The evolution of 5S RNA secondary structures

DAVID SANKOFF AND ANNE-MARIE MORIN

Centre de recherches mathématiques, Université de Montréal, Montréal (Qué.), Canada H3C 3J7

AND

R. J. CEDERGREN

Département de biochimie, Université de Montréal, C.P. 6128, Montréal (Qué.), Canada H3C 3J7

Received February 7, 1978

This paper is dedicated to the memory of the late Dr. G. Malcolm Brown

Sankoff, D., Morin, A.-M. & Cedergren, R. J. (1978) The evolution of 5S RNA secondary structures. *Can. J. Biochem. 56*, 440–443

We have applied the Pipas–McMahon algorithm based on free energy calculations to the search for a 5S RNA base-pair structure common to all known sequences. We find that a 'Y'-shaped model is consistently among the structures having the lowest free energy using 5S RNA sequences from either eukaryotic or prokaryotic sources. Comparison of this 'Y' structure with models which have recently been proposed show these models to be remarkably similar, and the minor differences are explicable based on the technique used to obtain the model. That prokaryotic and eukaryotic 5S RNA can adopt a similar secondary structure is strong support for its resistance to change during evolution.

Introduction

Although the primary structure of an RNA can be directly determined by laboratory techniques, its secondary structure, i.e., its base-pairing arrangement, is less accessible; relevant experiments provide only partial and indirect evidence. For example, the secondary 'cloverleaf' model of tRNA was confirmed mainly by showing that all known tRNA sequences were compatible with this conformation (1). More recently, X-ray crystallography data from tRNAs were shown to be consistent with the cloverleaf model (2, 3). Finally, and of particular interest here, Pipas and McMahon (4) carried out free energy calculations of all possible conformations of 62 tRNA sequences and found that the cloverleaf is thermodynamically preferred in most of the tRNAs examined.

Although many 5S RNA sequences are known, a generalized secondary structure has defied consensus. Quite similar models have, however, been proposed by a number of investigators (5–8). We report here the results of combining free energy calculations, using the Pipas and McMahon programme which was previously described, together with comparative considerations in inferring the base-pairing pattern of 5S RNA and its evolution across various phyla.

Definitions, Problem, and Methods

As indicated above, we take 'secondary structure' to be coterminous with base-pairing pattern, though the restrictions we impose on possible patterns preclude certain rare but feasible combinations of pairs. For example, if B_1, B_2, B_3, and B_4 are bases occurring in that order, not consecutively, but spaced out along a sequence, we exclude the possibility that B_1 and B_3 be paired at the same time as B_2 and B_4, any other combination being permitted. This restriction against 'knots' is reasonable since few such crossed-over pairs have been documented for RNA structures, and they are rather considered aspects of tertiary structure (2, 3). In addition, this restriction is the key to methods of finding optimal secondary structures, since it greatly limits the amount of structure searching which must be carried out. Of course, we impose additional restrictions on possible secondary structures, such that all pairs must be of the Watson–Crick or G-U type, that adjacent bases, or bases separated by only one or two intervening bases, are too close to be paired, etc.

We turn to the question of how to ascertain the correct secondary structure among all the possible base-pairing patterns. One approach is to choose the structure of lowest total free energy. This leads to the following three problems: (*a*) First, the free energy which we calculate for an isolated molecule, i.e., in solution, is not necessarily meaningful to *in vivo* conditions where certain conformations could be stabilized by association with ribosomal protein and by tertiary structure considerations. We might hope, nevertheless, that the structure of the molecule in solution has some bearing on its structure *in vivo* and to see whether a consistent pattern appears from species to species, suggesting evolutionary conservation of functional–structural aspects. (*b*) The 5S RNA molecule is much too large for exact free energy calculations. Here, we can try local free energy calculations as pioneered by Tinoco (9) and Tinoco *et al.* (10). A proposed secondary structure can be viewed in a unique way as being made up of various helical (base-paired) regions plus single-stranded regions. The free energy due to each of these portions of the molecule can be estimated based on experimental results on melting curves of various oligonucleotides, and these local energies can be added together in order to approximate the global free energy of the structure. (*c*) The task of carrying out the Tinoco calculation for each of the 10 or 100 000 possible structures which are compatible with any given sequence in order to find the thermodynamically preferred one seems prohibitive. But this can be solved by an electronic computer, a well thought-out algorithm, and careful programming.

Pipas and McMahon (4) devised and implemented the following procedure. First, all possible helical regions in the sequence which result in at least three consecutive Watson–Crick or G-U pairs when the molecule is folded back upon itself are compiled. There is an important restriction, however, that no such helix begin or terminate with a G-U pair. The number of possible helices in the tRNA study was aroud 20 for the tRNA sequences

they inventoried. The extension from the 80 nucleotides in tRNA to the 120 in 5S RNA increases the number of possible helices to about 125.

The second step is to decide whether two helices are compatible for a given sequence. They are obviously not if they overlap and involve identical terms of the sequence in different pairing arrangements. More generally, two helices are incompatible if they produce a 'knot' as defined above.

The third and most time-consuming part of the algorithm is to examine in effect all sets of mutually compatible helices to see whether the corresponding structure has a low free energy as calculated by the Tinoco approach. Roughly speaking, the number of different sets of mutually compatible helices is an exponential function of the number of helices themselves which explains why the task of examining all these sets is orders of magnitude more difficult for the 5S RNA molecule than for tRNA.

The 5S RNA sequences from *Escherichia coli, Bacillus megaterium, Bacillus stearothermophilus, Anacystis nidulans, Chlorella*, KB cells, *Xenopus*, and *Torulus utilus* which were used in this study can be found in Ref. 11. In addition, we have analyzed the 5S RNA sequences of rye (12), bean (12), *Drosphila* (13), *Saccharomyes carlbergensis* (14), *S. cerevisiae* (15), and chicken (16).

Results

We have used an option of the Pipas–McMahon programme to print out the 'best' 10, 20, or 50 structures having the lowest free energy, since free energy differences may be insignificant among the best structures. Examination of these best structures calculated from RNA sequences of different organisms showed that one and only one structural type was consistently present in the top 5 or 10 best structures. This structure is a Y-shaped model which is very similar to the models already cited and in fact not so different from the model suggested by Madison in 1968 based on only two sequences (17). For some 5S RNA sequences such as those from *Drosophila*, rye, *S. cerevisiae*, and *T. utilus*, the Y structure had the lowest overall free energy. The structure for rye which is also representative of the others is shown in Fig. 1. The free energy of these Y models is of the order of -40 ± 10 kcal/mol. For this same group of organisms, many of the other 'best' structures were closely related variants of the Y model.

For another set of 5S RNA sequences including those of *Anacystis, B. stearothermophilus, Chlorella*, and broad bean, the Y model, although not of lowest energy, was found among the best structures in the computer output. Typical of this set was *Anacystis* 5S RNA whose lowest energy form was a rather symmetrical cloverleaf (Fig. 2*a*). This model was not found for any of the other 5S RNA sequences. The fifth best structure, however, was the familiar Y model (Fig. 2*b*) whose free energy was only 5% higher than the lowest energy form. Similarily, the lowest free energy structure of *B. stearothermophilus* was the H form shown in Fig. 3*a*. Again, this H form was not found among the 'best' structures for any other 5S RNA sequence. The Y model shown in Fig. 3*b* was found to have about 5% higher free energy than the H structure.

There are a small number of 5S RNA sequences that behave anomalously and do not produce Y-shaped models; they are the *E. coli, S. carlbergensis*, and the three closely related animal sequences.[1] This anomalous behaviour can be traced to certain restrictions in the programme such as not allowing single unpaired bases in the middle of a short helix. These constraints are not completely justifiable thermodynamically and in rare cases they actually preclude examination of valid low energy conformations (without these generally reasonable re-

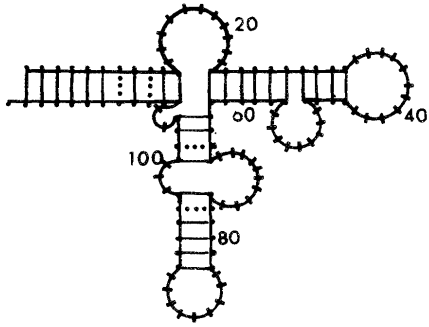

FIG. 1. Base-pair model of rye 5S RNA which has the lowest free energy. Each hatch mark represents a base. Base pairs (A-U or C-G) are indicated by continuous lines where G-U pairs are dotted.

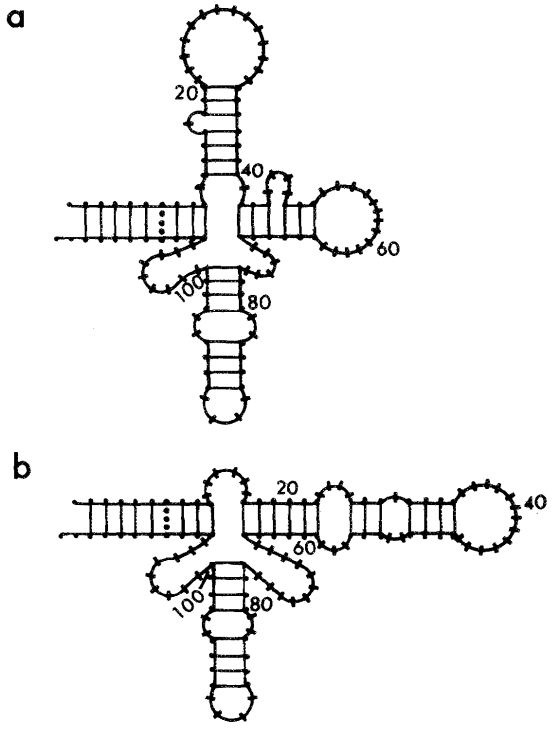

FIG. 2. (*a*) Base-pair model of *A. nidulans* 5S RNA having the lowest free energy; (*b*) a Y model outputed among the 'best' structures.

[1]The models obtained for these sequences are unique in that they are not found for other sequences.

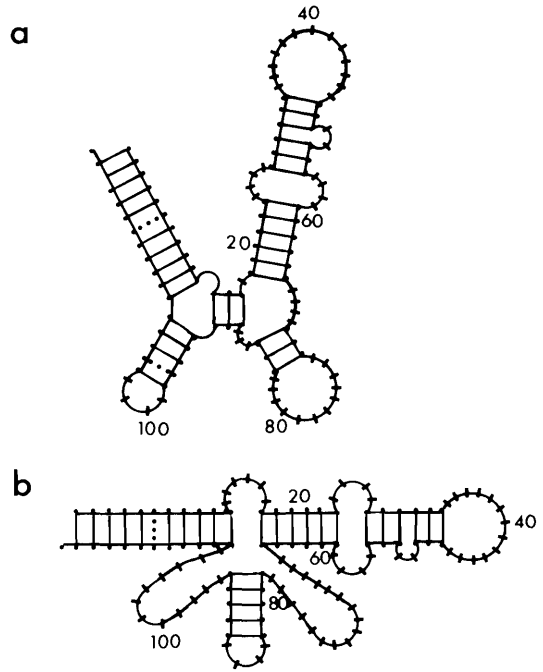

FIG. 3. (*a*) Base-pair model of *B. stearothermophilus* 5S RNA having the lowest free energy; (*b*) a Y model outputed among the 'best' structures.

strictions, however, the number of structures to be examined would greatly exceed the capabilities of even a high-speed computer). That this result is an artefact of the programme was proved by manual calculation of the Y free energy of the model for each of the anomalous RNAs. This calculation shows that the Y-shaped model actually has either a lower or nearly equivalent free energy value (depending on the sequence) than the lowest energy structure produced by the programme.

Discussion

Pipas and McMahon (4) found their programme to accurately predict the correct cloverleaf structure in about half of the tRNA sequences examined. In another 25% of the sequences, a close variation of the correct structure was found, and in the remaining cases, other types of structures were produced. In the present exercise on the considerably longer 5S RNA molecule, the results are comparable. About half of the sequences produced Y shapes with the lowest free energy structures. As Pipas and McMahon (4) observed, in some cases, the correct structure was not found because of assumptions inherent in the programme which are necessary to keep computing time within reasonable limits but which occasionally exclude valid structures. In the 5S RNA case, this occurred with *E. coli*, *S. carlbergensis*, and animal sequences, though all of these can be shown to have very low energy Y-shaped structures. In the remaining cases, as was also true with respect to the cloverleaf in the tRNA study, the Y shape, though not optimal, was within 10% of being optimal. This can be understood in terms of the approximate nature of free energy calculations and the empirical results on which they are based. It should be stressed that no type of structure other than the Y shape of rather stable dimension appears as a low energy solution in diverse species. It seems clear then that if 5S RNA secondary structure is to be at all comparable in different branches of evolutionary history, it must take on the Y shape.

Towards a Consensus on 5S RNA Secondary Structure

A Y shape was one of the first 5S RNA structural suggestions (17), and many recent proposals are variants of this shape. Most important among these are the model worked out largely for eukaryotes by Vigne and Jordan (19), the similar model for prokaryotes studied by Fox and Woese (11), and the somewhat different structure suggested by Nishikawa and Takemura (7). We consider all these, as well as the various free energy solutions discussed in the present paper, basically to be variants of a common structure. The Fox–Woese and Vigne–Jordan models represent a more conservative position towards the existence of paired bases, preferring to leave single-stranded those regions in which there is no comparative evidence for helical structure. The Takemura and Nishikawa model represents the other extreme, allowing as much base pairing as possible even if this is quite species specific. Our own calculations tend to fall somewhere between the two largely because of constraints against short helices and G-U pairs built into the programme. All these models find the same two hairpins closed by identical helices, though the existence and composition of helical regions near the centre of the molecules differs widely. This is illustrated in Fig. 4 with the three models for *A. nidulans* 5S RNA.

We feel that the biological significance of these model differences is doubtful and therefore consider them all to be roughly equivalent and refer to them collectively as the Y model. This model is supported by many lines of independent evidence. The number of base pairs predicted by the model is roughly equivalent to that calculated from nuclear magnetic resonance data (18). In addition, Vigne and Jordan (19) found that limited ribonuclease digests of 5S RNA isolated from six different species, eukaryotic and prokaryotic, cleave the RNA only in the single-stranded regions predicted from this model. It has equally been shown that a similar model is consistent with X-ray scattering experiments (20). Finally, our free energy calculations show that Y structures are consistently among the lowest energy forms possible. Results of complementary oligonucleotide binding to 5S RNA can also be seen to be consistent with the Y model (Erdmann, V. A., personal communication).

A 5S RNA model similar in all species is in good agreement with the large sequential homology of 5S RNA from different sources and emphasizes the evolutionarily conservative nature of prokaryotic and eukaryotic 5S RNA. We do remark one consistent evolutionary trend which distinguuishes prokaryotic and eukaryotic 5S RNA and which may be of practical significance. While

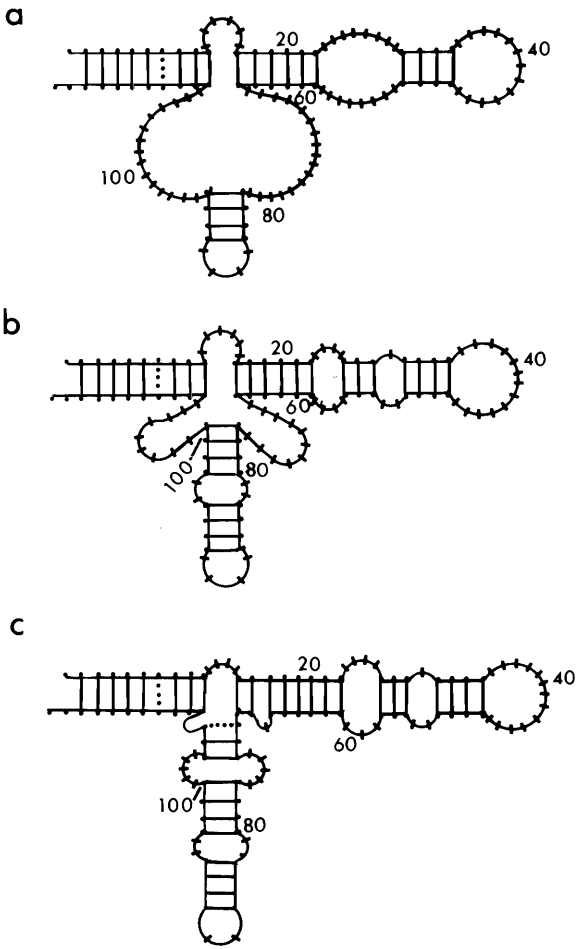

FIG. 4. (*a*) Fox and Woese model of *A. nidulans* 5S RNA; (*b*) free energy Y model as produced in this paper of *A. nidulans* 5S RNA; (*c*) Nishikawa and Takemura type model of *A. nidulans* 5S RNA.

the long axis of the molecule remains relatively unchanged, the so-called "prokaryote loop" (6) consists, in the prokaryotes, of a helix of four or five G-C pairs closing a hairpin loop of three or four pyrimidines attached to the main body of the molecule by two long single-stranded regions. These regions may be involved in some base pairing but not in a way which is similar from one bacteria to another. The eukaryotes, on the other hand, have some A-U or even G-U pairs closing the analogous loop which is generally longer than for prokaryotes. In addition, the eukaryotes all permit a good deal of base pairing in the portion of the 'prokaryote loop' which is proximal to the main axis, whereas this region is single stranded in prokaryotes.

Acknowledgements

The authors thank Dr. B. Jordan for providing results prior to publication and for many stimulating discussions. In addition, we thank Dr. V. Erdmann and T. Dyer for supplying data prior to publication and Dr. J. Pipas for providing their programme. This work was supported by a grant from le Ministère de l'Education du Québec.

1. Cramer, F. (1971) *Prog. Nucleic Acid Res. Mol. Biol. 11*, 391–421
2. Robertus, J. D., Ladner, J., Finch, J., Rhodes, D., Brown, R., Clark, B. & Klug, A. (1974) *Nature (London) 250*, 546–551
3. Kim, S., Suddath, F., Quigley, G., McPherson, A., Sussman, J., Wang, A., Seeman, N. & Rich, A. (1974) *Science 185*, 435–440
4. Pipas, J. M. & McMahon, J. E. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 2017–2021
5. Jordan, B. R., Galling, G. & Jourdan, R. (1974) *J. Mol. Biol. 87*, 755–774
6. Fox, G. E. & Woese, C. R. (1975) *Nature (London) 256*, 505–507
7. Nishikawa, K. & Takemura, S. (1974) *J. Biochem. 76*, 935–947
8. Hori, H. (1976) *Mol. Gen. Genet. 145*, 119–123
9. Tinoco, I., Uhlenbeck, O. C. & Levine, M. D. (1971) *Nature (London) 230*, 362–367
10. Tinoco, I., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M. & Gralla, J. (1973) *Nature (London) New Biol. 246*, 40–41
11. Fox, G. E. & Woese, C. R. (1975) *J. Mol. Evol. 6*, 61–76
12. Dyer, T. A., Bowman, C. M. & Payne, P. I. (1976) in *Nucleic Acids and Protein Synthesis in Plants* (Bogorad, L. & Weil, J. H., eds.), pp. 121–133, Plenum Press, New York
13. Benhamou, J. & Jordan, B. R. (1976) *FEBS Lett. 62*, 146–149
14. Hindley, J. & Page, S. M. (1972) *FEBS Lett. 26*, 157–160
15. Miyazaki, M. (1974) *J. Biochem. 75*, 1407
16. Brownlee, G. G. & Cartwright, E. M. (1975) *Nucleic Acid Res. 2*, 2279–2288
17. Madison, J. T. (1968) Annu. Rev. Biochem. *37*, 131–148
18. Kearns, D. R. & Wong, Y. P. (1974) *J. Mol. Biol. 87*, 755–774
19. Vigne, R. & Jordan, B. R. (1977) *J. Mol. Evol.*, in press
20. Österberg, R., Sjöberg, B. & Garrett, R. A. (1976) *Eur. J. Biochem. 68*, 481–487