect increased structural similarity.

would not interfere with determinations of the structures assumed by the middle two-thirds of the Y RNA sequences.

Secondary structure modeling

All computer applications were performed on a Digital 2100 Server (DEC 2100) containing two Alpha processors. The operating system was running OpenVMS AXP version 6.2.

Optimal and suboptimal secondary structures were predicted for each Y RNA sequence using the MFOLD program, version 2.2 (40) available in the Genetics Computer Group (GCG) version 8.0 program suite (41). MFOLD parameters were as follows: maximumloop = 40, temp = 37.0°C for human, cow, mouse, guinea pig, rabbit and iguana sequences; temp = 42.0°C for the duck sequence, and temp = 26.0°C for frog and trout sequences.

Sequence alignments from the GCG LINEUP module were used to adjust all sequences to a maximal length of *n* bases, in which gaps were inserted to maximize sequence similarity.

An adaptation of the PLOTFOLD (41) program was used to derive base pair information for the maximum number of structures [/window = 1.0] having free energies within 25% of that assigned to the most thermodynamically stable structure predicted by MFOLD. In this adaptation, called PFOLD, a single secondary structure is stored in an array of dimension *n* X *n*, where A_{ij} = 1 for a base-pair between bases *i* and *j*; A_{ij} = 0 otherwise. The matrix A is symmetric about the axis $i = j$.

In the suboptimal RNA analysis program (SORA), for a given RNA homologue, single optimal or suboptimal secondary structure predictions from each species are combined as a comparison set. The unique portion of matrices A in a particular comparison set are summed to give matrix A′. A score for any comparison set of structures is given by: score = BC + $\Sigma(a'_{ij}/N)^2$, where B is the number of bases unanimously unpaired in all the structures of a particular comparison set, C is a constant to modify the contribution of B to the score, and N is the number of homologues analyzed. Constant C is required for accurate scoring in this algorithm so that structure combinations do not score higher due to the presence of base pairs which may not be common to all structures in a comparison set. A value of C = 0.45 for PFOLD with a window = 1.0 was empirically found to be optimal among a range of values tested and was thus selected as a default parameter. The scores are computed for all possible combinations of sub-structures from MFOLD data produced for each RNA homologue. Up to 200 of the highest scoring comparisons are reported. An example score calculation for a single comparison of three hypothetical structures is shown in Figure 1.

RESULTS

Novel vertebrate Y RNA sequences

A general strategy of RT–PCR using primers constructed from human Y RNA end sequences was employed for obtaining novel vertebrate Y RNA sequence information. For the trout Y1 and duck Y3 RNAs, sufficient end sequence divergence was present to preclude hybridization with human primers. In these instances, approaches involving internal enzymatic RNA sequencing and homopolymer tailing of complementary DNA (cDNA) allowed RT–PCR to be conducted such that all but the 3′ ends of these RNAs should be resolved. In total, partial sequences for six different Y RNAs from five different species were revealed. This is the first elucidation of Y RNA sequence information from the birds and fish. These data are summarized in Figure 2.

Most interesting was the trout Y1 RNA homologue, with 32 differences from the human Y1 (hY1) RNA sequence noted

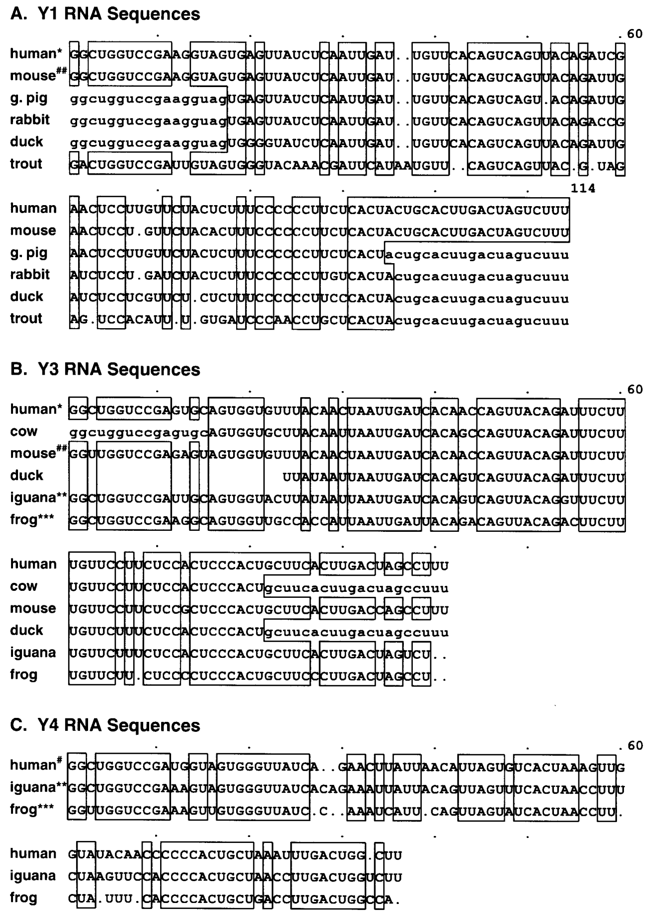

A. Y1 RNA Sequences
60
human* GGCUGGUCCGAAGGUAGUGAGUUAUCUCAAUUGAU..UGUUCACAGUCAGUUACAGAUCG
mouse## GGCUGGUCCGAAGGUAGUGAGUUAUCUCAAUUGAU..UGUUCACAGUCAGUUACAGAUUG
g. pig ggcugguccgaagguagUGAGUUAUCUCAAUUGAU..UGUUCACAGUCAGU.ACAGAUUG
rabbit ggcugguccgaagguagUGAGUUAUCUCAAUUGAU..UGUUCACAGUCAGUUACAGACCG
duck ggcugguccgaagguagUGGGUAUCUCAAUUGAU..UGUUCACAGUCAGUUACAGAUUG
trout GACUGGUCCGAUUGUAGUGGGUACAAACGAUUCAUAAUGUU..CAGUCAGUUAC.G.UAG
114
human AACUCCUUGUUCUACUCUUUCCCCCCUUCUCACUACUGCACUUGACUAGUCUUU
mouse AACUCCU.GUUCUACACUUUCCCCCCUUCUCACUACUGCACUUGACUAGUCUUU
g. pig AACUCCUUGUUCUACUCUUUCCCCCCUUCUCACUacugcacuugacuagucuuu
rabbit AUCUCCU.GAUCUACUCUUUCCCCCCUUGUCACUAcugcacuugacuagucuuu
duck AUCUCCUCGUUCU.CUCUUUCCCCCCUUCCCACUAcugcacuugacuagucuuu
trout AG.UCCACAUU.U.GUGAUCCCAACCUGCUCACUAcugcacuugacuagucuuu
B. Y3 RNA Sequences
60
human* GGCUGGUCCGAGUGCAGUGGUGUUUACAACUAAUUGAUCACAACCAGUUACAGAUUUCUU
cow ggcugguccgagugcAGUGGUGCUUACAAUUAAUUGAUCACAGCCAGUUACAGAUUUCUU
mouse## GGUUGGUCCGAGAGUAGUGGUGUUUACAACUAAUUGAUCACAACCAGUUACAGAUUUCUU
duck UUAUAAUUAAUUGAUCACAGUCAGUUACAGAUUUCUU
iguana** GGCUGGUCCGAUUGCAGUGGUACUUAUAAUUAAUUGAUCACAGUCAGUUACAGGUUUCUU
frog*** GGCUGGUCCGAAGGCAGUGGUUGCCACCAUUAAUUGAUUACAGACAGUUACAGACUUCUU
human UGUUCCUUCUCCACUCCCACUGCUUCACUUGACUAGCCUUU
cow UGUUCCUUCUCCACUCCCACUgcuucacuugacuagccuuu
mouse UGUUCCUUCUCCGCUCCCACUGCUUCACUUGACCAGCCUUU
duck UGUUCUUUCUCCACUCCCACUgcuucacuugacuagccuuu
iguana UGUUCUUUCUCCACUCCCACUGCUUCACUUGACUAGUCU..
frog UGUUCUU.CUCCCCUCCCACUGCUUCCCUUGACUAGCCU..
C. Y4 RNA Sequences
60
human# GGCUGGUCCGAUGGUAGUGGGUUAUCA..GAACUUAUUAACAUUAGUGUCACUAAAGUUG
iguana** GGCUGGUCCGAAAGUAGUGGGUUAUCACAGAAAUUAUUACAGUUAGUUUCACUAACCUUU
frog*** GGUUGGUCCGAAAGUUGUGGGUUAUC.C..AAAUCAUU.CAGUUAGUAUCACUAACCUU.
human GUAUACAACCCCCCACUGCUAAAUUUGACUGG.CUU
iguana CUAAGUUCCACCCCACUGCUAACCUUGACUGGUCUU
frog CUA.UUU.CACCCCACUGCUGACCUUGACUGGCCA.

Figure 2. Y RNA sequences and alignments. Sequence alignments of Y1 (**A**), Y3 (**B**) and Y4 (**C**) are shown. Sequence derived from oligonucleotide primers is shown in lowercase letters. Nucleotides which are conserved across all species, with the exception of primer regions, are boxed. *, from ref. 4; **, from ref. 22; ***, from ref. 9; #, from ref. 5 and ##, from ref. 23. All other sequences are from this study.

over the interval homologous to hY1 nucleotides 1–93. The duck and rabbit Y1 RNAs were found to exhibit seven and five sequence differences from hY1, respectively, within the 72 nt interval spanning the primer regions. Similarly, the guinea pig Y1 RNA differed from the hY1 18–92 sequence in two positions, with one nucleotide deletion and one substitution. The cow and duck Y3 homologues were found to contain three and five differences, respectively, from the human Y3 (hY3) RNA sequence.

Y RNA sequence alignments

Alignment of these novel Y RNA sequences with those already known revealed certain conserved segments within each of the Y1, Y3 and Y4 RNAs (boxed areas in Fig. 2; primer regions are indicated by lower case letters). Several of these conserved segments are ≥7 nt in length (colored boxes). A 9 nt segment at Y1 positions 3–11 is invariantly conserved and occurs, less 1 nt, in the Y3 and Y4 RNAs where it occurs in both at positions 4–11. These sequences fall within oligonucleotides protected from ribonuclease digestion by Ro protein and further define part of the Ro binding site (24). Although the conservation of the 3′ half of the Ro binding site in Y1 RNAs could not be evaluated, the corresponding segment in the Y4 sequences is well conserved, with an 8 nt invariant segment present at positions 85–92 (Fig. 2). Nucleotides at the 3′ end of hY3 which have been protected from RNase digestion by Ro protein, Y3 positions 91–97, are less well conserved but point to specific bases which may directly contact the Ro protein.

A number of other invariantly conserved segments of sequence were noted in each of the Y RNAs, which may mark protein or nucleic acid binding sites. Particularly interesting is an 8 nt segment in the Y1 RNAs, CAGUCAGU, beginning at position 44 in Figure 2. If gaps in the guinea pig and trout sequences may be ignored, then this sequence is a portion of 17 contiguous, invariant nucleotides. Aside from a 6 nt fragment near the 5′ end of Y1 (part of the Ro binding site), no other fragment longer than 2 or 3 nt is completely conserved in Y1; this is excluding the 3′ end, which could not be evaluated. The presence of such a conserved segment amid non-conserved nucleotides suggests that this region is important in Y1 function. Furthermore, if Y1 directly binds another cellular constituent, this is the most likely site that would be involved.

There are four completely conserved segments 7 nt in length in the Y3 sequences not including the Ro protein binding sites, though a 6 nt fragment near the 5′ end may not be excluded as potentially important for function (Fig. 2). Interestingly, one of these sites (Y3 positions 56–65, UUCUUUGUUC) overlaps with a potentially homologous segment in *Caenorhabditis elegans* Y RNA (UUUCUUU) by 6 nt, possibly marking a key functional site in the Y3 RNA (31).

Finally, three contiguous stretches of conserved nucleotides were found in the Y4 sequences, occurring at hY4 positions 17–26, 47–53 and 69–78. Further analysis of more disparate Y4 sequences will be required to assess the likely importance of each of these regions.

Secondary structure modeling

Sequences were first aligned to maximize sequence similarity using the Genetics Computer Group (GCG) Lineup program (41), and alignments were saved with the Print command. Gap positions were identical to those in Figure 2; however, the guinea pig Y1 and cow Y3 sequences shown in Figure 2 were not included in initial analyses. RNAs having one or more ends with undetermined sequence were analyzed as artificial hybrid sequences in which the first 15–17 and last 19–20 nt were substitutions from the human sequences (Materials and Methods); these included guinea pig, rabbit, duck and trout Y1 RNAs and the cow Y3 RNA. A lack of sufficient unique duck Y3 sequence information, owing to an apparent 5′ end truncation during reverse transcription, precluded this RNA from analysis.

Individual complete or hybrid sequences were then folded with the MFOLD program version 2.2 on the GCG 8.0 program suite (40,41), using the /temp qualifier to allow the usage of adjusted energy rules for the normal body temperatures of the animal species from which the sequences were taken. MFOLD data sets consisting of those structures occurring within 25% of the predicted free energy minimum, were then created using PFOLD, an adaptation of PLOTFOLD as discussed in Materials and Methods. PFOLD outputs the data in two formats—one format (.cnfld files) is read by our SORA, while the second format (.fld files) allows typical PLOTFOLD outputs of the structures for

Table 1. tRNA^{gln} SORA solutions

Solution	MFOLD Structure Number						SORA
No.	*E.coli*	*A. laidlawii*	*M. pneumoniae*	*S. lividans*	*T. thermophila*	*H. influenzae*	Score
1	1	1	2	2	2	6	23.56
2	1	1	7	2	2	6	23.39
3	1	1	2	11	2	6	23.14
4	1	1	3	2	2	6	23.12
5	1	1	2	2	2	16	23.08

The five highest scoring solutions are shown. MFOLD structure number is the ranking by free energy of a particular secondary structure computed by GCG MFOLD for tRNA^{gln} from various species. Designations are as follows: *E.coli, Escherichia coli*; *A.laidlawii, Acholeplasma laidlawii*; *M.pneumoniae, Mycoplasma pneumoniae*; *S.lividans, Streptomyces lividans*; *T.thermophila, Thermus thermophila*; *H.influenzae, Haemophilus influenzae*.

manual perusal of the data sets. For our analyses, the structures in the so obtained data sets were numbered consecutively and are referred to by these MFOLD structure numbers. SORA jobs were run in batch mode; the larger jobs required increasing certain VMS parameters such as Working Set Quota, which was increased to 50 000.

An initial test was conducted with available tRNA glutamine sequences from various genera of Eubacteria [*Escherichia* (GenBank K00182), *Acholeplasma* (X61067), *Mycoplasma* (X17113), *Streptomyces* (X58873), *Thermophilus* (M35400) and *Haemophilus* (U32783)], since the *Escherichia coli* tRNA-gln structure is known (42). The number of structures in the resulting P25 (free energy within 25% of the predicted minimum) datasets ranged from 8 to 30, and a total of 13 271 040 comparisons were made by SORA. The results, by MFOLD structure number, are shown in Table 1. Only one structure type was found in the five best answers, and this SORA solution structure completely agrees with the secondary structure derived from the crystal structure.

SORA was then applied to the Y1, Y3 and Y4 sequence alignments in initial analyses, and those results are summarized in Table 2. The MFOLD data sets used in these analyses allowed a maximum loop size of 40 since similar analyses conducted with the default maximum loop size of 30, gave Y4 and Y3 model structures which were not unanimously homologous and/or had lower SORA scores (data not shown). The total number of comparisons made for the Y1, Y3 and Y4 analyses were 71 662 500, 1 477 980 and 3696, respectively. The longer run consumed ~3 h of CPU time, while the shortest run was completed in seconds.

The Y1 SORA solution structures from individual species, along with their associated MFOLD structure numbers are presented in Figure 3. The final solution required manual unpairing of certain base pairs which were not unanimously conserved in order to maximize the similarity of structures between species. These manual changes are highlighted on the original MFOLD-derived structures shown in Figure 3. Remarkably, the Y1 SORA solution structure is nearly identical to a model proposed for the Y1 RNA from chemical and enzymatic probing data (39).

Table 2. Y1, Y3 and Y4 RNA SORA solutions from initial analyses

Y1 Results

	MFOLD Structure Number					SORA
Solution No.	Human	Mouse	Rabbit	Duck	Trout	Score
1	3	4	2	27	45	45.73
2	3	4	2	1	45	45.61
3	3	5	2	27	45	45.57
4	3	5	2	1	45	45.45
5	7	11	3	1	45	45.40

Y3 Results

	MFOLD Structure Number				SORA
Solution No.	Human	Mouse	Iguana	*Xenopus*	Score
1	15	6	1	4	43.84
2	15	6	1	12	43.51
3	15	1	1	4	42.50
4	1	1	1	4	42.24
5	15	1	1	12	42.18

Y4 Results

	MFOLD Structure Number			SORA
Solution No.	Human	Iguana	*Xenopus*	Score
1	16	1	1	39.56
2	16	6	1	38.43
3	16	1	8	38.43
4	16	16	1	37.98
5	16	20	1	37.98

The five highest scoring solutions are shown. MFOLD structure number is the ranking by free energy of a particular secondary structure computed by GCG MFOLD for Y RNAs from various species.

The presence of stem IIIa was supported by chemical probing data in that model, but not by the enzymatic data. Perhaps this stem breathes in the Y1 structure or is present in a subset of Y1

Table 3. Y1 and Y3 RNA SORA solutions from expanded analyses

Y1 Results

Solution No.	MFOLD Structure Number						SORA Score
	Human	Mouse	Guinea Pig	Rabbit	Duck	Trout	
1	3	4	1	2	27	45	46.12
2	3	4	1	2	1	45	46.04
3	3	5	1	2	27	45	45.90
4	3	5	1	2	1	45	45.82
5	3	4	1	2	10	45	45.67

Y3 Results

Solution No.	MFOLD Structure Number					SORA Score
	Human	Cow	Mouse	Iguana	*Xenopus*	
1	15	12	6	1	4	43.91
2	15	12	6	1	12	43.54
3	15	12	1	1	4	42.57
4	1	1	1	1	4	42.44
5	15	12	1	1	12	42.20

The five highest scoring solutions from each analysis are shown. MFOLD structure number is the ranking by free energy of a particular secondary structure computed by GCG MFOLD for Y RNAs from various species. The guinea pig Y1 data set used in the Y1 analysis contained a forced structure generated with the following command: MFOLD /FORC=1,108,8 /FORC2=13,94,9 /FORC3=24,55,4 /FORC4=30,49,8 /FORC5=57,69,4 /MAXL=40.

molecules. All five stems present in the Y1 SORA solution structure are recurrent motifs in nearly all of the MFOLD P25 data sets tested, with prevalences ranging from 13 to 97%. Exceptions include stems IIIa and IIIb in the trout data set and stem IV in the trout and duck data sets. While the trout IIIb and duck IV stems occurred only once in their respective data sets, true stems IIIa and IV did not occur at all in the trout P25 data set. Moreover, these stems were not found in the entire set of trout MFOLD structures produced at a window of 1.0. However, the trout structure found in the Y1 SORA solution set (Fig. 3), did contain pseudo-stems IIIa and IV, denoted IIIa′ and IV′. A number of base changes occur in the predicted helical regions of the Y1 solution structures, providing additional evidence for their existence. Compensatory changes support stems II and IV, and these occur at rabbit base pair U59–A66 and duck base pair G20–C87, respectively.

Interestingly, when SORA was run with a duck Y1 MFOLD data set folded at /temp = 37°C, rather than the normal Avian body temperature of 42°C, a duck structure similar to the Y1 SORA solution structure was not produced. This illustrates a utility of the temperature modified energy rules of Turner and colleagues (43). The isolated base pairs between human Y1 nucleotides 71 and 86 may be unlikely to exist; although allowable in MFOLD version 2.2 (30), isolated base pairs are disallowed in MFOLD version 2.3 (M.Zuker, personal communication). These two isolated base pairs and 11 others (Fig. 3) are not completely conserved in the Y1 solution structures and should be manually unpaired to maximize structural similarity.

The Y3 and Y4 SORA solution structures (Fig. 4) are essentially identical to models previously proposed by us using fewer species and a completely different comparison algorithm (11,22). Similar to the Y1 SORA solution, the predicted helical regions of the Y3 and Y4 secondary structure models are recurrent motifs in the MFOLD data sets, and stems I and II are supported by a number of base changes, though none is compensatory. Thus, although these Y3 and Y4 secondary structures are the best predictions which can be made given the current data, sequences of more disparate Y3 and Y4 homologues will be required to test these solutions.

Manual examination of several of the best scoring SORA solutions for all of the analyses conducted confirmed that the SORA scores accurately reflected the degrees of homology among the combinations of structures compared.

Following these initial analyses, two additional partial sequences became available, namely the guinea pig Y1 and cow Y3 sequences. After the construction of hybrid sequences for these RNAs as discussed in Materials and Methods, they were added to the Y1 and Y3 SORA analyses. The number of comparisons examined in the Y1 analysis was thus increased to 2 866 500 000, while the number of Y3 comparisons increased to 67 987 080. While the expanded Y3 SORA analysis yielded the same SORA solution structure as before (Table 3), the expanded Y1 analysis found a somewhat different solution structure, with a SORA score of 45.67 (data not shown).

There is a simple explanation for this finding. The best structure from the earlier analysis for Y1 was not present in the PLOTFOLD output and was not available for SORA. Once this structure was included (i.e., forced), then it was found to be the best solution in the revised analysis. Perhaps, this problem would have been avoided if a smaller energy increment was allowable by PLOTFOLD, thereby producing more of the possible structures for SORA to consider.

In particular, manual examination of the guinea pig Y1 data set of Y1 structures revealed an absence of the structure representing the initial solution for Y1 depicted in Figure 3, although such a structure (with a predicted free energy of –19.6 kcal) was easily forced using the /force MFOLD qualifier. This structure was also not found in the entire PLOTFOLD data set of Y1 structures, though all stem elements of the structure were present as recurrent motifs in the data set. Upon the addition of this forced structure to the guinea pig Y1 data set submitted to SORA, the SORA analysis yielded the same solution shown in Figure 3, with a score of 46.12, exceeding the score obtained without the inclusion of the forced structure (Table 3). Therefore, the structure initially found was the best of those known. Also, though the PLOTFOLD algorithm will not produce all possible RNA secondary structures *per se*, it has the capacity of predicting multiple structures and a superior capacity to predict stem motifs.

The invariantly conserved residues at the 5′ ends of each of the Y RNAs (Y1 nucleotides 3–11 and Y3 and Y4 nucleotides 4–11) are found to participate in a stem proposed by others (4,39), and a previously proposed bulged cytidine (24,39) is unpaired in all of the structures predicted.

The core of the Y1 invariantly conserved sequence (hY1 nucleotides 42–49, CAAGUCAGU) is located entirely within a predicted helical region, apparently unavailable for base pairing with other cellular nucleic acids. In contrast, a number of other invariantly conserved sequences, reside at least partially in looped regions in the Y RNA molecules. The Y3 conserved segment at hY3 positions 56–65, which overlaps a potentially homologous region of the *C.elegans* Y RNA, is completely unpaired in the Y3 SORA solution. Similarly, 5 of 10 invariant nucleotides in the Y4 RNA, at hY4 positions 17–26, are unpaired. Though many other of the conserved blocks of nucleotides are primarily located in predicted helices of the Y RNA SORA

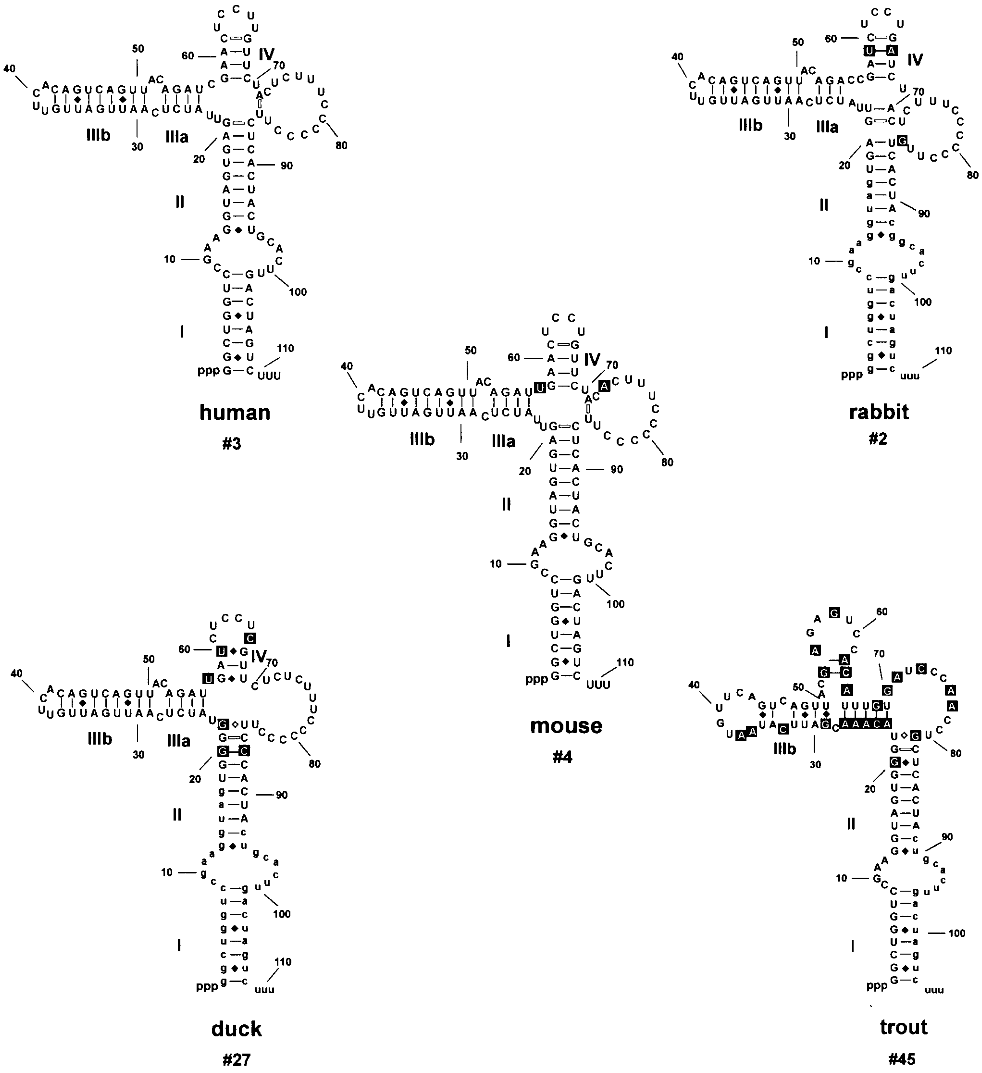


Figure 3. Y1 SORA solution structures. Secondary structures are from the highest scoring SORA solution shown in Table 2. Nucleotides differing from the human Y1 sequence are reverse shaded. Rankings, corresponding to free energies predicted by MFOLD, are labeled beneath species designation. Lowercase letters indicate bases derived from oligonucleotide primers. Covariation was observed in stem II of duck Y1 (pair between G^{20} and C^{87}) and in stem IV of rabbit Y1 (pair between U^{59} and A^{66}). Bars between nucleotides denote canonical base pairs. Diamonds between nucleotides denote non-canonical base pairs. Hollow bars and hollow diamonds between nucleotides represent base pairs which are not conserved and were manually unpaired to maximize structural similarity.

solution structures, most have 2 or 3 nt dangling into looped regions. Since only a few unpaired nucleotides may be required to initiate base pairing between a folded RNA and another nucleic acid (44); however, the possibility that these conserved Y RNA segments base pair to other cellular nucleic acids cannot be discounted.

DISCUSSION

An objective method for phylogenetic secondary structure analysis has been developed, in which the novel SORA, described here, is combined with the already widely available MFOLD program, allowing automated phylogenetic analysis. In addition to accurately predicting the Eubacterial tRNA^{gln} secondary structure, which completely agrees with the *E.coli* tRNA^{gln} crystal structure, independent chemical and enzymatic probing evidence strongly suggests that the Y1 RNA model described herein is accurate as well (39). Both Y1 RNA models predict stems I, II, IIIb and IV. While stem IIIa is not supported by the enzymatic cleavage data from that study, it is somewhat in agreement with the chemical probing data, suggesting that this stem could be forming in a subset of purified Y1 RNAs in solution. The peculiar pyrimidine-rich loop which was unusually resistant to both single- and double-stranded RNases in that study, does occur in the Y1 SORA solution structure as well, but is not conserved in sequence. The use of the MFOLD/SORA method to successfully predict both transfer and Y RNA structures demonstrates its broad applicability to general problems of RNA secondary structure comparison.

The trout Y RNA molecule described herein shares 66% sequence identity to the human Y1 RNA molecule. With Y1 identities spread throughout its length (Fig. 2A), it is reasonable to assume that this trout Y RNA is a Y1 homologue. Given this assumption, however, this RNA could not be folded into a secondary structure absolutely identical to the Y1 RNAs of other

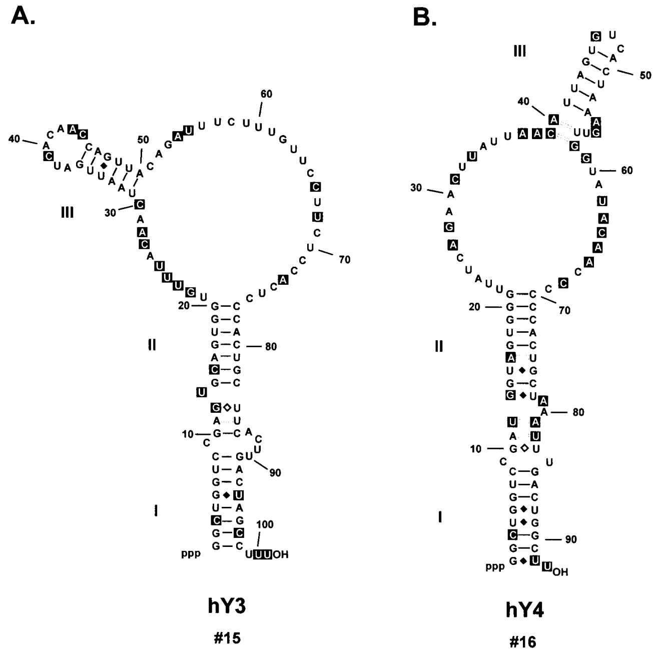


Figure 4. Y3 and Y4 SORA solution structures. Only the human Y3 and Y4 structures are shown. Reverse shaded letters are positions of either nucleotide substitutions or gaps in sequence alignment, as compared with homologous non-human sequences. Other designations are as in Figure 3. Covariation was evident in stem III of iguana Y4 (pair between A^{39} and U^{58}, numbered according to human Y4).

species (Fig. 3), suggesting that the trout Y1, as the Y1 sequence itself, has diverged to some degree. Alternatively, it remains a formal possibility that the trout 'Y1' sequence is not a Y1 homologue but either arose by convergent evolution or has been lost in other species. There may be a precedent for at least one of these pathways with the existence of a *Xenopus* Y RNA, which shares sequence similarity with other known Y RNAs only at the 60 kDa Ro binding region and is otherwise not identifiable as a homologue of any other known Y RNA (9). Notably, SORA analysis conducted using various Y1 RNAs in the absence of the trout MFOLD data set yields the same solution presented in Figure 3. Thus, even if the trout Y RNA is not a Y1 homologue, the structural model has not been skewed by its inclusion in the analysis.

While the MFOLD/SORA method is a valuable tool for comparative analyses, it has uncovered certain characteristics of the MFOLD algorithm which should be noted. Though MFOLD may predict all the possible stem motifs of a particular RNA, as viewed in an energy dot plot, it does not necessarily predict all possible combinations of these motifs in single particular structures, as viewed by PLOTFOLD. Therefore, users of these programs should be aware that all possible RNA secondary structures *per se* are not predicted by MFOLD/PLOTFOLD. One way of handling this problem, which was illustrated by the guinea pig Y1 data set in this work, may be the use of a Monte Carlo-styled approach to the SORA/MFOLD phylogenetic analysis. Thus, in addition to analyzing data sets from all homologues at once (N = 6 in the case of Y1), separate analyses using combinations of all homologues less one (N = 5 in the case of Y1) could be conducted. In this way, multiple candidate solution structures could be generated, and the data sets of the homologues missing from each N–1 analysis could be examined for the true lack of ability to fold into their respective N–1 SORA solution structures. If the structures in question could be forced to fold, they would be added to their respective data sets, then the entire group of expanded structure sets (N′) could be re-analyzed using SORA. Hence, any inability of MFOLD to predict a crucial homologous structure could be circumvented.

Like all phylogenetic comparative analyses, the MFOLD/SORA method requires the use of sequences having sufficient homology for reliable sequence alignment. Therefore, the most recently characterized *C.elegans* Y RNA could not be incorporated into the current analysis. The only two recognizable Y RNA homologies within this RNA are the Ro protein binding site and a possible 6 nt homology (UUCUUU, at hY3 positions 56–61) with an invariantly conserved segment of the Y3 RNA identified herein. The invariance of this sequence likely marks an element important to or even essential for Y3 RNA function. Interestingly, this sequence is completely unpaired in the proposed Y3 RNA SORA solution structure, leaving open the possibility that the Y3 RNA may base pair with another cellular nucleic acid(s) at this site.

The presence of an invariant block of 8 nt in the Y1 sequence (at hY1 positions 42–49) amid surrounding nucleotides which are not conserved suggests that this is a key functional site for this RNA. However, its location within a paired region of the Y1 RNA favors the interpretation that it is a site of protein interaction or a necessary structural element, rather than a nucleic acid binding site.

In addition to those discussed above, all blocks of invariant sequence identified in this study are potentially important to Ro function and may be involved in sequence-specific contacts with protein or nucleic acid. Furthermore, those located at least partially in looped regions in the Y RNA secondary structures may be particularly interesting in this regard since they are more likely to have sequence-specific contacts exposed.

Regardless of the nature of any proposed contacts between the Y RNAs and other cellular constituents, the existence of invariantly conserved nucleotides in the Y RNAs is consistent with a direct role for Y RNAs in Ro function.

ACKNOWLEDGEMENTS

The authors thank Joanne Gross and Ricki Galchen for expert technical assistance, and Les Cummings, Tom Barton and the Management Information Services staff at OMRF. We also thank Peggy Anderson and the OASIS staff at OMRF. This work was supported in part by the Netherlands Foundation for Chemical Research (SON) with financial aid from the Netherlands Organization for Scientific Research (NWO). The work of G.J.M.P has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. The SORA program is available by return email as FORTRAN executable source code. Please send requests to gerald-koelsch@omrf.ouhsc.edu.

REFERENCES

1 Hendrick,J.P., Wolin,S.L., Rinke,J., Lerner,M.R. and Steitz,J.A. (1981) *Mol. Cell. Biol.*, **1**, 1138–1149.
2 Slobbe,R.L., Pluk,W., van Venrooij,W.J. and Pruijn,G.J. (1992) *J. Mol. Biol.*, **227**, 361–366.
3 Boire,G. and Craft,J. (1990) *J. Clin. Invest.*, **85**, 1182–1190.
4 Wolin,S.L. and Steitz,J.A. (1983) *Cell*, **32**, 735–744.
5 O'Brien,C.A. and Harley,J.B. (1990) *EMBO J.*, **9**, 3683–3689.

6 Kato,N., Hoshino,H. and Harada,F. (1989) *Biochem. Biophys. Res. Commun.*, **108**, 363–370.

7 Deutscher,S.L., Harley,J.B. and Keene,J.D. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 9479–9483.

8 Ben-Chetrit,E., Gandy,B.J., Tan,E.M. and Sullivan,K.F. (1989) *J. Clin. Invest.*, **83**, 1284–1292.

9 O'Brien,C.A., Margelot,K. and Wolin,S.L. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 3683–3689.

10 Van Horn,D.J., Eisenberg,D., O'Brien,C.A. and Wolin,S.L. (1995) *RNA*, **1**, 293–303.

11 Farris,A.D. (1995) Dissertation, University of Oklahoma, Oklahoma City, OK, pp. 113–117.

12 Gottlieb,E. and Steitz,J.A. (1989) *EMBO J.*, **8**, 841–850.

13 Gottlieb,E. and Steitz,J.A. (1989) *EMBO J.*, **8**, 851–862.

14 Scherly,D., Stutz,F., Lin-Marq,N. and Clarkson,S.G. (1993), *J. Mol. Biol.*, **231**, 196–204.

15 Yoo,C.J. and Wolin,S.L. (1994) *Mol. Cell. Biol.*, **14**, 5412–5424.

16 Slobbe,R.L., Pruijn,G.J., Damen,W.G., van der Kemp,J.W. and van Venrooij,W.J. (1991) *Clin. Exp. Immunol.*, **86**, 99–105.

17 Reddy,R., Tan,E.M., Henning,D., Nohga,K. and Busch,H. (1983) *J. Biol. Chem.*, **258**, 1383–1386.

18 Mamula,M.J., O'Brien,C.A., Harley,J.B. and Hardin,J.A. (1989) *Clin. Immunol. Immunopathol.*, **52**, 435–446.

19 Craft,J., Mamula,M., Ohosone,Y., Boire,G., Gold,H. and Hardin,J. (1990) *Clin. Rheum.*, **9**, 10–19.

20 Itoh,Y., Kriet,J.D. and Reichlin,M. (1990) *Arth. Rheum.*, **33**, 1815–1821.

21 Pruijn,G.J.M., Wingens,P.A.E.T.M., Peters,S.L.M., Thijssen,J.P.H. and van Venrooij,W.J. (1993) *Biochim. Biophys. Acta*, **1216**, 395–401.

22 Farris,A.D., O'Brien,C.A. and Harley,J.B. (1995) *Gene*, **154**, 193–198.

23 Farris,A.D., Gross,J.K., Hanas,J.S. and Harley,J.B. (1996) *Gene*, **174**, 35–42.

24 Wolin,S.L. and Steitz,J.A. (1984) *Proc. Natl Acad. Sci. USA*, **81**, 1996–2000.

25 Farris,A.D., Puvion-Dutilleul,F., Puvion,E., Harley,J.B. and Lee,L.A. (1996) *Proc. Natl Acad. Sci. USA*, **94**, 3040–3045.

26 O'Brien,C.A. and Wolin,S.L. (1994) *Genes Dev.*, **8**, 2891–2903.

27 Shi,J., Obrien,C.A., van Horn,D.J. and Wolin,S.J. (1996) *RNA*, **2**, 769–784.

28 Ramakrishnan,S., Sharma,H.W., Farris,A.D., Kaufman,K.M., Harley,J.B., Collins,K., Pruijn,G.J.M, van Venrooij,W.J., Martin,M.L. and Narayanan,R. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 10075–10079.

29 Maniatis,T. and Reed,R. (1987) *Nature*, **325**, 673–678.

30 Lerner,M.R., Boyle,J.A., Mount,S.M., Wolin,S.L. and Steitz,J.A. (1980) *Nature*, **283**, 220–224.

31 Reddy,R. and Busch,H. (1988) In Birnsteil,M.L. (ed.), *Structure and Function of the Major and Minor Small Nuclear Ribonucleoprotein Particles*. Springer, Berlin, pp. 14–15.

32 Wu,J. and Manley,J.L. (1991) *Nature*, **352**, 818–821.

33 Datta,B. and Weiner,A.L. (1991) *Nature*, **352**, 821–824.

34 Steitz,J.A. (1992) *Science*, **257**, 888–889.

35 James,B.D., Olsen,G.J., Liu,J. and Pace,N.R. (1988) *Cell*, **52**, 19–26.

36 Fox,G.E. and Woese,C.R. (1975) *Nature*, **256**, 505–507.

37 Noller,H.F. and Woese,C.R. (1981) *Science*, **212**, 403–411.

38 Forman,M.S., Nakamura,M., Mimori,T., Gelpi,C. and Hardin,J.A. (1985) *Arth. Rheum.*, **28**, 1356–1361.

39 van Gelder,C.W.G., Thijssen,J.P.H.M., Klaassen,E.C.J., Sturchler,C., Krol,A., van Venrooij,W.J. and Pruijn,G.J.M. (1994) *Nucleic Acids Res.*, **22**, 2498–2506.

40 Zuker,M. (1989) *Science*, **244**, 48–52.

41 Devereux,J. (1989) *The GCG Sequence Analysis Software Package, Version 8.0*. Genetics Computer Group, Inc., Madison, WI.

42 Rould,M.A., Perona,J.J., Soll,D. and Steitz,T.A. (1989) *Science*, **246**, 1135–1142.

43 Jaeger,J.A., Turner,D.H. and Zuker,M. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 7706–7710.

44 Tomizawa,J. (1990) *J. Mol. Biol.*, **212**, 683–694.