A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison

Michael Zuker, John A.Jaeger[1] and Douglas H.Turner[1]*
Institute for Biological Sciences, National Research Council of Canada, Ottawa, Ontario K1A OR6, Canada and [1]Department of Chemistry, University of Rochester, Rochester, NY 14627, USA

Received January 29, 1991; Revised and Accepted March 27, 1991

ABSTRACT

This article describes the latest version of an RNA folding algorithm that predicts both optimal and suboptimal solutions based on free energy minimization. A number of RNA's with known structures deduced from comparative sequence analysis are folded to test program performance. The group of solutions obtained for each molecule is analysed to determine how many of the known helixes occur in the optimal solution and in the best suboptimal solution. In most cases, a structure about 80% correct is found with a free energy within 2% of the predicted lowest free energy structure.

INTRODUCTION

The number of known RNA sequences is rapidly increasing. Insight into function, however, depends on methods for decoding the information in the sequences. One piece of information is the three dimensional structure of the RNA. In principle, it is possible to predict structure from sequence (1). In practice, it is not only difficult to predict structure (2), but also to determine it experimentally (3,4). Thus only the three dimensional structures of tRNAs are known in detail (5−7).

A first step in modelling the structure of an RNA is determination of the secondary structure, since this provides many constraints. Unfortunately, a huge number of secondary structures are possible for any given sequence. For example, when A, C, G, and U occur randomly with equal probability, the number of valid secondary structures is greater than 1.8^N, where N is the number of nucleotides (8). Thus a sequence of 400 nucleotides has about 10^{102} possible foldings. To make the connection between structure and function, it is necessary to determine the one or few foldings that actually exist. One method that can be used to restrict the number of secondary structures considered is free energy minimization (1,2,9). In principle, this method can predict the equilibrium secondary structure. In practice, only limited experimental data are available for parameterization (2,10), and small changes in energy parameters often result in large changes in predicted foldings. Thus the problem is 'ill-conditioned' in a mathematical sense (11). Moreover, a cell is not at equilibrium, so there is no fundamental reason why the lowest free energy and biologically important structures have to be the same. To cope with these ambiguities, Williams and Tinoco (12) and Zuker (9) developed algorithms to generate a range of suboptimal foldings close to the minimum free energy. Both algorithms use dynamic programming methods (13,14). The Williams and Tinoco approach is based on making alternative choices during the traceback algorithm (15). The Zuker algorithm, MFOLD, relies on predicting all base pairs that are possible in all foldings close to the minimum free energy (9). The programs also use different criteria to exclude similar foldings from the ensemble generated.

This article explores the current effectiveness of free energy minimization as implemented by the algorithm of Zuker (9) with the energy parameters of Freier et al. (10) and the loop model of Jaeger et al. (16). Sequences with secondary structures known from phylogenetic comparisons are folded. Comparisons are made between the phylogenetic structure and both the predicted lowest free energy ('optimal') structure and the suboptimal structure of the computer selected ensemble that is closest to the phylogenetic structure ('single best structure'). A revised criterion is used to limit the number of suboptimal foldings considered. The results suggest that for a sequence with about 400 nucleotides, a structure about 80% correct can be found within an ensemble of about 20 structures that have free energies within 5% of the lowest free energy structure. Most of the time, the single best structure is found within 2% of the free energy of the optimal structure. Thus free energy minimization can provide a reasonable number of rough working models that can be refined with additional experimental data, such as chemical and enzyme modification results (4,17), site directed mutagenesis (18,19), and phylogenetic comparisons (20,21).

MATERIALS AND METHODS

Version 2 of MFOLD was used in this work. It corrects a number of small bugs in the original VAX/VMS version, and has two new features. First, a new program, newtemp, has been added

* To whom correspondence should be addressed

to compute energy files for folding at arbitrary temperatures. An energy computation program, efn, was added to compute the energy of a given folding. It can be used to compute the energy of a phylogenetically determined structure, to reevaluate the stability of a folding at different temperatures or, if modified, with different energy rules. The second feature is a change in the distance criterion used to generate foldings that are not too close to one another. Version 2 runs in VAX/VMS and UNIX environments. It has been ported to the Silicon Graphics personal IRIS model 25S workstation under IRIX 3.2 as well as to Digital Equipment Corporation's DEC 3100 workstation. The UNIX port to the IRIS uses the Silicon Graphics graphics library for the interactive energy dot plot and P-Num plots, while the DECstation version creates these plots with X-window software.

Newtemp (22) creates free energy, ΔG°, files between 0 and 100°C, from published thermodynamic values (2,10,23−25). Unmeasured terminal mismatch enthalpies, ΔH°, are approximated by making the corresponding 3′ dangling end ΔG° more stable by 0.3 kcal/mol at 0°C, and assuming the ΔG° at 37°C as given by Turner, et al. (2). Unmeasured ΔH°'s for two 5′ dangling end sequences were approximated as the average of the measured ΔH°'s for other 5′ dangling ends. The stabilizing ΔG°s for base pairs, dangling ends, and terminal mismatches are extrapolated to t°C with:

$$\Delta G^\circ_t = \Delta H^\circ - (\Delta H^\circ - \Delta G^\circ_{37})(t+273.15)/(37+273.15) \qquad (1)$$

Free energies for bulge, hairpin, internal, and multibranch loops, and asymmetry penalties for internal loops are considered purely entropic, and extrapolated by:

$$\Delta G^\circ_t = \Delta G^\circ_{37}\ (t+273.15)/(37+273.15) \qquad (2)$$

The tetraloop (26,27) ΔH° was approximated from hydrogen bond measurements of Turner et al. (24), extrapolated to the zero stacking limit. The ΔG° for the extra stability of the tetraloops was calculated with eq 1. Free energies of loops of more than 30 nucleotides were calculated with the temperature dependent equation of Jacobson and Stockmayer (28). These approximations give a reasonable prediction of the melting of the self splicing Group I large subunit intron of *Tetrahymena thermophila* (22).

MFOLD uses a distance measure between foldings to ensure that no two foldings are 'too close' to one another. The distance is set by the user. Choosing a small distance might result in the prediction of hundreds or even thousands of suboptimal foldings, many of them similar to one another. A large distance will result in fewer predicted foldings, with the risk of missing some correct folding motifs. The original version of MFOLD (9,29) used a distance criterion defined by Zuker (30), and illustrated by Jaeger et al. (29). In this criterion, the distance between two base pairs, i.j and i′.j′ is defined as max {|i-i′|,|j-j′|}. Two foldings are said to be within a distance d from one another if for every base pair i.j from one folding, there is a base pair i′.j′ from the other within a distance d of i.j. The distance between the two foldings is the smallest d that satisfies this condition. With this distance measure, two foldings can be a large distance apart while differing by only one or two base pairs. To avoid this, the criterion has been modified to demand that the above condition hold for all but d of the base pairs from each folding. Thus, if two foldings are more than a distance d apart, then one of the foldings must contain at least d+1 base pairs that are not within a distance of d of any base pair in the other folding. The distance criterion, called the 'window' parameter in MFOLD, ensures that the automatic generation of suboptimal foldings will not yield structures too close to one another as the minimum distance is increased. The new criterion was inspired by the Prokhorov metric (31) from probability theory.

A scoring program was used to determine the number of helixes and the number of base pairs in the predicted foldings that are in the phylogenetic model. A helix is defined as a double stranded region of at least 3 base pairs interrupted by bulge or interior loops containing at most two unpaired bases each (16). A phylogenetically determined helix is said to be in a computed folding if the computed folding contains all the base pairs of the given helix with the exception of at most two base pairs. Pseudoknots (32) were not included because they cannot be predicted by the algorithm. Computed foldings are ranked first by the number of correct helixes they contain. Different foldings containing the same number of helixes are then ranked by the number of correct base pairs. The scoring program ranked all the suboptimal foldings and automatically determined 'single best' foldings for each sequence.

The entire 1542 nucleotides of E. coli 16S rRNA were folded with the constraint that all the phylogenetic base pairs must occur in the optimal structure. The program uses bonus energies to force constrained base pairs. The default value of the bonus energy is −50.0 kcal/mole of forced base pair. The forcing of several hundred base pairs with such a large bonus energy causes integer overflow because folding energies are stored as two byte integers to save space. The minimum energy of any structure, including all the bonus energies, is therefore bounded below by −3276.7 kcal/mole. Thus, for practical reasons, the bonus energy was set at −2.0 kcal/mole of forced base pair. This was altered to −3.0 kcal/mole when −2.0 proved insufficient to force base pairs G725-C732, C726-G731, and the stem loop structures between nucleotides 289−311 and 316−337.

RESULTS AND DISCUSSION

Table I lists results from folding a representative set of sequences. The percentage of helixes correct for optimal folding of most of the sequences has been discussed previously (16). The exceptions are unmodified yeast phenylalanine tRNA, the entire yeast B1 intron, and the entire E. coli 16S rRNA. These are discussed below.

Previous foldings of tRNAs with MFOLD have made use of the information that some modified nucleotides cannot base pair (16). This is often a severe constraint. Recently Sampson and Uhlenbeck (33) showed that unmodified yeast phenylalanine tRNA also folds into the cloverleaf structure. As shown in Table I, the cloverleaf is also the optimal structure predicted by MFOLD. Thus the algorithm performs well on this tRNA sequence without constraints.

The 2 domains of the yeast B1 intron (34) and the 4 domains of E. coli 16S rRNA (4) have previously been folded separately with MFOLD giving 68 and 63%, respectively, of the phylogenetic helixes in the optimal structures (16). The results in Table I show that only 45 and 58%, respectively, of the known helixes are predicted correctly when the entire 768 and 1542 nucleotides of B1 and 16S are folded. The best suboptimal structure for B1, however, contains 65% of the known helixes and is only 0.6% away in free energy. The results suggest prior knowledge of domain structure can be a useful constraint for folding algorithms.

Figures 1−3 show comparisons of the predicted optimal and best suboptimal structures with the phylogenetically deduced

Table I. Comparison of structures predicted by energy minimization and determined by phylogenetic comparisons[a].

		optimal folding					single best folding			
RNA	folded nucleotides	-ΔG° kcal/mol	correct + slipped/ phylo. helixes[b]	correct/ phylo. b.p.[b]	distance parameter	no. of structures computed[b]	position on list[b]	% difference from opt. ΔG°[b]	correct + slipped/ phylo. helixes[b]	correct/ phylo. b.p.[b]
unmod tRNA(phe)	1-76		4/4	21/21	3	7	1	0.0	4/4	21/21
spinach chloroplast 5S	1-122	32.2	4(5)/5	(30(35)/37)	2	10(1)	1(1)	0.0(0.0)	4(5)/5	(30(35)/37)
E. coli 16S rRNA										
Domain 1	26-557	148.7	15+2(20+2)/24	98(124)/155	5	107(89)	8(8)	0.9(0.5)	20+1(21+3)/24	125(126)/155
Domain 2	561-913	108.5	10+2(11+1)/15	79(85)/104	5	37(13)	3(4)	1.1(1.8)	13+2(14+1)/15	94(100)/104
Domain 3	913-1397	126.4	13+1(10)/22	79(49)/136	5	105(101)	23(30)	4.4(3.6)	17+1(16+4)/22	102(98)/136
Domain 4	1397-1542	42.5	3(3)/4	32(32)/39	5	4(4)	1(1)	0.0(0.0)	3(3)/4	32(32)/39
Total			41+5/65	288/434					53+4/65	353/434
E. coli 16S rRNA	1-1542	432.6	38(38)/65	247/434	10	145	3	0.9	47(50)/65	325/434
C.r. chloroplast										
Domain 1	27-509	136.2	13+2/24	75/138	5	113	13	2.0	17+3/23	103/138
Domain 2	515-857	92.1	9/17	57/102	5	27	5	2.2	17/17	100/102
Domain 3	866-1326	114.5	2+1/21	14/128	5	100	13	1.9	14/21	79/128
Domain 4	1329-1476	46.7	2+1/4	27/39	5	6	2	1.9	3+1/4	34/39
T4 td (I)	-8-94+851-1021[c]	63.9	10+1/12	61/77	5	16	1	0.0	10+1/12	61/77
Yeast B1 (II)	1-768	104.6	14+2/31	122/219	10	35	3	0.6	20+3/31	167/219

[a] Phylogenetic structures are given in the indicated references: unmodified yeast tRNA (phe) (33), spinach chloroplast 5S (43), protein free E. coli 16S rRNA (4), Chlamydomonas reinhardii (C.r.) 16S like rRNA (35), T4 td group I self-splicing intron (44), Saccharomyces cerevisiae B1 group II intron (34).

[b] The number following a + sign is the number of slipped or shortened helixes in the predicted structure. Numbers in parentheses are for foldings constrained with chemical modification data.

[c] Nucleotides 95-850 are replaced by NNN

Table II. Additional base pairs formed by energy minimization of E. coli 16S rRNA with phylogenetic base pairs forced.

48-5'-CU-3'	61-5'-GUC-3'	518-5'-CC^{A}GC-3'	570-5'-G-3'	
362-3'-GG-5'	110-3'-CAG-5'	530-3'-GG_CG-5'	880-3'-C-5'	
716-5'-AU-3'	778-5'-AU-3'	811-5'-CG-3'	828-5'-U-3'	921-5'-UG-3'
723-3'-UG-5'	801-3'-UG-5'	818-3'-GC-5'	872-3'-A-5'	1396-3'-AC-5'
935-5'-AC-3'	957-5'-UAA-3'	922-5'-UG-3'	1117-5'-A-3'	
1348-3'-UG-5'	984-3'-AUU-5'	1213-3'-AU-5'	1183-3'-U-5'	
1125-5'-UU-3'	1158-5'-CU-3'	1239-5'-AU-3'	1258-5'-GC-3'	
1146-3'-AA-5'	1178-3'-GG-5'	1298-3'-UG-5'	1277-3'-CG-5'	
1301-5'-UCCGG-3'	1373-5'-GAA-3'	1501-5'-AAGG-3'		
1339-3'-AGGCU-5'	1382-3'-CUU-5'	1541-3'-UUCC-5'		

structures for domain 1 of E. coli 16S rRNA, domain 2 of C. reinhardii chloroplast 16S like rRNA (35), and the yeast B1 intron (34). Predicted helixes present in the phylogenetic structure are boxed. No single best suboptimal structure is drawn for domain 2 of chloroplast 16S because it is essentially identical to the phylogenetic structure. The differences observed between the predicted and known structures are typical for cases where the optimal structure is not particularly good. Similar comparisons for domain 2 of E. coli 16S rRNA and the self splicing Group I large subunit intron from *T. thermophila* have been presented previously (16,22,29).

One common difference between predicted and phylogenetic structures is the presence of one or two helixes with the base pairing region slipped or shorter than the phylogenetic helix (see Table I). In some cases, the phylogenetic structure may reflect a required intermediate in a dynamic process, rather than the equilibrium structure, or depend on interactions with proteins. Based on chemical modification data, the latter has been suggested for the region between nucleotides 289–311 in E. coli 16S rRNA (4). In most cases, however, the thermodynamic parameters for loops probably need small adjustment.

Examples suggesting modifications of loop parameters include helixes with hairpin loops of 4 nucleotides at positions 81 in E. coli 16S rRNA, and 110 and 182 in yeast B1 (see Figures 1 and 3). These helixes end in CUUG or GUAA 'tetraloops.' Some tetraloop sequences are known to have unusual stability (26; Haney & Uhlenbeck, unpublished data) and structure (36,37). Eight of these sequences are given an additional 2 kcal/mol of stability in the folding algorithm. CUUG and GUAA are relatively rare, however, and are not currently included in the set with extra stability. Phylogenetic comparisons suggest CUUG, GUAA and several other loop sequences may also be unusually stable (27,38). Thus it may be necessary to expand the list of hairpin sequences given extra stability in the folding algorithm. Additionally, studies on oligonucleotides indicate the approximations for the sequence dependence of internal loop stability are oversimplified (39,40). In particular, internal loops terminated with GA mismatches are unusually stable due to hydrogen bonding. Several such loops are missed in the predicted structures. The results suggest prediction of structure will improve as we learn more about the sequence dependence of stabililty for single stranded regions.

Another common difference between predicted and phylogenetic structures is the number, size, and location of multibranch loops. For example, the predicted optimal structure for domain 1 of E. coli 16S rRNA misses the multibranch loops containing nucleotides 50, 110, and 180; in domain 2 of C. reinhardii, the loop containing nucleotide 525 is missed; in yeast

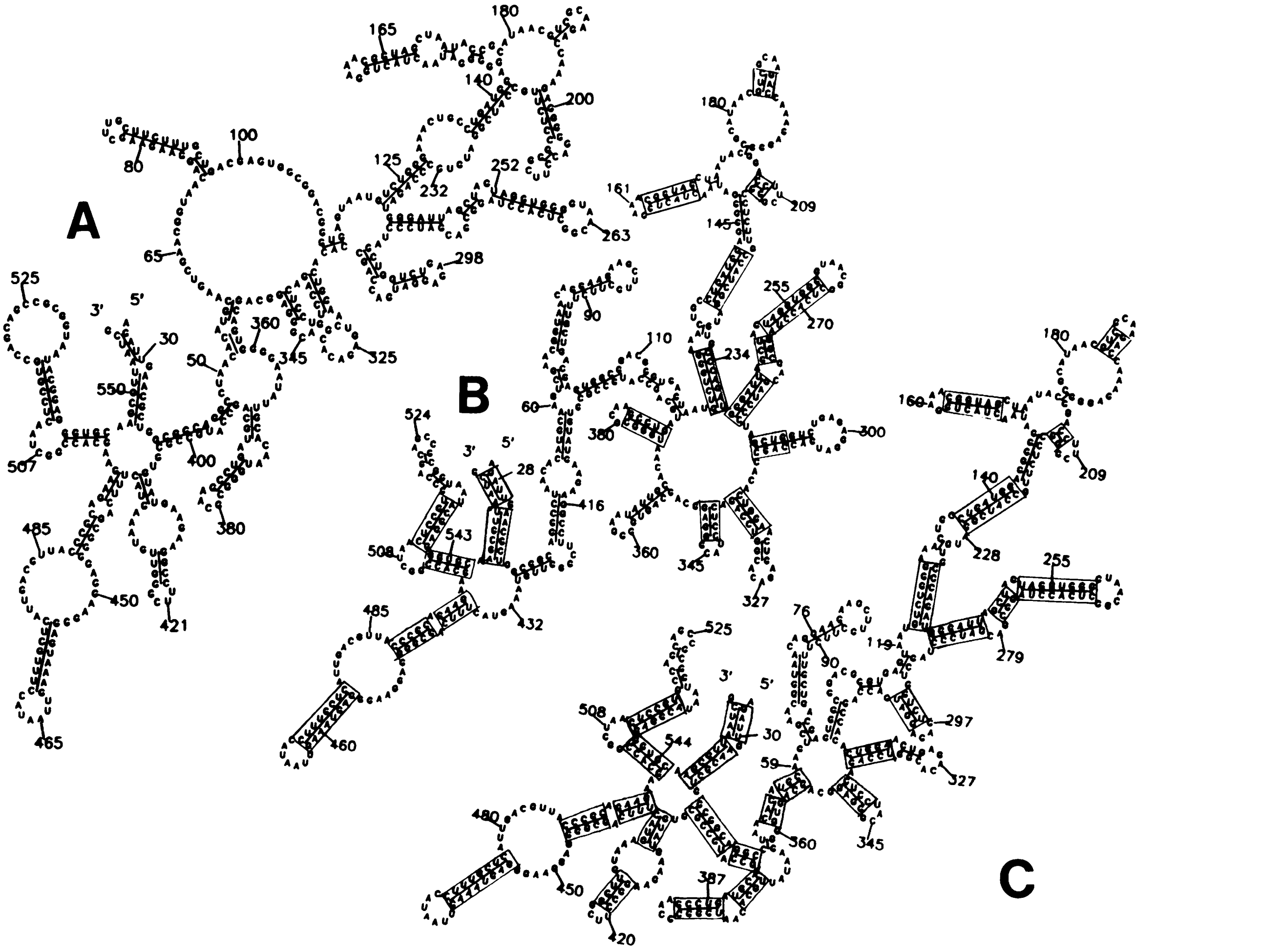


Figure 1. Phylogenetic (A), predicted optimal (B), and best predicted suboptimal (C) structures for domain 1 of E. coli 16S rRNA (4). Phylogenetic helixes in predicted structures are boxed.

B1, the loops containing nucleotides 30, 215, 510, and 630 are missed. This is not surprising because there is no experimental data for sequence and length effects on the stabilities of multibranch loops. Moreover, the algorithm uses an unrealistic linear approximation for the length dependence in order to ensure true energy minimization (9). Melting studies of a circular RNA also have indicated the linear approximation is unrealistic for large loops (22). The results suggest studies of factors affecting the stabilities of multibranch loops should help improve predictions of structure.

The results in Figures 1−3 suggest that even when the predicted optimal structure is not particularly good, a reasonable structure can be found that is only modestly higher in free energy. This is generalized and quantified in Table I. Optimal and suboptimal structures within a 10% window of free energy were generated for 13 sequences. The distance parameter, d, was chosen to reasonably limit the number of structures generated. For the sequences with less than 800 nucleotides, on average the 'single best' structure generated has 80% of the phylogenetically known helixes. The 'single best' structure is always within 5% of the free energy of the optimal structure and always within the first 25 structures generated. Interestingly, domain 3 of E. coli 16S rRNA has the largest difference in free energy between optimal and 'single best' structures, 4.4%. There is evidence that the phylogenetic structure of this domain may not be appropriate for the RNA in the absence of ribosomal proteins. In particular, G951 is strongly modified by kethoxal although it occurs in the middle of a 10 base pair helix (4). Thus it may not be appropriate to compare the predicted structures for this domain with the phylogenetic structure. Eliminating this sequence, the 'single best' folding for sequences less than 800 nucleotides has a free energy on average only 1% higher than the optimal folding.

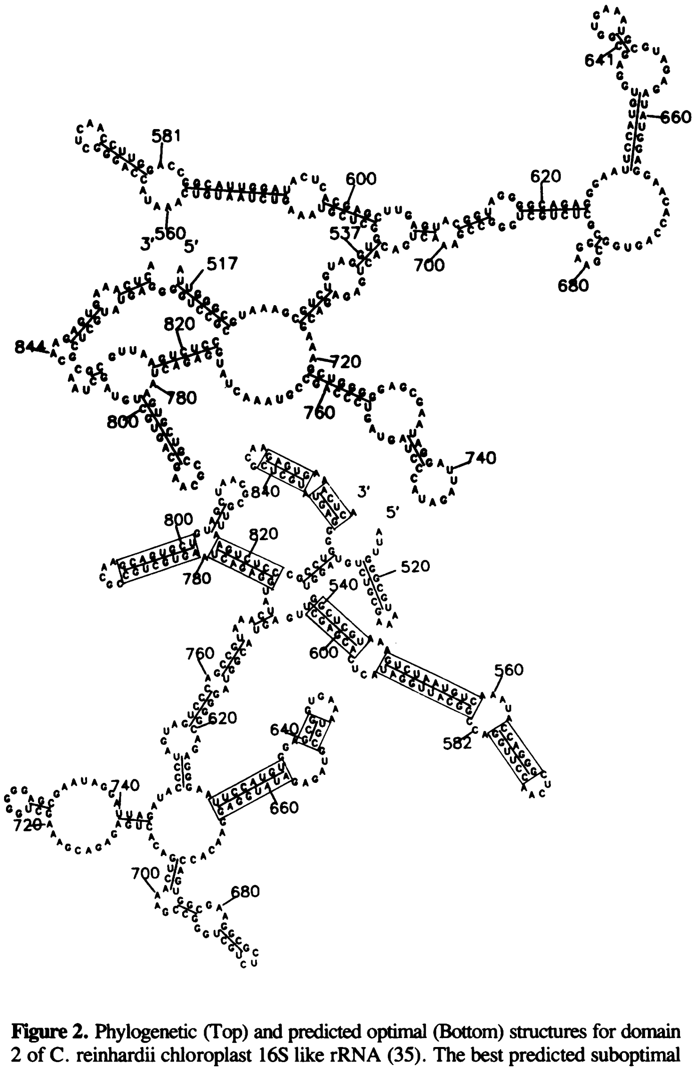


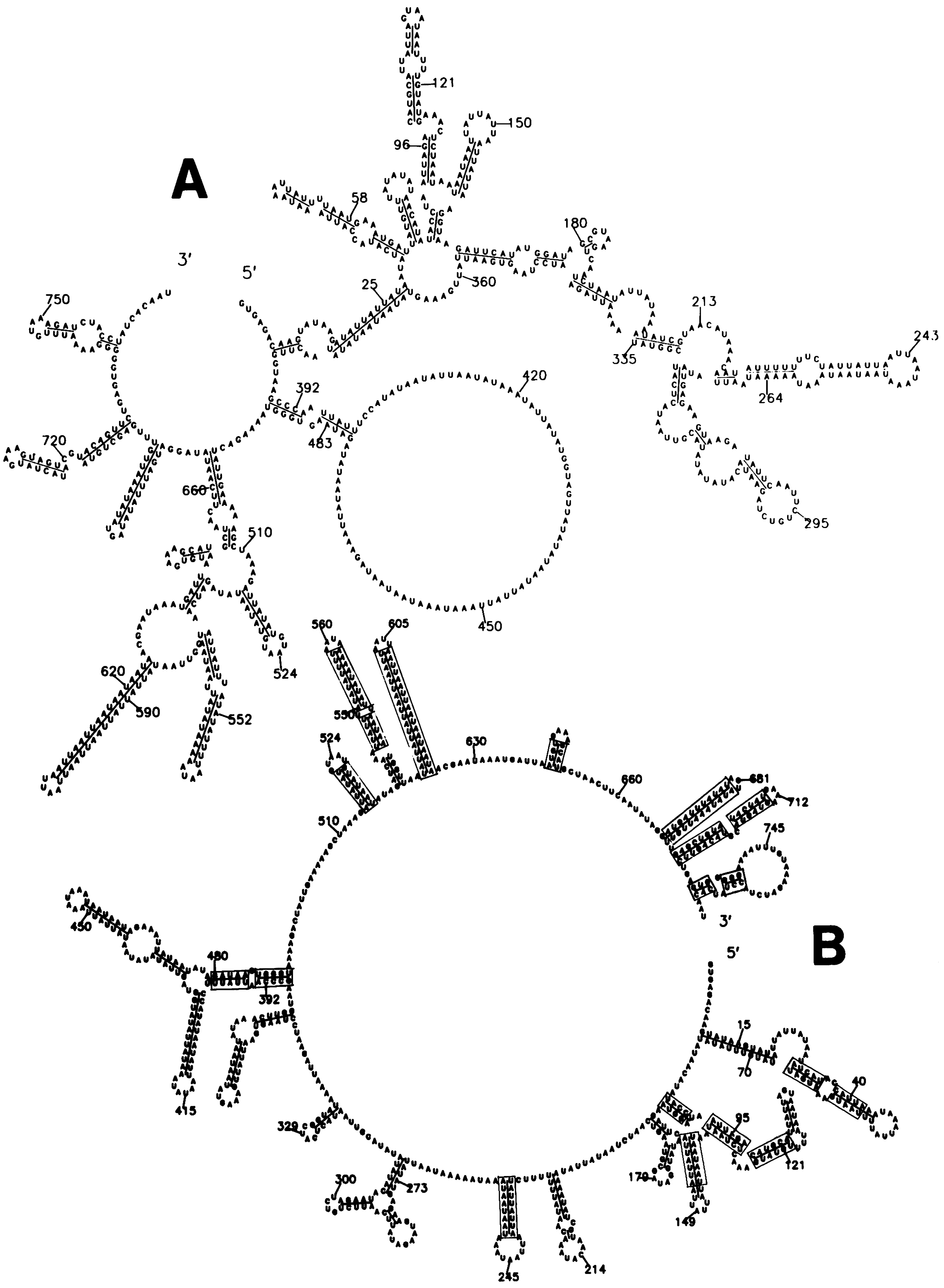
Figure 2. Phylogenetic (Top) and predicted optimal (Bottom) structures for domain 2 of C. reinhardii chloroplast 16S like rRNA (35). The best predicted suboptimal structure is essentially identical to the phylogenetic structure. Phylogenetic helixes in predicted structure are boxed.

At first, it is surprising that such a small free energy window is required to find the single best structure. The energy parameters in the folding model are known to 10% accuracy at best, since little is known about the sequence dependence of stabililty for single stranded regions (2,10,16,41). The nearest neighbor approximation for base pair stability is only accurate to about 5% (10,42). It must be remembered, however, that the free energy parameters are additive in the nearest neighbor model, whereas their random errors propagate as the square root of the sum of the squared errors, i.e. $\Delta G^\circ_T = \Delta G^\circ_1 + \Delta G^\circ_2 + \bullet\bullet\bullet + \Delta G^\circ_N$, whereas $\sigma_T = (\sigma_1{}^2 + \sigma_2{}^2 + \bullet\bullet\bullet + \sigma_N{}^2)^{1/2}$. Here σ_i is the error in ΔG_i°. For example, if $N = 100$ and $\Delta G^\circ_1 = \Delta G^\circ_2 = \bullet\bullet\bullet = \Delta G^\circ_{100} = \Delta G^\circ$, $\sigma_1 = \sigma_2 = \bullet\bullet\bullet = \sigma_{100} = 0.1\Delta G^\circ$, then $\Delta G^\circ_T = 100\ \Delta G^\circ \pm \Delta G^\circ$. Thus for this case, the error in each ΔG°_N is 10%, but because the errors are random, the error in ΔG°_T is 1%. The empirical results in Table I are thus consistent with the errors being largely random. This implies that for large sequences such as those shown in Table I, it should be sufficient to search suboptimal structures within a few percent of the free energy of the optimal structure.

The results shown in Table I seem less favorable than those presented previously (16), where a free energy window of 10% was sufficient to include a structure more than 90% homologous with the phylogenetic structures. This is because in the previous work, correct helixes were gathered from a group of suboptimal foldings. In this work, only single computer selected structures have been used in the comparisons with phylogenetically determined ones. This is perhaps more realistic in terms of how this program is likely to be used. A molecular biologist examining a single RNA without the benefit of a known structure or closely related RNAs will most likely choose a single structure from a list of computed ones based on closest agreement with chemical modification, enzyme cleavage, or other experimental data. The process of selecting compatible helixes from several suboptimal structures can then be used for refinement as more data becomes available.

Additional data should also restrict the number of structures that need to be considered. Table I lists in parentheses results for chloroplast 5S and E. coli 16S rRNA when chemical modification data are used as a constraint. For the chloroplast 5S RNA, 10 structures were generated in the absence of constraints within a free energy window of 10% using a distance of 2. The optimal and 'single best' structures had 4 of 5 known helixes. When the chemical modification data of Romby et al. (43) were added as constraints, only 1 structure was generated with the same parameters. This structure agreed with the phylogenetic structure except for the absence of base pairs G70•U109 and A30•U56, the latter because A30 is constrained to be single stranded from the modification data.

For 16S rRNA, each domain was folded with the constraint that nucleotides strongly modified in protein free 16S rRNA (4) were forced to be single stranded. Substantially fewer structures


A
3'
5'
121
150
96
58
180
25
360
213
243
750
335
264
392
420
483
720
660
510
524
295
450
560
605
620
590
552
B
550
524
630
660
681
712
745
510
3'
5'
450
490
392
15
70
40
415
329
95
179
121
149
300
273
245
214

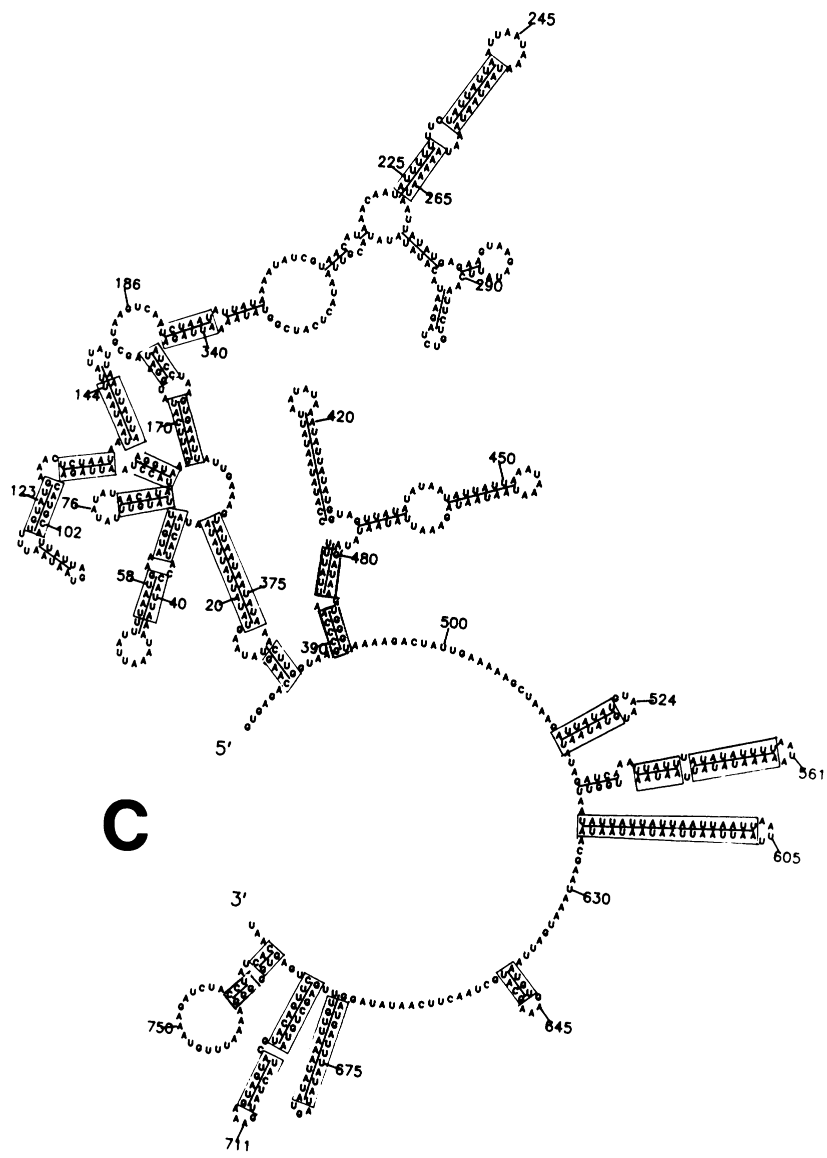


Figure 3. Phylogenetic (A), predicted optimal (B), and best predicted suboptimal (C) structures for yeast B1 intron (34). Phylogenetic helixes in predicted structures are boxed. The phylogenetic structure should contain additional pairing between nucleotides 733–736: 762–765.

(13 vs. 37) were generated only for domain 2. Domains 1 and 2 gave improved foldings, but domain 3 gave a substantially worse optimal folding (see Table I). This is partly due to the strong reactivity observed at nucleotide 951 which occurs in the middle of a 10 base pair helix in the phylogenetic structure (4). Nevertheless, a structure much like the phylogenetic structure and consistent with the chemical modification data is found with a free energy 3.6% less favorable than the lowest free energy structure. The 6 helixes missing in this structure are replaced by relataively similar helixes. In particular, the helix between nucleotides 923–933: 1384–1393 is replaced by 926–933: 1384–1391; nucleotides 946–955 pair with 978–980+1222–1224 instead of 1225–1235; 1058–1067 pair with 1187–1199 instead of 1189–1199; 1113–1116 pair with 1179–1182 instead of 1184–1187; 1128–1132 pair with 1139–1143 forming a hairpin loop of 6 instead of 1128–1135: 1139–1144 forming a hairpin of 3; the helix 1046- 1057: 1203–1211 is shortened to 1046–1053: 1205–1211 followed by an internal loop of 3.

To gain further insight into the suitability of the nearest neighbor model and the energy parameters used in the folding algorithm, the free energies of the phylogenetic and optimal structures for the entire 16S rRNA were computed. The values are −343.8 and −432.6 kcal/mol, respectively. This 20% difference could reflect tertiary or protein interactions that force the RNA to fold into a suboptimal secondary structure. Another possibility, however, is that the single stranded regions of the phylogenetic structure may form additional base pairs that are not conserved. To determine the possible magnitude of this effect, the phylogenetic structure was energy minimized as described in Materials and Methods. Forty-seven additional base pairs formed, and are listed in Table II. Only 5 of the 94 nucleotides involved are strongly hit by chemical modification reagents (4), suggesting many of these base pairs may occur. The free energy of the phylogenetic structure with these additional base pairs is −384.4 kcal/mol, only 11% higher than for the predicted optimal structure. Thus, there is no indication that tertiary or protein interactions are strong enough to grossly alter the folding determined by secondary structural interactions.

Consideration of the free energy window required to find at least one structure for each of the phylogenetically determined base pairs in 16S rRNA provides an indication that energy minimization can also eliminate unreasonable base pairs while retaining feasible base pairs. In particular, the worst base pair, U1056-A1204, is first found in a structure that is 2.7% higher in free energy than the optimal structure. This compares with an 8.6% window required to find each of the 444,768 possible base pairs that the sequence could form. While the 2.7% energy window contains structures with 92,217 different base pairs, it nevertheless excludes 79% of the possible base pairs.

The fact that all base pairs in the phylogenetic model can be found in at least one structure within 2.7% of the lowest free energy structure should not be confused with the fact that the energy of the extended phylogenetic structure is 11% from optimal. If multiple structures were automatically generated at the 2.7% level with a window value of 0, the results might contain thousands of structures. Altogether, these structures would contain all the phylogenetically proven base pairs, but they would be scattered among many different structures. Combining all the correct base pairs into a single structure forces the energy to rise. This is why energy minimization is more successful at limiting the number of structures that need be considered rather than the number of base pairs. The results indicate that for large sequences at least, inspection of individual base pairs as in energy dot plots (9) should be confined to just a few percent from optimal.

The large number of RNA sequences being determined leads to a demand for rapid methods to determine RNA structure. The results in this paper show that while a vast number of structures are possible, energy minimization can filter out most of them and provide a reasonable set of working models. Selection among these models can then be made with phylogenetic comparisons, chemical modification data, site directed mutagenesis, and other methods.

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health Grant GM22939. Michael Zuker is a Fellow of the Canadian Institute for Advanced Research. We thank Philip C.Bevilacqua for suggesting the error analysis.

REFERENCES

1. Tinoco, I. Jr., Uhlenbeck, O.C., and Levine, M.D. (1971) *Nature (London)* **230**, 363–367.
2. Turner, D.H., Sugimoto, N., and Freier, S.M. (1988) *Annu. Rev. Biophys. Biophys. Chem.* **17**, 167–192.
3. Heus, H.A., Uhlenbeck, O.C., and Pardi, A. (1990) *Nucleic Acids Res.* **18**, 1103–1108.
4. Moazed, D., Stearn, S. and Noller, H.F. (1986) *J. Mol. Biol.* **187**, 399–416.
5. Kim, S.-H., Quigley, G.J., Suddath, F.L., McPherson, A., Sneden,D., Kim, J.J., Weinzierl, J., and Rich, A. (1973) *Science* **179**, 285–288.
6. Robertus, J.D., Ladner, J.E., Finch, J.T., Rhodes, D., Brown, R.S. et al. (1974) *Nature (London)* **250**, 546–551.
7. Moras, D., Comarmond, M.B., Fischer, J., Weiss, R., Thierry, J.C., Ebel, J.P., and Giege, R. (1980) *Nature (London)* **288**, 669–674.
8. Zuker, M. and Sankoff, D. (1984) *Bull. Math. Bio.* **46**, 591–621.
9. Zuker, M. (1989) *Science* **244**, 48–52.
10. Freier, S.M., Kierzek, R., Jaeger, J.A., Sugimoto, N., Caruthers, M.H., Neilson, T., and Turner, D.H. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 9373–9377.
11. Zuker, M. (1986) *Lect. Math. Life. Sci.* **17**, 87–124.
12. Williams, A. Jr., and Tinoco, I. Jr. (1986) *Nucleic Acids Res.* **14**, 299–315.
13. Nussinov, R., and Jacobson, A.B. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 6309–6313.
14. Zuker, M., and Steigler, P. (1981) *Nucleic Acids Res.* **9**, 133–148.
15. Waterman, M.S., and Beyers, T.H. (1985) *Math. Biosci.* **77**, 179–184.
16. Jaeger, J.A., Turner, D.H., and Zuker, M. (1989) *Proc. Natl. Acad. Sci. U.S.A.* **86**, 7706–7710.
17. Ehresmann, C., Baudiǹ, F., Mougel, M., Romby, P., Ebel, J.P., and Ehresmann, B. (1987) Nucleic Acids Res. **15**, 9109–9128.
18. Burke, J.M., Irvine, K.D., Kaneko, K.J.,Kerker, B.J., Oettgen, A.B., Tierney, W.M., Williamson, C.L., Zaug, A.J., and Cech, T.R. (1986) *Cell (Cambridge, Mass.)* **45**, 167–176.
19. Michel, F., Ellington, A.D., Couture, S., and Szostak, J.W. (1990) *Nature (London)* **347**, 578–580.
20. Noller, H.F., and Woese, C.R. (1981) *Science* **212**, 403–411.
21. James, B.D., Olsen, G.J., and Pace, N.R. (1989) *Methods Enzymol.* **180**, 227–239.
22. Jaeger, J.A., Zuker, M., and Turner, D.H. (1990) *Biochemistry* **29**, 10147–10158.
23. Freier, S.M., Burger, B.J., Alkema, D., Neilson, T., and Turner, D.H. (1983) *Biochemistry* **22**, 6198–6206.
24. Turner, D.H., Sugimoto, N., Kierzek, R., and Dreiker, S. (1987) *J. Am. Chem. Soc.* **104**, 3783–3785.
25. Sugimoto, N., Kierzek, R., Freier, S.M., and Turner, D.H. (1986) *Biochemistry* **25**, 5755–5759.
26. Tuerck, C., Gauss, P., Thermes, C., Grobe, D.R., Gayle, M., Guild, N., Stormo, G., d'Aubenton-Carafa, Y., Uhlenbeck, O.C., Tinoco, I. Jr., Brody, E.N., and Gold, L. (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85**, 1364–1368.
27. Woese, C.R., Winker, S., and Gutell, R.R. (1990) *Proc. Natl. Acad. Sci. U.S.A.* **87**, 8467–8471.
28. Jacobson, H., and Stockmayer, W.H. (1950) *J. Chem. Phys.* **18**, 1600–1606.
29. Jaeger, J.A., Zuker, M., and Turner, D.H. (1990) *Methods Enzymol.* **183**, 281–306.
30. Zuker, M. (1989) in *'Mathematical Methods for DNA Sequences' (M.S. Waterman, ed.)* 159–184, CRC Press.
31. Prokhorov, Y.V. (1956) *Theory Prob. Appl.* **1**, 157–214.
32. Pleij, C.W.A., van Belkum, A., Rietreld, K., and Bosch, L. (1986) *Structure and Dynamics of RNA (van Knippenberg, P.H. and Hilbers, C.W., eds)*, 87–98 (Plenum).
33. Sampson, J.R. and Uhlenbeck, O.C. (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85**, 1033–1037.
34. Michel, F., and Jacquier, A. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52**, 201–212.
35. Gutell, R.R., Weiser, B., Woese, C.R., and Noller, H.F. (1985) *Prog. Nucleic Acid Res. Mol. Biol.* **32**, 155–216.
36. Cheong, C., Varani, G., and Tinoco, I. Jr. (1990) *Nature (London)* 680–682.
37. Varani, G., Cheong, C., and Tinoco, I. Jr. (1991) *Biochemistry*, in press.
38. Grobe, D.R., (1988) *Ph.D. Thesis, University of Colorado, Boulder.*
39. SantaLucia, J., Kierzek, R., and Turner, D.H. (1990) *Biochemistry* **29**, 8813–8819.
40. SantaLucia, J., Kierzek, R., and Turner, D.H., (1991) *J. Amer. Chem. Soc.*, in press.
41. Grobe, D.R., and Uhlenbeck, O.C. (1988) *Nucleic Acids Res.* **16**, 11725–11735.
42. Kierzek, R., Caruthers, M.H., Longfellow, C.E., Swinton, D., Turner, D.H. and Freier, S.M. (1986) *Biochemistry* **25**, 7840–7846.
43. Romby, P., Westhof, E., Toukifimpa, R., Mache, R., Ebel, J., Ehresmann, C., and Ehresmann, B. (1988) *Biochemistry* **27**, 4721–4730.
44. Shub, D.A., Gott, J.M., Xu, M.-Q., Lang, B.F., Michel, F., Tomaschewski, J., and Belfort, M. (1988) *Proc. Natl. Acad. Sci. U.S.A.* **85**, 1151–1155.