

## A Method for Predicting Common Structures of Homologous RNAs

SHU-YUN LE, KAIZHONG ZHANG,\* AND JACOB V. MAIZEL, JR.

*Laboratory of Mathematical Biology, Division of Cancer Biology, Diagnosis and Centers, National Cancer Institute, NIH, Building 469, Room 151, Frederick, Maryland 21702; and*

*\*Department of Computer Science, University of Western Ontario, London, Ontario, Canada N6A 5B7*

Received June 28, 1994

We have developed a procedure, composed of a set of computer programs, for predicting common RNA structures of homologous sequences. Given a set of homologous RNAs, these programs perform a multiple sequence alignment, generate a list of possible helical stems that are thermodynamically favored in RNA folding from a selected individual sequence, establish a conserved stem list by inspecting the equivalent base pairings and/or conserved helical stems from the derived alignment of homologous RNAs, and build common RNA secondary structures with the maximum scores (i.e., compensatory base changes and number of base pairs, etc.). The approach is a combination of phylogenetic and thermodynamic methods and has been applied to the prediction of common folding structures of the 5' untranslated regions in a number of positive RNA viruses. © 1995 Academic Press, Inc.

### INTRODUCTION

A complete understanding of the function of RNA molecules requires a knowledge of their three-dimensional (3D) structures. The determination of RNA structure is a limiting step in the study of RNA structure-function relationships because it is very difficult to crystallize and/or get nuclear magnetic resonance spectrum data for large RNA molecules. Reliable prediction of RNA tertiary structure from the primary sequence is therefore highly desirable. Numerous common foldings deduced from phylogenetic comparisons and from the sensitivity of nucleotides to chemical modification have demonstrated that specific sets of RNA secondary structure interactions, which are conserved during the course of evolution, are essential elements in RNA functions (1, 2, 5). An important step toward the determination of RNA 3D structure is the prediction of common RNA secondary structures. Based on a reliable RNA secondary structure, the possible tertiary interactions that occur between secondary structural elements and between these elements and single-stranded regions of RNA can be characterized. Examples include the structural studies

of group I self-splicing introns (1), 16S rRNAs (2, 3), 5S rRNAs (4), and other RNA enzymes such as RNase P RNAs (5).

Currently, common RNA secondary structures are derived by phylogenetic comparative methods. The power of these methods has been demonstrated by predictions of RNA secondary structures for 16S rRNAs, RNase P RNAs, and other RNAs. The determination of a RNA secondary structure requires a set of homologous RNAs from diverse organisms; among which the sequence varies. Generally, the sequence similarities of 60–80% are favorable for inspecting the equivalent base pairings that occur in these sequences (5). In the process, a large number of possible base pairings need to be stored and examined for their possible conservation; however, phylogenetic comparative analysis is very tedious since it is performed manually.

Although dynamic programming and energy minimization methods (6–15) for predicting RNA structure are not as successful as phylogenetic comparative methods, they can be performed automatically by computer. With the improvement of the dynamic programming algorithm and parameters for the free energy of formation of RNA structural elements, thermodynamically favored stems predicted by EFFOLD contain approximately 90% or more of the phylogenetically known helical stems in tRNAs, 16S rRNA, and group I self-splicing introns (14). A similar accuracy level for predicting RNA structure has also been obtained by Zuker's suboptimal folding (10) and by using Turner energy rules (16). Thus, these predicted stems from improved thermodynamic approaches can provide a base pairing pool with which to construct common RNA structures of homologous sequences.

In this paper, we describe a set of algorithms and computer programs for predictions of RNA common folding of homologous sequences. Contrary to common phylogenetic comparative methods, our procedure does not require manual inspection for equivalent base pairings in a set of sequences (26). The uncertainty in the alignment of multiple sequences can be partially circumvented by means of a window. The predicted RNA structure is not as biased to a specific sequence as those produced by energy minimization methods. The computed structure can offer a robust working model for further refinement of RNA structures by experimental analyses.

## METHODS

This procedure includes the following steps: (a) align a set of homologous sequences; (b) generate a list of possible thermodynamically favored helical stems; (c) examine equivalent base pairings in the set of sequences for each stem of the above list and score the stem conservation and the compensatory base changing encountered; and (d) build a RNA structure with the maximum score from those conserved stems found in these sequences. A number of multiple alignment (MAL) programs have been developed. In our procedure, we use Zuker's MAL program (17). The possible thermodynamically favored helical stems are deduced by the program EFFOLD (14), a RNA folding simula-

tion accomplished by fluctuating free energy parameters within the range of experimental errors of these thermodynamic parameters. The stem list can also be derived using Zuker's RNA suboptimal folding program (10, 15) and a stem comparison program, NEWTREE, as reported in the previous publication (18). The programs MATCH and BUILD are newly developed and perform steps (c) and (d), respectively, in our procedure.

### *MATCH Algorithm*

Given a stem list from the  $j$ th sequence and an alignment of  $M$  sequences, the MATCH program examines the conservation of each stem in the list by using MAL. To inspect the conservation of base pairs in the helical stem, we need a suitable window whose size is based on the predetermined allowable shifts of the stem position at both the 5' and 3' sides, as well as on the stem size. Using the window, the corresponding positions in other  $M - 1$  sequences for the base pairing region of the  $i$ th stem in the  $j$ th sequence can be determined by ignoring any inserted gaps in MAL when gaps are encountered. The stem conservation is represented by the recurring ratio of this stem in the  $M$  sequences. Two types of scores for compensatory base changes are employed. One score is used to evaluate if the stem is supported by covariation. The other score is used to count the bonus contributed by diverse equivalent base pairings. If the input stem list includes  $n$  stems, the algorithm requires  $O(nM)$  time and  $O(n + NM)$  space. MATCH was written in standard Pascal and runs in both VAX/VMS and IRIS/UNIX environments.

### *BUILD Algorithm*

The input of BUILD is the conserved stem list produced by MATCH. For simplicity, we assume the input is a list of  $n$  stems with information that is represented by the position of the 5' end, the position of the 3' end, the size of the stem, and the score for compensatory base changes (5end, 3end, size, score). This can be easily extended to the situation where, for each stem, there are several scores. Given a stem (5end, 3end, size, score), we call the sequence from nucleotide 5end+size to 3end-size the induced sequence of this stem. If all the base pairs of a stem are within a RNA fragment, we say this stem is "inside the sequence."

Our task is to build a RNA secondary structure using the stems from the conserved stem list. Since there are many possible RNA structures based on the list, our goal is to find the RNA secondary structure with the maximum score. This means that we would like to find the optimal solution among all the possible RNA structures from the conserved stem list. This is not a trivial task, since the stems in the conserved stem list may have conflicts. Sorting the conserved stems by their scores and then always choosing the next stem to be the one with the largest score is not a successful approach.

Studnicka (7) reported a search method to generate the best overall structure based on a "hyperstructure matrix." Our idea is to use the dynamic program-

ming method to build the secondary structures from the bottom. For each stem in the conserved stem list, we consider the optimal secondary structure with this stem to be the root of the structure. This means that we have to find the optimal substructure induced by each stem in the conserved stem list. Each stem is represented as a node; thus we use a tree or forest to represent a RNA structure (18). Since this is a bottom-up process, we have to choose a suitable order; that is, we process the conserved stem list in such a way that when we consider a new stem, all stems inside the sequence that is induced by the new stem have already been processed.

The sketch of the algorithm is described as follows:

1. We start by sorting all the stems by their 3' end positions. We call this list  $T[]$  and its length  $N$ . This sorted list has the property that if  $i < j$  then  $T[i].3end \leq T[j].3end$ . This property will guarantee that when we consider stem  $T[i]$ , all of the stems inside the sequence induced by  $T[i]$  have already been considered.
2. Consider each of the stems  $T[k]$  ( $T[k].5end$ ,  $T[k].3end$ ,  $T[k].size$ ,  $T[k].score$ ) in the sorted order. We compute the optimal substructure from nucleotide  $T[k].5end$  to  $T[k].3end$  by using this stem as the root. This is the same as computing the optimal substructure for the sequence induced by  $T[k]$ . The score of the resulting optimal substructure will be stored in  $M[k]$ .
  - 2a. Determine the list of stems that are inside the sequence induced by the stem  $T[k]$ . Call the new list  $t[]$  and its length  $n$ .
  - 2b. Scan the list  $t[]$  and do the following for each index  $l$ :
    - (i) Find the largest index  $ll$  such that  $t[ll].3end < t[l].5end$ .
    - (ii) Let  $lll$  be the index such that stem  $t[l]$  and stem  $T[lll]$  are the same.
    - (iii) Let  $m[l] = \max\{m[l-1], m[ll] + M[lll]\}$ .  
(Since  $lll$  is less than  $k$ , we know that  $M[lll]$  is available.)
  - 2c.  $M[k] = m[n] + T[k].score$ .
3. Determine the optimal secondary structure. The maximum score is stored in  $m[N]$ . This step is similar to 2b.
 

Scan the list  $T[]$  and do the following for each index  $l$ :

  - (i) Find the largest index  $ll$  such that  $T[ll].3end < T[l].5end$ .
  - (ii) Let  $m[l] = \max\{m[l-1], m[ll] + M[l]\}$ .

The above sketch only describes how to find the score of the optimal folding. We can recover the actual optimal folding by a simple backtracking. The structure with the maximum score is produced in  $O(N^2)$  time and  $O(N)$  space. The program was written in C and runs in a UNIX environment.

## RESULTS AND DISCUSSION

In the prediction of possible thermodynamically favored helical stems for RNA folding, uncertainties of energy parameters for estimating the formation of RNA secondary structural elements are assumed to follow a normal distribu-

tion. The free energy parameters from Turner's energy rule are perturbed about their tabular values according to the normal distribution within the ranges of the experimental errors. In general, 50 "simulated energy rules" may be a suitable sample size for the simulation of a RNA folding (14). The lowest free energy structure is searched for each "simulated energy rule." In the simulation, the computed lowest free energies (say 50 observations) follow an approximately normal distribution. All thermodynamically favored helices that occurred in the simulation are compiled and called the "stem list." The program EFFOLD requires a considerable amount of computing time for a large RNA molecule. Currently, the EFFOLD can be performed on VAX, CRAY, and MasPar parallel computer systems.

The program MATCH requires two input data: an alignment of homologous RNAs generated by Zuker's MAL program or other available programs and a stem list as mentioned as above. Based on the MAL of homologous RNAs, the phylogenetic conserved helices are selected from the input stem list. During examination, the sequence position of each helix is allowed to shift within a small adjustable range so that the bias in the sequence alignment can be partially eliminated. The information from compensatory base changes in these phylogenetic helices is saved. This program produces a list that includes all conserved stems among these homologous sequences found in the input stem list. Table 1 shows two input data files used in the program MATCH. The first input is a MAL of 34 prokaryotic 5S rRNA sequences (19), and the second is a stem list of thermodynamically favored helical stems computed in the RNA folding simulation of *Bacillus subtilis* (*B. sub*) 5S rRNA. Table 2 shows its output, a phylogenetically conserved stem list. In this example, we generated the stem list from both thermodynamically favored and phylogenetically conserved helical stems among 34 5S rRNAs.

The program BUILD reads a stem list file generated by the program MATCH, counts scores of stem size (S1), compensatory base changes of the stem (S2), and the bonus of covariation found in the stem (S3), and builds a RNA structure with the maximum score. Currently, BUILD has four choices. First, it constructs a RNA secondary structure with the maximum S1 in the list. Second, it constructs a RNA secondary structure with the maximum score of S3 prior to S1. Third, it constructs a RNA secondary structure with the maximum score in the order of S2, S3, and S1. Fourth, it constructs a RNA secondary structure with the maximum score in the order of S2, S1, and S3. The outputs of BUILD are shown in Table 3.

Using this method, we have computed several common RNA foldings. Our examples include the conserved RNA structures folded in the internal ribosomal entry sites of picornavirus (20, 21) and infectious bronchitis virus RNAs (22). In this paper, we show an example for folding 34 prokaryotic 5S RNAs. The stem lists of common RNA foldings from 34 5S RNAs computed using four optimized techniques are listed in Table 3. The four models of RNA secondary structures are almost identical to each other. The computed common RNA secondary structure of 5S RNA is consistent with the classic phylogenetic structure model (23). The deduced common folding is shown in Fig. 1.

TABLE 1

TWO INPUT DATA FILES FOR RUNNING THE PROGRAM MATCH

(a)	10	20	30	40	50	60	70
#14 B.sub	...	uuuggggggg	gaaagcgaag	gggucCaCccg	.uucccaUacCGAaCac	gggaaguu	lagcucucag
#1 E.coli	...	ugccuggggc	gaaagcggcg	ggggucCaCcu	ga. ccccaUgcCGAaCuc	agaaguu	laacgcggcuag
#2 P.vulgar	...	ugcugggggc	cauagcgcag	ggggucCaCcu	ga. ucccaUgcCGAaCuc	agaaguu	laacgcggcuag
#3 Photobac	...	ugcuuggggc	cauagcggcu	uagggaccCaCcu	ga. ucccaUgcCGAaCuc	agaaguu	laacgcggcuag
#4 Beneckia	...	ugcuuggggc	cauagcggcu	uagggaccCaCcu	ga. ucccaUgcCGAaCuc	agaaguu	laacgcggcuag
#5 P. fluor	...	uuuuugagc	acauagagca	uugggaaCaCcu	ga. ucccaUccCGAaCuc	agaaguu	laacgcggcuag
#6 Azo.vine	...	ugcuugagc	cauagagc	gguugaaCaCcu	ga. ucccaUccCGAaCuc	agaaguu	laacgcggcuag
#7 P.aerugi	...	ugcuugagc	cauagagc	gguugaaCaCcu	ga. ucccaUccCGAaCuc	agaaguu	laacgcggcuag
#10 Th.aquat	...	aaucoccccg	ugcccauagc	ggcgugggaaCaCccg	.uucccaUccCGAaCac	gggaaguu	lagcgcggcag
#11 Th.therm	...	aaucoccccg	ugcccauagc	ggcgugggaaCaCccg	.uucccaUccCGAaCac	gggaaguu	lagcgcggcag
#8 Paraco	...	gucucggggc	caaaagcagc	ggcaaaaCaCccg	.uucccaUccCGAaCuc	ggggcgguu	laggucggcuag
#9 R.rubrum	...	uggcguggggc	cauuagcggg	cucgaaaCaCccg	.uucccaUccCGAaCuc	ggggcgguu	laggaaagccuag
#15 B.lichen	...	uuuggggggc	gaaagcgaag	gggucCaCccg	.uucccaUgcCGAaCac	gggaaguu	lagcucucag
#16 B.stearo	...	ccuagugac	cauagcggg	agggaaCaCccg	.uucccaUccCGAaCac	gggaaguu	lagcucucag
#19 Lac.brev	...	uugggggc	gaaagcggc	gaaCaCcu	ga. ucccaUgcCGAaCac	agaaguu	lagcucucag
#20 Strept.f	...	uugggggc	gaaagcggc	gaaCaCcu	ga. ucccaUgcCGAaCac	agaaguu	lagcucucag
#18 Lac.viri	...	uugggggc	gaaagcggc	gaaCaCcu	ga. ucccaUgcCGAaCac	agaaguu	lagcucucag
#17 B.acidoc	...	ucugggggc	cauuagcggg	gggcaCaCccg	.uucccaUccCGAaCac	ggggcgguu	lagcgcggcag
#21 B.brevis	...	ucugggggc	cauuagcggg	gggcaCaCccg	.uucccaUccCGAaCac	ggggcgguu	lagcgcggcag
#23 Spiropla	...	ucugggggc	cauuagcggg	gggcaCaCccg	.uucccaUccCGAaCac	ggggcgguu	lagcgcggcag
#24 Mycop.ca	...	uuggggg	gaaagcgaag	gggucCaCcu	ga. ucccaUgcCGAaCac	agaaguu	lagcucucag
#25 Mycop.my	...	uuggggg	gaaagcgaag	gggucCaCcu	ga. ucccaUgcCGAaCac	agaaguu	lagcucucag
#22 C.pasteu	...	uccaguguc	cauagcau	aggguaaCaCucc	.uucccaUccCGAaCac	gggaaguu	lagcucucag
#26 Micrococ	...	guuacggggc	cauuagcggg	gggaaCaCccg	.gccguUauCGAaCcc	gggaaguu	lagcgcggcag
#27 Streptom	...	guuucggggc	cauuagcggg	gggaaCaCccg	.guuUauCGAaCcc	gggaaguu	lagcgcggcag
#12 Anacysti	...	ucugggggc	cauuagcggg	gggaaCaCccg	.uucccaUccCGAaCac	ggggcgguu	lagcgcggcag
#13 Prochlor	...	uccgggggc	cauuagcggg	gggaaCaCccg	.uucccaUccCGAaCac	ggggcgguu	lagcgcggcag
#29 Wbb.sm12	...	uagggggggc	gaaagcggg	gaaCaCcu	ga. ucccaUccCGAaCac	agaaguu	lagcucucag
#30 Wsp.hung	...	uccaaagcggc	gcaagcggg	gucCaCcu	ga. ucccaUccCGAaCac	agaaguu	lagcucucag
#31 Hb.salini	...	uuuaa.ggg	ggcccaagc	gggguaCuCccg	.uucccaUccCGAaCac	gggaaguu	lagcgcggcag
#32 Hc.morrh	...	uuuaa.ggg	ggcccaagc	gggguaCuCccg	.uucccaUccCGAaCac	gggaaguu	lagcgcggcag
#33 T.acidoc	...	ggca.c.ac	ggggaagc	gggcaCaCccg	.uucccaUccCGAaCac	gggaaguu	lagcgcggcag
#34 S.acidoc	...	ggccaccggg	caagcggg	gggcaCaCccg	.gacucaUccCGAaCcc	gggaaguu	lagcgcggcag
#28 Wheat.mt	...	aaaccgggca	caagcggg	gggcaCaCccg	.uucccaUccCGAaCac	gggaaguu	lagcgcggcag

	80	90	100	110	120	130
#14 B.sub	cg...c	cgauagg	uagucgggg	.guuuc	ccccuGmg	gagaguu
#1 E.coli	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#2 P.vulgar	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#3 Photobac	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#4 Beneckia	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#5 P. fluor	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#6 Azo.vine	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#7 P.aerugi	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#10 Th.aquat	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#11 Th.therm	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#8 Paraco	cg...c	caauagg	uacu.gcg	.g	ccaaagcggc	gucgag
#9 R.rubrum	cg...c	caauagg	uacu.gcg	.g	ccaaagcggc	gucgag
#15 B.lichen	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#16 B.stearo	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#19 Lac.brev	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#20 Strept.f	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#18 Lac.viri	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#17 B.acidoc	cg...c	caauagg	uacu.gcg	.g	ccaaagcggc	gucgag
#21 B.brevis	cg...c	caauagg	uacu.gcg	.g	ccaaagcggc	gucgag
#23 Spiropla	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#24 Mycop.ca	cg...g	ugaaaga	uaau.....	acugau	gudaga	aaauagc
#25 Mycop.my	cg...g	ugaaaga	uaau.....	acugau	gudaga	aaauagc
#22 C.pasteu	ug...c	ugauagg	uacugcagg	.gg	aaagcc	cuugGmg
#26 Micrococ	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#27 Streptom	cg...c	cgauagg	uagucgggg	.g	ggucucccc	cauGmg
#12 Anacysti	cg...g	caacga	uagcuccgg	.gg	uaagcc	ggucGmg
#13 Prochlor	cg...g	caacga	uagcuccgg	.gg	uaagcc	ggucGmg
#29 Wbb.sm12	cguaau	.g	cgauagg	uagucgggg	.g	ggucucccc
#30 Wsp.hung	cguaau	.g	cgauagg	uagucgggg	.g	ggucucccc
#31 Hb.salini	cguaau	.g	cgauagg	uagucgggg	.g	ggucucccc
#32 Hc.morrh	cguaau	.g	cgauagg	uagucgggg	.g	ggucucccc
#33 T.acidoc	cguaau	.g	cgauagg	uagucgggg	.g	ggucucccc
#34 S.acidoc	cguaau	.g	cgauagg	uagucgggg	.g	ggucucccc
#28 Wheat.mt	cguaau	.g	cgauagg	uagucgggg	.g	ggucucccc

TABLE 1—Continued

Number	5-end	3-end	Size	Ratio	Freq
(b)					
All possible stems occurring in RNA foldings					
1	1	115	10	0.420	21
2	2	114	9	0.580	29
3	14	66	2	1.000	50
4	16	63	6	0.980	49
5	16	63	5	0.020	1
6	22	58	1	0.020	1
7	24	57	1	0.020	1
8	25	53	2	0.780	39
9	27	53	2	0.020	1
10	28	49	2	0.560	28
11	29	49	4	0.440	22
12	31	45	4	0.560	28
13	33	43	2	0.400	20
14	68	81	5	0.980	49
15	68	104	2	0.020	1
16	71	101	2	0.020	1
17	75	97	1	0.020	1
18	77	95	2	0.020	1
19	80	92	5	0.020	1
20	82	90	3	0.980	49
21	91	104	5	0.980	49
The frequently recurring stems in the list. The occurring ratio is greater than 0.06.					
1	1	115	10	0.420	21
2	2	114	9	0.580	29
3	14	66	2	1.000	50
4	16	63	6	0.980	49
5	25	53	2	0.780	39
6	28	49	2	0.560	28
7	29	49	4	0.440	22
8	31	45	4	0.560	28
9	33	43	2	0.400	20
10	68	81	5	0.980	49
11	82	90	3	0.980	49
12	91	104	5	0.980	49

Note. (a) Multiple sequence alignment of 34 prokaryotic 5S rRNAs. (b) Stem list of thermodynamically favored helical stems in a RNA folding simulation of *B. sub* 5S rRNA computed by the program EFFOLD. In this simulation, 50 "simulated energy rules" for RNA folding were generated.

To compare our approach with the classical phylogenetic comparative procedure, we generated a stem list based on base pairing information without considering the free energy of formation. This stem list contained 562 stems, in which a stem composed of only 1 base pair was not included. Also, only one stem in a set of consecutive overlapping stems was included to represent this set of

TABLE 2

CONSERVED STEM LIST IN WHICH THE RATIO OF THE FREQUENTLY RECURRING HELICAL  
STEMS IN 34 SEQUENCES IS GREATER THAN 75%

The multiple sequence alignment file: bsu5s.aln

The stem list file (RNA structure): bsu.lst

Matching factor for assessing a legal helix in multiple sequence is: 0.80

The minimum size of a legal helical stem is: 2

The maximum shift size in 5-end\_left is: 2

The maximum shift size in 5-end\_right is: 2

The maximum shift size in 3-end\_left is: 2

The maximum shift size in 3-end\_right is: 2

Master seq. no.	Pos 5'	Pos 3'	Length	Ratio	Covariations	
					Bp. No.	Bonus
1	75	97	1	1.000	1	1.0
1	27	53	2	0.971	2	3.0
1	77	95	2	0.971	2	3.0
1	16	63	6	0.912	19	31.5
1	16	63	5	0.912	16	26.0
1	22	58	1	0.912	1	1.0
1	29	49	4	0.912	3	4.0
1	80	92	5	0.882	16	27.0
1	82	90	3	0.882	7	11.5
1	1	115	10	0.824	15	23.0
1	2	114	9	0.824	20	32.0
1	14	66	2	0.824	3	4.0
1	68	104	2	0.794	1	2.0
1	33	43	2	0.765	0	0.0

*Note.* This is an output file from the program MATCH. The matching factor, minimum size, and maximum shift sizes of the stems are input parameters. The matching factor for assessing a legal helix allows the helix to span in a small range. The stem with one base pair was processed in such a way that exact matching was required. For example, when the matching factor is set to greater than or equal to 0.8, a helix of five base pairs (length is five) is considered to be matched in the other sequence if only four or more conserved or equivalent base-pairs out of five base pairs can be found. The minimum size of a legal helical stem of 2 means that there are at least two successive base pairs in the matching helix. For example, the stem lengths of matching patterns 11011 and 10111 of a stem in two other sequences are assigned to 4 and 3 respectively, where 1 means base pairing and 0 means no base pairing. The master sequence number (Master seq. No.) means that the *B. sub* is located at the first line in the multiple sequence alignment. The ratio of the frequently recurring helical stem in 34 sequences is listed in the fifth column. The base pair number (bp. No.) and bonus in the column of covariations represent the number of equivalent base pairs and the specific base-pairing bonus of compensatory base changes found in 34 sequences for the stem. The bonus of bp G-C changed to C-G is assigned to 2, G-C to U-A 2, G-C to A-U 1, G-C to U-G 1.5, A-U to U-A 2, A-U to U-G 1.5, and U-G to G-U 1.5.

TABLE 3  
 OUTPUTS OF COMMON RNA FOLDING FOR 34 5S RNAs BY BUILD

0	13	33	43	2	0.765000	0.000000
1	6	29	49	4	0.912000	4.000000
2	1	27	53	2	0.971000	3.000000
3	5	22	58	1	0.912000	1.000000
4	4	16	63	5	0.912000	26.000000
5	3	16	63	6	0.912000	31.500000
6	11	14	66	2	0.824000	4.000000
7	8	82	90	3	0.882000	11.500000
8	7	80	92	5	0.882000	27.000000
9	2	77	95	2	0.971000	3.000000
10	0	75	97	1	1.000000	1.000000
11	12	68	104	2	0.794000	2.000000
12	10	2	114	9	0.824000	32.000000
13	9	1	115	10	0.824000	23.000000

$i = 0$  weight = 2

$i = 1$  weight = 6

$i = 2$  weight = 8

$i = 3$  weight = 9

$i = 4$  weight = 14

$i = 5$  weight = 14

$i = 6$  weight = 16

$i = 7$  weight = 3

$i = 8$  weight = 5

$i = 9$  weight = 7

$i = 10$  weight = 8

$i = 11$  weight = 10

$i = 12$  weight = 35

$i = 13$  weight = 36

final weight = 36

(a) Final tree for stem size: 36

0	9	1	115	10	0.824000	23.000000
1	11	14	66	2	0.824000	4.000000
2	3	16	63	6	0.912000	31.500000
3	1	27	53	2	0.971000	3.000000
4	6	29	49	4	0.912000	4.000000
5	13	33	43	2	0.765000	0.000000
6	12	68	104	2	0.794000	2.000000
7	0	75	97	1	1.000000	1.000000
8	2	77	95	2	0.971000	3.000000
9	7	80	92	5	0.882000	27.000000

Final weight = 107.500000 35

(b) Final tree weight = 107.500000 35

0	10	2	114	9	0.824000	32.000000
1	11	14	66	2	0.824000	4.000000
2	3	16	63	6	0.912000	31.500000
3	1	27	53	2	0.971000	3.000000
4	6	29	49	4	0.912000	4.000000
5	13	33	43	2	0.765000	0.000000
6	12	68	104	2	0.794000	2.000000
7	0	75	97	1	1.000000	1.000000
8	2	77	95	2	0.971000	3.000000
9	7	80	92	5	0.882000	27.000000

TABLE 3—Continued

Final weight: b_no = 10 bonus = 103.000000 size = 35						
(c) Final tree weight = 10 103.000000 35						
0	10	2	114	9	0.824000	32.000000
1	11	14	66	2	0.824000	4.000000
2	4	16	63	5	0.912000	26.000000
3	5	22	58	1	0.912000	1.000000
4	1	27	53	2	0.971000	3.000000
5	6	29	49	4	0.912000	4.000000
6	13	33	43	2	0.765000	0.000000
7	12	68	104	2	0.794000	2.000000
8	0	75	97	1	1.000000	1.000000
9	2	77	95	2	0.971000	3.000000
10	7	80	92	5	0.882000	27.000000
Final weight: b_no = 10 size = 36 bonus = 94.000000						
(d) Final tree weight = 10 36 94.000000						
0	9	1	115	10	0.824000	23.000000
1	11	14	66	2	0.824000	4.000000
2	4	16	63	5	0.912000	26.000000
3	5	22	58	1	0.912000	1.000000
4	1	27	53	2	0.971000	3.000000
5	6	29	49	4	0.912000	4.000000
6	13	33	43	2	0.765000	0.000000
7	12	68	104	2	0.794000	2.000000
8	0	75	97	1	1.000000	1.000000
9	2	77	95	2	0.971000	3.000000
10	7	80	92	5	0.882000	27.000000

*Note.* (a) The common RNA secondary structure was deduced by the maximum score of stem size. The size of the total base pairs in the structure was 36. The 5'-end and 3'-end positions of the stem and stem size are listed in the third, fourth, and fifth column, respectively. The number in the second column represents the order of the stem in the input stem list. The numbers in the sixth and seventh columns represent the ratio and bonus of compensatory base changes (see Table 2). (b) The common RNA secondary structure was deduced by the maximum scores in the order of equivalent base pairing bonus and stem size. The maximum bonus (see legend to Table 2) was 107 and the total stem size of the structure was 35. (c) The common RNA secondary structure was deduced by the maximum scores in the order of covariation, bonus of compensatory base changes, and stem size. The 10 of 11 stems found in the common secondary structure were supported by an inspection of the covariation found in 34 5S RNAs. The bonus from equivalent base pairing and the total stem size found in the structure were 103 and 35, respectively. (d) The common RNA secondary structure was deduced by the maximum scores in the order of covariation, stem size, and equivalent base pairing bonus. The 10 of 11 stems of the common secondary structure were supported by an inspection of the covariation found in 34 5S RNAs. The total stem size of the structure was 36 and the bonus was 94.

stems. For instance, if we had a set of stems, 1-115(10), 2-114(9), ..., 8-108(3), 9-107(2), only the stem 1-115(10) with the maximum size (10 base pairs) was included in the list. The conserved stem list derived by the program MATCH included 101 stems. A possible common RNA folding computed by the program BUILD with the maximum stem size is listed in Table 4. The predicted common RNA folding for 34 prokaryotic 5S rRNAs is similar to that displayed in Fig.

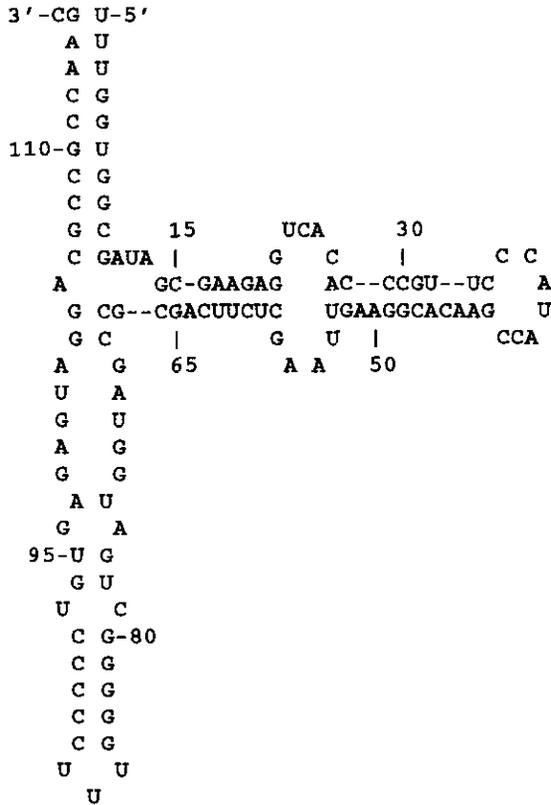


FIG. 1. A common RNA secondary structure model derived from 34 5S RNA. Sequence data in the figure are based on *B. sub* 5S rRNA. The model is based on the data displayed in Table 3a.

1. This comparison shows that the stem list generated by EFFOLD is a good pool of base pairing for constructing a common structure supported by phylogenetic comparative analysis.

The program MATCH uses a window with a small shift in the positions of the stem under inspection. A little uncertainty in the MAL for a set of homologous RNAs can be eliminated. The predicted conserved stems may have a little bias toward the selected sequence in the input stem list. As a rule of thumb, the most similar sequence compared with other sequences in a set of homologous RNAs is selected. When an unreliable MAL with large uncertainty is produced from a MAL computer program, the program will not give a reasonable result. For example, we failed to generate a correct phylogenetic structure model for RNase P RNAs. Obviously, a robust MAL is crucial in predicting a reliable common folding.

Given a score scheme, the program BUILD can produce the optimal solution with respect to the score scheme. In the current version of BUILD, we have introduced three scores. A more precise score function is being examined by

TABLE 4  
THE OUTPUT OF A COMMON RNA FOLDING FOR 34 5S RNAs BY BUILD

0	50	2	14	3	0.824000	14.000000
1	51	5	15	2	0.824000	0.000000
2	83	6	16	2	0.765000	0.000000
3	81	1	19	3	0.765000	14.000000
4	47	1	20	3	0.824000	14.000000
5	65	1	21	3	0.794000	14.500000
6	25	21	29	3	0.882000	10.000000
7	40	21	30	2	0.853000	0.000000
8	1	3	31	3	0.971000	17.500000
9	85	10	33	3	0.765000	16.000000
10	71	21	33	3	0.794000	15.000000
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
96	9	94	111	3	0.971000	18.000000
97	98	83	112	3	0.765000	16.000000
98	3	6	113	5	0.971000	20.000000
99	49	1	115	10	0.824000	23.000000
100	93	31	116	3	0.765000	7.000000

$i = 0$  weight = 3

$i = 1$  weight = 2

$i = 2$  weight = 2

$i = 3$  weight = 5

$i = 4$  weight = 5

$i = 5$  weight = 5

.....

.....

.....

$i = 96$  weight = 3

$i = 97$  weight = 6

$i = 98$  weight = 31

$i = 99$  weight = 35

$i = 100$  weight = 22

Final weight = 35

Final tree for stem size: 35

0	49	1	115	10	0.824000	23.000000
1	55	14	66	2	0.824000	4.000000
2	15	16	63	6	0.912000	31.500000
3	4	27	53	2	0.971000	3.000000
4	16	29	49	4	0.912000	4.000000
5	76	68	104	2	0.794000	2.000000
6	77	75	102	2	0.794000	0.000000
7	63	78	97	3	0.824000	14.000000
8	35	81	93	4	0.882000	18.000000

*Note.* The common RNA secondary structure was deduced by the maximum score of the stem size. The input conserved stem list (containing 101 stems) was generated by MATCH, based on the same MAL listed in Table 1a and a stem list that includes all possible stems with the minimum size of two base pairs. The stem list includes 562 stems and is not shown here. The output was edited to save space. For further details, see the legend to Table 3.

using a large number of divergent RNA sequences. Possible tertiary interactions in the common folding can be added, based on a statistical test that evaluates their statistical significance (24). This algorithm is currently being extended so that it can find a set of structures within a certain level of scores and include tertiary interactions.

Recently, an algorithm for the prediction of common folding structures of homologous RNAs has been reported (25). However, this algorithm is totally dependent on a correct multiple sequence alignment. Also, this algorithm is not guaranteed to give an optimal solution for building the RNA structure. The prediction of RNA folding requires a progressive refinement process, as our knowledge of RNA folding is limited. Our programs can provide a tool for predicting a robust working model of RNA common folding for further refinement. They maintain the advantage of dynamic programs and almost completely automate current phylogenetic procedures.

#### ACKNOWLEDGMENTS

We are very grateful to Jih-H. Chen for his useful discussion. Jun Zhang performed part of the programming work. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

#### REFERENCES

1. MICHEL, F., AND WESTHOF, E. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.* **216**, 585 (1990).
2. GUTELL, R. R., WEISER, B., WOESE, C. R., AND NOLLER, H. F. Comparative anatomy of 16S-like ribosomal RNA. *Prog. Nucleic Acid Res. Mol. Biol.* **32**, 155 (1985).
3. STERN, S., WEISER, B., AND NOLLER, H. F. Model for the three-dimensional folding of 16S rRNA. *J. Mol. Biol.* **204**, 447 (1988).
4. WESTHOF, E., ROMBY, P., ROMANIUK, P. J., EBEL, J.-P., EHRESMANN, C., AND EHRESMANN, B. Computer modelling from solution data of spinach chloroplast and of *Xenopus laevis* somatic and oocyte 5S rRNAs. *J. Mol. Biol.* **207**, 417 (1989).
5. JAMES, B. D., OLSEN, G. J., AND PACE, N. R. Phylogenetic comparative analysis of RNA secondary structure. *Methods Enzymol.* **180**, 227 (1989).
6. WATERMAN, M. S., AND SMITH, T. F. RNA secondary structure: A complete mathematical analysis. *Math. Biosci.* **42**, 257 (1978).
7. STUDNICKA, G. M., RAHN, G. M., CUMMINGS, I. W., AND SALSER, W. A. Computer method for predicting the secondary structure of single-stranded RNA. *Nucleic Acids Res.* **5**, 3365 (1978).
8. NUSSINOV, R., AND JACOBSON, A. B. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. USA* **77**, 6309 (1980).
9. ZUKER, M., AND STEIGLER, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9**, 133 (1981); ZUKER, M. On folding all suboptimal folding of an RNA molecule. *Science* **244**, 48 (1989).
10. SANKOFF, D., KRUSKAL, J. B., MAINVILLE, S., AND CEDERGREEN, R. J. Fast algorithms to determine RNA secondary structures containing multiple loops. In "Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison" (D. Sankoff and J. B. Kruskal, Eds.), pp. 93-120. Addison-Wesley, New York, 1983.
11. MARTINEZ, H. M. An RNA folding rule. *Nucleic Acids Res.* **12**, 323 (1984).

12. WILLIAMS, A. L., AND TINOCO, I., JR. A dynamic programming algorithm for finding alternative RNA secondary structures. *Nucleic Acids Res.* **14**, 299 (1986).
13. LE, S.-Y., CHEN, J.-H., CURREY, K. M., AND MAIZEL, J. V., JR., A program for predicting significant RNA secondary structures. *CABIOS* **4**, 153 (1988).
14. LE, S.-Y., CHEN, J.-H., AND MAIZEL, J.V., JR. Prediction of alternative RNA secondary structures based on fluctuating thermodynamic parameters. *Nucleic Acids Res.* **21**, 2173 (1993).
15. JAEGER, J. A., TURNER, D. H., AND ZUKER, M. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA* **86**, 7706 (1989).
16. FREIER, S. M., KIERZEK, R., JAEGER, J. A., SUGIMOTO, N., CARUTHERS, M. H., NEILSON, T., AND TURNER, D. H. Improved free-energy parameters for prediction of RNA duplex stability. *Proc. Natl. Acad. Sci. USA* **83**, 9373 (1986).
17. LE, S.-Y., AND ZUKER, M. Common structures of the 5' non-coding RNA in enteroviruses and rhinoviruses: Thermodynamical stability and statistical significance. *J. Mol. Biol.* **216**, 729 (1990).
18. LE, S.-Y., OWENS, J., NUSSINOV, R., CHEN, J.-H, SHAPIRO, B., AND MAIZEL, J. V., JR. RNA secondary structures: Comparison and determination of frequently recurring substructures by consensus. *CABIOS* **5**, 205 (1989).
19. WATERMAN, M. S. Computer analysis of nucleic acid sequences. *Methods Enzymol.* **164**, 765 (1988).
20. LE, S.-Y., CHEN, J.-H., SONENBERG, N. AND MAIZEL, J. V., JR. Conserved tertiary structural elements in the 5' untranslated region of human enteroviruses and rhinoviruses. *Virology* **191**, 858 (1992).
21. LE, S.-Y., CHEN, J.-H., SONENBERG, N., AND MAIZEL, J. V., JR. Conserved tertiary structural elements in the 5' nontranslated region of cardiovirus, aphthovirus and hepatitis A virus RNAs. *Nucleic Acids Res.* **21**, 2445 (1993).
22. LE, S.-Y., SONENBERG, N., AND MAIZEL, J. V., JR. Distinct structural elements and internal entry of ribosomes in mRNA3 encoded by infectious bronchitis virus. *Virology* **198**, 405 (1994).
23. SPECHT, T., WOLTERS, J., AND ERDMANN, V. A. Compilation of 5S rRNA and 5S rRNA gene sequences. *Nucleic Acids Res.* **18**(Suppl.)2215 (1990).
24. CHEN, J.-H., LE, S.-Y., AND MAIZEL, J. V., JR. A procedure for RNA pseudoknot prediction. *CABIOS* **8**, 243 (1992).
25. HAN, K., AND KIM, H.-J. Prediction of common folding structures of homologous RNAs. *Nucleic Acids Res.* **21**, 1251 (1993).
26. LE, S.-Y., AND ZUKER, M. Predicting common foldings of homologous RNAs. *J. Biomol. Struct. Dyn.* **8**, 1027 (1991).