Volume 10 Number 1 1982 Nucleic Acids Research

Computer-aided prediction of RNA secondary structures

Philip E.Auron*, Wayne P.Rindone**+, Calvin P.H.Vary†, James J.Celentano† and John N.Vournakis†

*Massachusetts Institute of Technology, Department of Biology, Cambridge, MA 02139, **Bolt Beranek and Newman Inc., 10 Moulton St., Cambridge, MA 02238, and †Syracuse University, Department of Biology, Syracuse, NY 13210, USA

Received 29 September 1981

ABSTRACT

A brief survey of computer algorithms that have been developed to generate predictions of the secondary structures of RNA molecules is presented. Two particular methods are described in some detail. The first utilizes a thermodynamic energy minimization algorithm that takes into account the likelihood that short-range folding tends to be favored over long-range interactions. The second utilizes an interactive computer graphic modelling algorithm that enables the user to consider thermodynamic criteria as well as structural data obtained by nuclease susceptibility, chemical reactivity and phylogenetic studies. Examples of structures for prokaryotic 16S and 23S ribosomal RNAs, several eukaryotic 5S ribosomal RNAs and rabbit β-globin messenger RNA are presented as case studies in order to describe the two techniques. An argument is made for integrating the two approaches presented in this paper, enabling the user to generate proposed structures using thermodynamic criteria, allowing interactive refinement of these structures through the application of experimentally derived data.

INTRODUCTION

Current methods for computer-aided prediction of RNA secondary structure (1-3) are designed to accept primary sequence information and generate the minimum free energy structure by using a general approach pioneered in the laboratory of I. Tinoco, Jr. (4). This empirical approach uses base stacking, hydrogen bonding, and loop free energy data obtained in many laboratories during the past 15 years from studies of physical-chemical properties of RNA oligomers and homopolymers (5). The three methods mentioned in references 1 to 3 vary in either the free energy values used or in the programming strategy.

Early attempts at predicting the most stable RNA secondary structure from sequence data alone were initiated by Pipas and McMahon (1). An algorithm was developed which carries out the analysis of possible secondary structures for any RNA primary structure in three sequential steps. This algorithm generates a list of all possible helical regions and then uses this list to calculate

© IRL Press Limited, 1 Falconberg Court, London W1V 5FG, U.K.

all possible compatible structures (i.e., those containing only non-overlapping helices). Finally, every structure obtained is evaluated for its total free energy of formation using a completely extended chain polymer as a reference state, and the minimum free energy structure is selected.

Studnicka et al. (2) have developed a similar method for predicting the most energetically favorable secondary structure for an RNA molecule using free energy values slightly different from those of Pipas and McMahon. This algorithm lists all possible double helical regions and then examines every pair of mutually incompatible regions to determine whether parts of those regions can be combined by branch migration to form a pair of compatible new subregions which together are more stable than either of the original regions separately. An additional improvement by Studnicka et al. is the formulation of a hyperstructure matrix which contains the topological relationship of all pairs of regions. The most stable structure can be derived directly from this matrix, thereby reducing the time spent examining all possible structures for the most stable structure, the most time consuming part of the problem.

The recent development of approaches that evaluate solution structure of RNA molecules using structure specific chemical and enzymatic probes has demonstrated that neither of these two landmark algorithms can be relied on to generate RNA secondary structures that are consistent with the data elicited by use of such probes. The reasons for this are not clear, but may have to do with assumptions made in assigning the free energy values to a particular secondary structure, the lack of sufficient experimentally derived free energy parameters, and the number of alternative secondary structures whose free energies are very nearly equal to the minimal free energy obtained for the most stable RNA structure. Some discrepancies are no doubt due not to inaccuracies in the prediction algorithms, but rather to an incomplete understanding of the behavior of the probes.

A THERMODYNAMIC FOLDING ALGORITHM

Despite the shortcomings of empirical thermodynamic energy minimization algorithms, predictive methods based upon molecular theories and surface exposure values (6) are probably premature and will have to await a better understanding of the forces involved. It is, however, possible to approach folding in a manner which is not purely thermodynamic. A kinetic parameter is probably essential for describing folding in that it is likely, in view of the stability of base-paired duplexes (7), that during the course of folding the molecule can become trapped within some localized energy minimum from which it cannot readily escape. This could prevent the molecule from continuing on to

a state which would represent the global minimum. Such a localized minimum would take the form of a large number of relatively small hairpins folding as a result of short-range interactions, and would not involve many extensive stretches of long range associations, except those which could occur at the expense of single stranded or easily dissociated double stranded regions. This kinetic argument is based on well understood collision theory and the presumed stability of reasonably well formed duplexes, which has been measured, to a first approximation, by determining proton exchange rates in nucleic acids (7). This type of folding scheme has also been suggested in the past on the basis of purely thermodynamic considerations (8).

Recently a secondary structure model has been presented for E. coli 16S ribosomal RNA based on chemical, enzymatic, and phylogenetic data (9,10,11). A remarkable feature of this structure is that 85 percent of the base pair interactions involve short range associations (i.e., hairpin loops closed by no more than 100 nucleotides). Such a pattern is consistent with the short-range folding model described above. Additionally, it is reasonable to expect that such folding be seen in RNA, since the molecules are synthesized in a 5' to 3' direction by RNA polymerase in vivo, resulting in a restricted folding pattern which would lead to proximal association of potential pairing regions, as in the short-range model. This idea of 5' to 3' gradient folding involving nascent RNA chains is supported by evidence accumulated by Spillman et al. (12) on E. coli 23S rRNA in which the protein assembly gradient was shown to parallel RNA synthesis. Since the 5' to 3' gradient folding scheme is itself a type of short-range folding, it is reasonable to expect that a short range interaction approach to RNA folding might better describe in vivo folding.

A program has been described (3) which utilizes the Tinoco rules to yield a single minimum energy structure, incorporating significant algorithmic improvements that allow the analysis of sequences that are longer than those that could be readily handled by earlier approaches. This program will be referred to below by the name RNA1. Recently a modification has been made to this program (M. Zuker, personal communication) that permits a parameter to be set which limits the number of nucleotides permitted to close any hairpin and can thereby be used to examine the short range type of folding scheme suggested by the combined kinetic-thermodynamic approach described above. This FORTRAN program will be referred to as RNA2. RNA2 was adapted to run on a VAX 11/780 and then used in conjunction with the original long range program and experimental data to obtain a structure for the 16S molecule which is

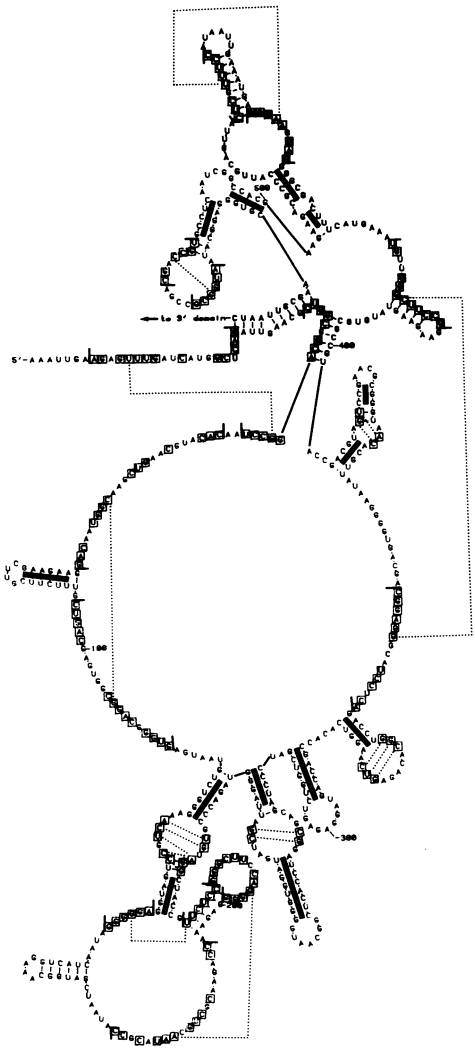

Figure 1, part A. 5' domain of secondary structure for E. coli 16S ribosomal RNA as proposed by Noller and Woese (11). Those base pairs in the Noller-Woese structure that were also found by the RNA2 program are indicated by a thick bar perpendicular to the hydrogen bonds between the bases. Residues that are involved in base pairs suggested by RNA2 that differ from the base pairs suggested by Noller and Woese are boxed. Complementary regions containing the base pairs found by RNA2 are joined by dotted lines.

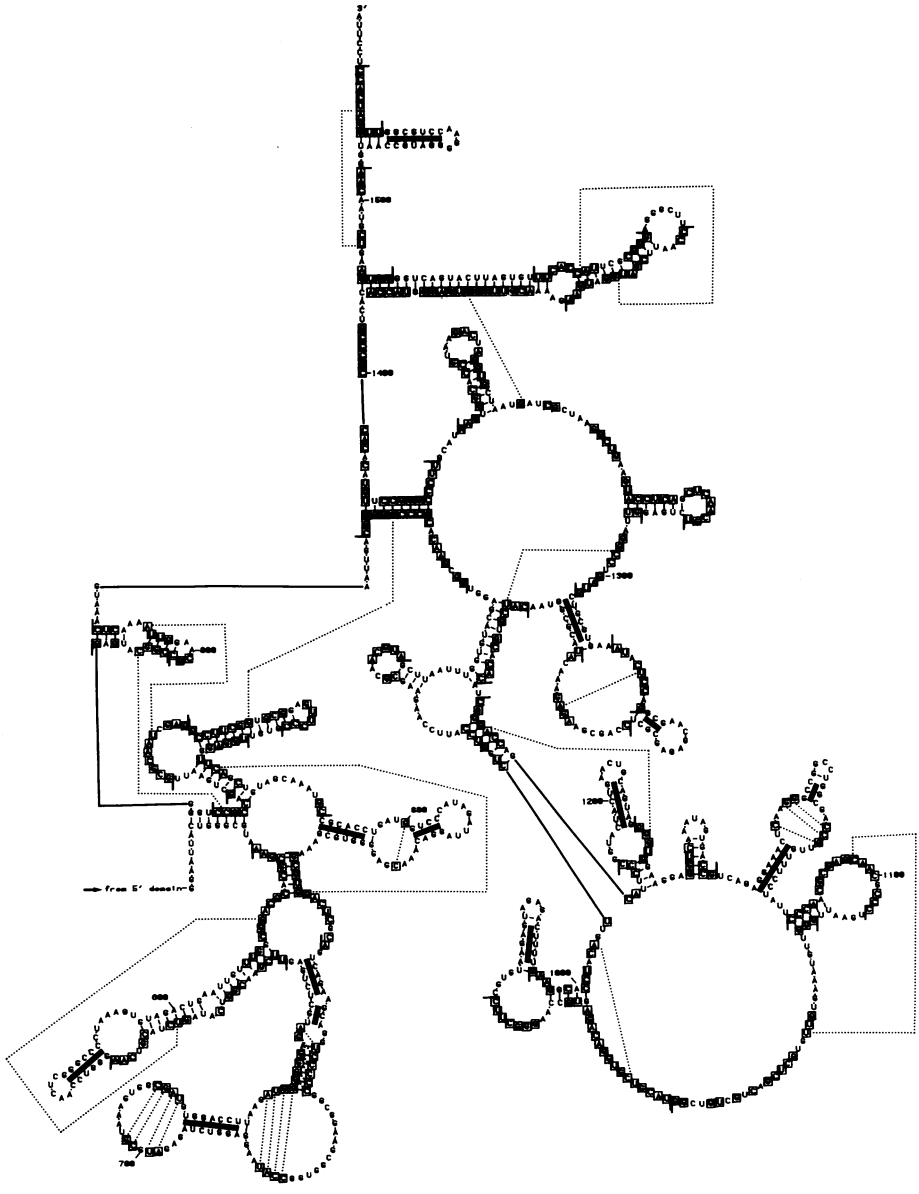

Figure 1, part B. 3' domain of secondary structure for E. coli 16S ribosomal RNA as proposed by Noller and Woese (11). Special symbols are described in the caption for Figure 1, part A.

quite similar to that described by Noller and Woese (10).

Figure 1, which was prepared by display software on the PROPHET computer system, described below, shows the Noller-Woese 16S ribosomal RNA structure. Superimposed upon it are the results obtained from computer folding using RNA2. The structure was folded in three overlapping pieces (residues 1-600, 470-1069, and 942-1542). What is immediately obvious in figure 1 is the large number of hairpin structures predicted by RNA2 which agree with the experimentally derived structure. Almost half of the total hairpin base pairs in the Noller-Woese structure are predicted out of the millions of possible base pairs that could be predicted. What is especially impressive is that these predictions were made using only the primary sequence and thermodynamic rules. In addition many of the base pairs predicted by the program, but not indicated in the Noller-Woese structure, agree very well with the experimental data presented by these authors (10). Studies in which the program was provided with smaller overlapping fragments in order to mimic 5' to 3' gradient folding yielded a few additional hairpins which agreed with the Noller-Woese structure. These occurred most notably in the vicinity of residues 840 and 1420, where long helices possessing phylogenetic support were found by the RNA2 program (results not shown). Analysis by the RNA1 program generated a few of the long-range helical regions in the Noller-Woese structure that RNA2 did not find. Of course the non-standard A-G base pairs suggested in the Noller-Woese model were not found by either RNA1 or RNA2 since these are not explicitly assigned free energy values within these programs.

On the basis of this study it appears that this thermodynamic short-range folding program can determine an adequate first approximation to an RNA secondary structure. One of several additional studies we have carried out to help confirm this conclusion involved folding rabbit β-globin mRNA using both RNA1 and RNA2. Figure 2 shows the RNA2 folding pattern of β-globin mRNA, as displayed in the program's output, with enzymatic digestion data superimposed as indicated in the caption to the figure. The folding scheme fits well with the data for both single and double strand specific nucleases. The fit with the single strand nuclease data was so good that rerunning RNA2 with the single strand data incorporated into the folding program (see 3) did not significantly alter the folding pattern. Folding of this molecule with RNA1 did not yield a good fit to the data, however, further supporting the short-range interaction approach.

These data present a fairly compelling argument for the validity of

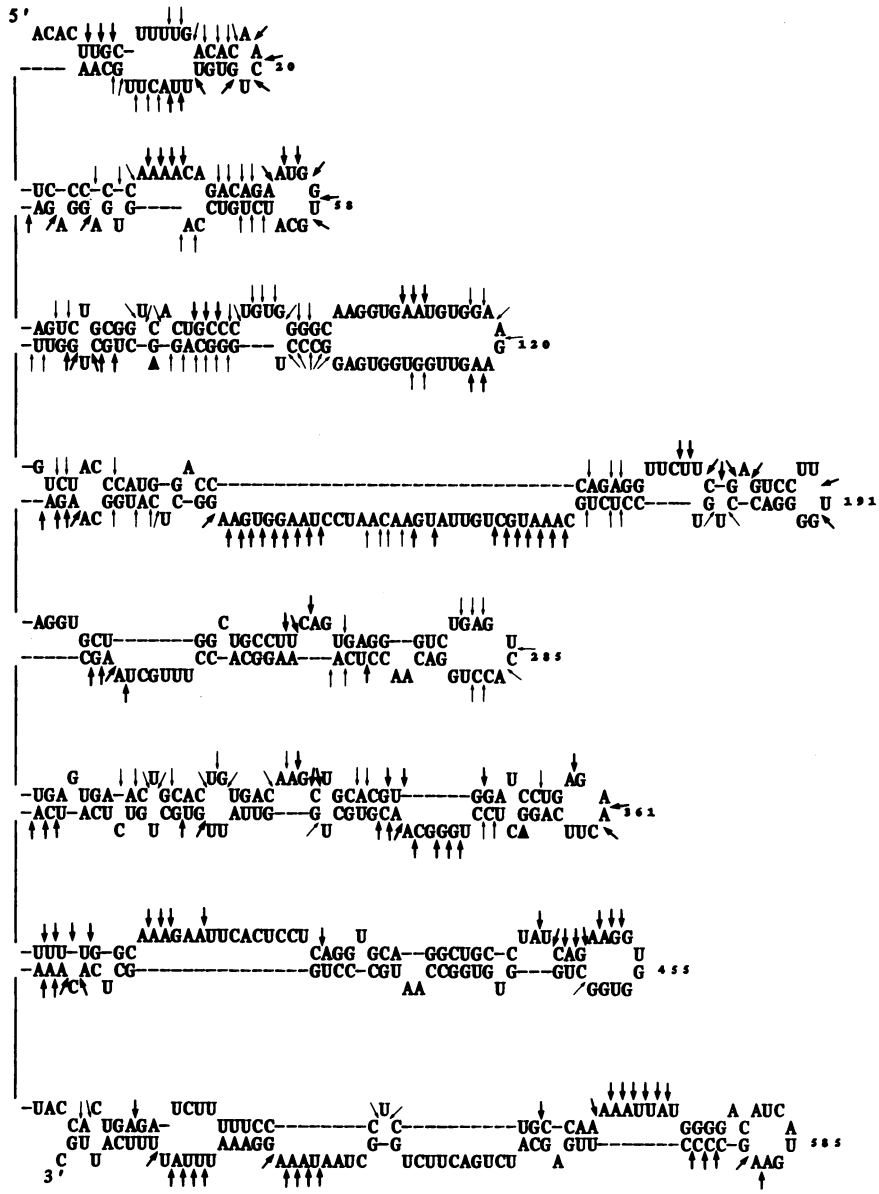

Figure 2. Secondary structure for rabbit β-globin messenger RNA proposed by RNA2 program. Bonds that are susceptible to cleavage by cobra venom ribonuclease are indicated by thin arrows, while those that are susceptible to cleavage by S1 nuclease are indicated by thick arrows. The two triangles indicate the positions where the two intervening sequences are excised in processing the messenger RNA. Compare with Figure 6.

thermodynamic RNA folding which is weighted toward short-range interactions. The thermodynamic rules, as determined from model oligonucleotide studies, probably represent a good starting point for RNA structure prediction. However, they are not in themselves complete, as indicated by the lack of ability to predict the entire 16S rRNA structure. Better results can be obtained by taking into account the relative imprecision of the underlying energy parameters and examining a small number of structures that are nearly as good as the single 'best' structure that the algorithm has picked. Nevertheless, accurate RNA predictions are at this time limited by our lack of understanding of the role of tertiary interactions and ordered single stranded structures in determining folding.

AN INTERACTIVE APPROACH

The approach described in the previous section takes into account thermodynamic criteria and the kinetics of short range folding, but it does not enable the user to interactively direct the generation of a proposed secondary structure based on his experimentally determined enzymatic digestion data or phylogenetic comparisons. A generalized nucleotide sequence handling facility has been developed that gives the user complete interactive control over the secondary structures that are generated through high resolution graphic displays and the use of a light pen and digitizing tablet. This facility is available on the PROPHET computer system (13), a national resource that is implemented on a DEC PDP-10 mainframe and provided by the NIH Division of Research Resources to some 700 biological scientists at 37 institutions throughout the United States.

The PROPHET secondary structure determination facility, which is written in PROPHET's own programming language, known as PL/PROPHET, begins with a structure derived through the application of a simple heuristic: those base-paired regions that are originally shown are chosen from all possible base-paired regions with stems greater than a minimum length specified by the user according to those regions that do not conflict with any longer base-paired regions. The user may also restrict the base-paired regions that are shown to any desired range of interaction lengths; thus he may take into account short-range kinetic interaction criteria. It is also possible to explicitly specify the initial base-paired regions he would like to show for all or part of the structure. PROPHET displays the secondary structure that it has generated or that the user has specified, and then allows the user to specify changes in base-paired regions, either by pointing to the regions in question with the light pen on the data tablet, or by typing residue indices.

The user may direct PROPHET to display any number of secondary structures for a given molecule, ultimately converging on a structure that is consistent with experimental, thermodynamic or phylogenetic data. Planned enhancements to the PROPHET facility will allow the user to specify regions that must be single or double stranded according to experimentally determined nuclease or chemical susceptibility results. Three examples are presented here in order to illustrate the manner in which this facility may be utilized.

1. E. coli 23S rRNA fragment

One of the first RNA molecules whose secondary structure was studied utilizing the tools available in PROPHET was a 3' fragment of Escherichia coli 23S ribosomal RNA isolated from an E. coli RNAse E mutant (14). PROPHET's built in heuristic that picks the longest possible base paired regions was used to generate a starting secondary structure, shown in the PROPHET display reproduced in Figure 3. We have added to the figure arrows that indicate sites where cobra venom ribonuclease, which is double-strand specific (15), cleaves the molecule and which are inconsistent with the initial candidate secondary structure. In addition, a large arrow is used to indicate a stem which is inconsistent with the fact that the bases in this region are susceptible to S1 nuclease, which is single-strand specific (16,17).

The first step the user took to modify this structure was to indicate, by pointing to the display using the light pen on the data tablet, three new base-paired regions, consistent with enzyme cleavage data, which are shown with brackets and the numbers 1, 2 and 3 in Figure 3. In addition, the user specified that he wanted to delete the short stem that was inconsistent with the S1 data by pointing to it in the interactive quiz. The resulting structure, which was generated by PROPHET using the three base paired regions specified by the user and then picking the longest possible paired regions from the remaining list of possible matches, is shown in Figure 4.

From this point, the user proceeded to delete base paired stems that were inconsistent with nuclease susceptibility data or which otherwise seemed unlikely for thermodynamic considerations, and letting PROPHET generate other potential base paired regions from its exhaustive list. After several intermediate steps that are not shown, the structure shown in Figure 5 was generated. On this figure, both S1 and cobra venom ribonuclease cleavage sites are indicated. Those parts of the molecule for which nuclease susceptibility data are available are quite consistent with these data, while the structure shown for the remainder of the molecule is of necessity quite arbitrary.

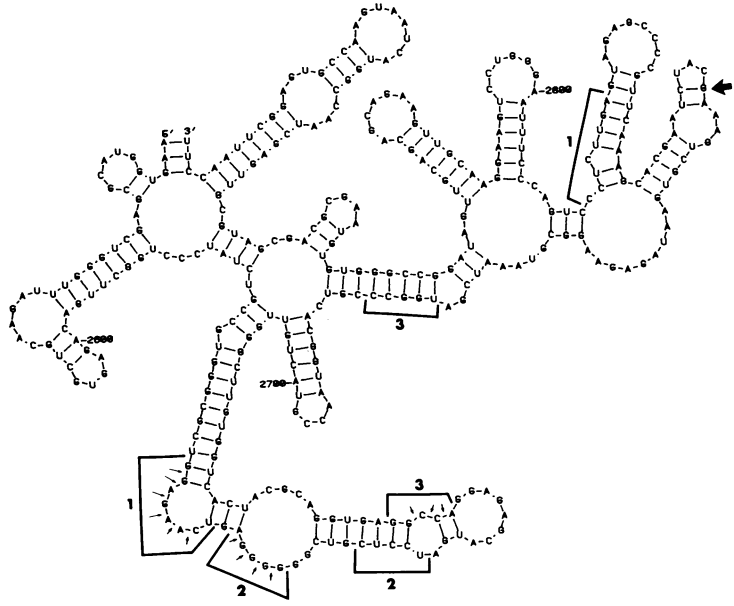

Figure 3. A PROPHET display of a heuristically chosen starting secondary structure for an E. coli 23S ribosomal RNA fragment. Sites that are susceptible to cobra venom ribonuclease cleavage and that are inconsistent with the secondary structure shown are indicated with small arrows. In making the next candidate secondary structure, shown in Figure 4, the user added three base-paired regions indicated by brackets and the numbers 1, 2 and 3 and deleted the short base-paired region indicated by the large arrow.

2. Rabbit β-globin mRNA

A slightly different approach was taken in determining a plausible secondary structure for rabbit β-globin messenger RNA. When we were studying the secondary structure of the 23S rRNA fragment, we entered the sequence we had determined into PROPHET using an interactive sequence editor. The rabbit β-globin gene sequence, on the other hand, was already available on PROPHET in a sequence database supplied by Los Alamos Scientific Laboratory. We retrieved this sequence from the database, excised the non-coding intervening sequences using the PROPHET sequence editor, and then transformed the DNA to the mRNA sequence by converting all T residues to U.

Since the β-globin sequence is considerably longer than the 23S rRNA fragment, we used PROPHET's selective display capability to display only designated parts of the molecule at any one time when we were carrying out the interactive editing of the base-paired regions. Thus, although the complete molecule was involved in generating each candidate secondary structure, only

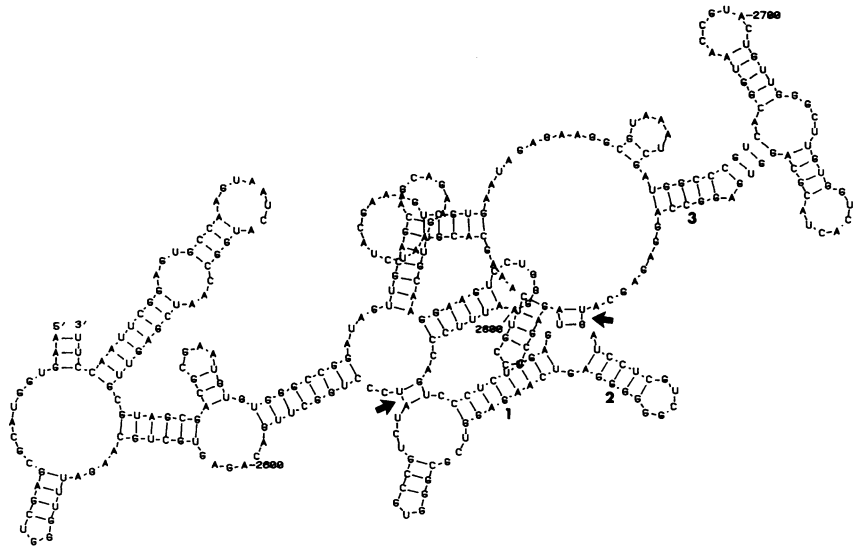

Figure 4. Intermediate secondary structure for an E. coli 23S ribosomal RNA fragment generated from the structure in Figure 3 after the user added the three base-paired regions indicated by the numbers 1, 2 and 3. The large arrows indicate two short base-paired regions that the user deleted in proceeding to the next intermediate structure (not shown).

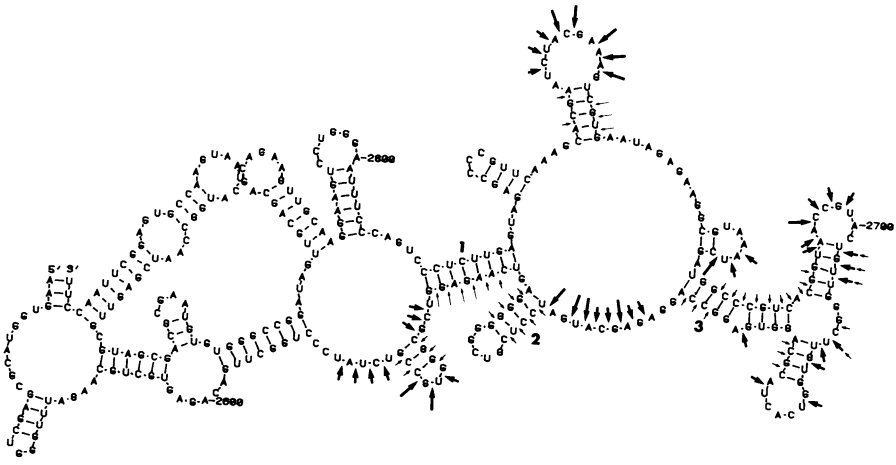

Figure 5. Final secondary structure for the E. coli 23S rRNA fragment at the end of the interactive PROPHET session. Cobra venom ribonuclease digestion sites are indicated by thin arrows, and S1 nuclease digestion sites are indicated by thick arrows. Long arrows represent strong cleavages, and short arrows represent weak cleavages. Several bonds are weakly susceptible to cleavage by both enzymes, as indicated.

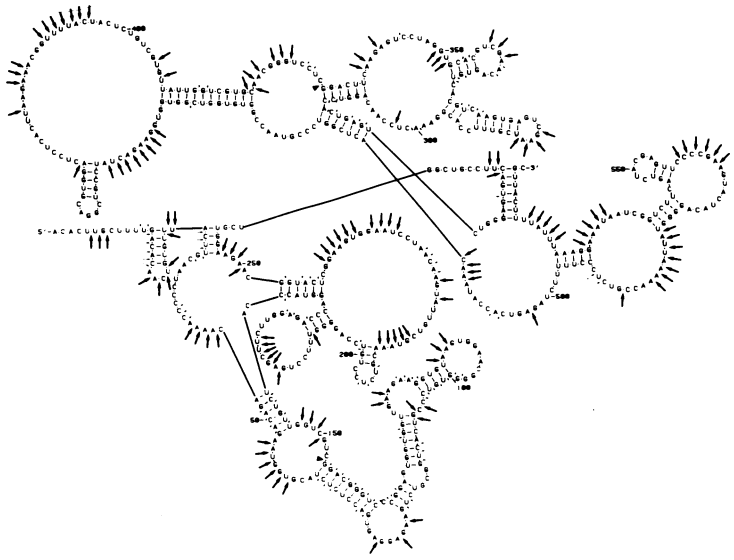

Figure 6. Proposed secondary structure for rabbit β-globin messenger RNA generated during an interactive PROPHET session. The six selectively displayed components of the molecule as shown by PROPHET are connected by hand-drawn lines. Arrows that indicate nuclease digestion sites were also added by hand. Thick arrows indicate S1 nuclease cleavage sites, while thin arrows indicate cobra venom ribonuclease digestion sites. The two triangles indicate the positions where two intervening sequences are excised in processing the messenger RNA. Compare with Figure 2.

the portion that was currently selected was displayed on the screen. The iterations to a final structure were otherwise quite similar to those used in generating the final 23S rRNA fragment structure. S1 nuclease and cobra venom ribonuclease susceptibility data were used to add and subtract base-paired regions until a structure that was consistent with the data had been reached. Figure 6 shows the final structure as displayed in six selective parts on PROPHET, with enzyme cleavage sites and important features added by hand.

The hypothetical β-globin structure presented in Figure 6 is the first structure proposed for eukaryotic messenger RNA which utilizes extensive data acquired on the disposition of residues in helical and single standed regions. This structure differs significantly from previous formulations of the secondary structure of the 5' proximal regions of the rabbit β-globin messenger RNA (18) for two reasons: first, only a portion of the total length of the globin messenger RNA had been considered in the earlier structures, thereby eliminating from consideration many possible alternative structures,

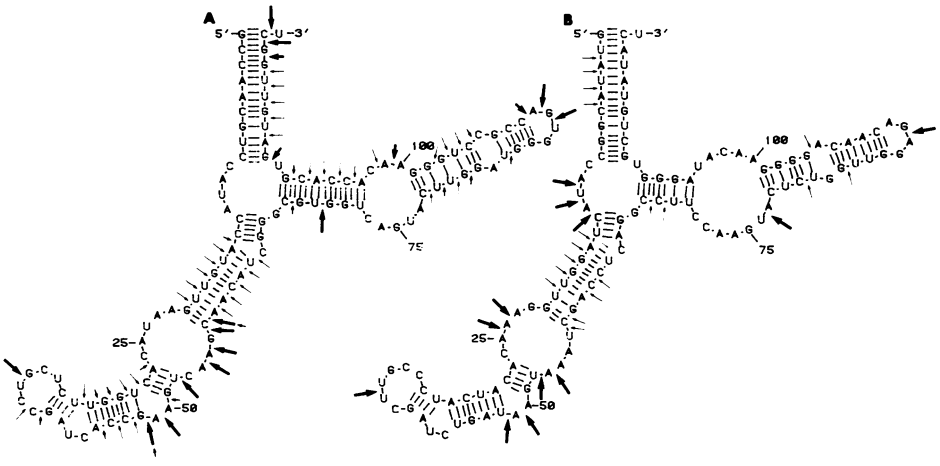

Figure 7. PROPHET generated secondary structures for (A) Bombyx mori and (B) Dictyostelium discoideum 5S ribosomal RNA molecules. S1 nuclease cleavage sites are indicated by thick arrows, and cobra venom ribonuclease cleavage sites are indicated by thin arrows.

and second, double-stranded cobra venom ribonuclease data were not available at the time the earlier structure was generated. Note that the β-globin structure generated by RNA2 in Figure 2 is also largely consistent with the S1 nuclease and cobra venom ribonuclease data.

3. Eukaryotic 5S rRNA molecules

Our final example involves the use of PROPHET to generate a consensus secondary structure for several eukaryotic 5S ribosomal RNA molecules that takes into account enzymatic and phylogenetic data. As described elsewhere (19), we sequenced two 5S rRNA molecules from Bombyx mori and Dictyostelium discoideum and then determined S1 nuclease and cobra venom ribonuclease cleavage sites for the two molecules. Using PROPHET, it was possible to readily determine candidate secondary structures for each molecule that were consistent with the available enzymatic data. One such structure for each of these molecules is shown in Figure 7, and the similarity between the proposed secondary structures for the two molecules is quite striking, particularly in the highly conserved sequences in some of the proposed single stranded regions.

We carried this analysis one step further by using PROPHET to analyze the secondary structures of several other 5S rRNA molecules, using phylogenetic similarities to direct the base-paired regions that were added or deleted, by,

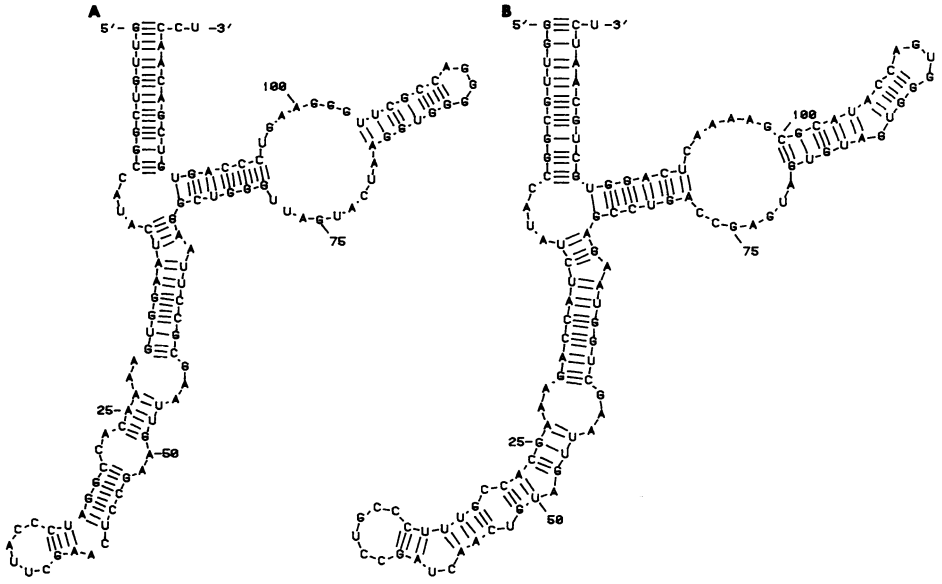

Figure 8. PROPHET generated secondary structures for (A) Tetrahymena thermophila and (B) Saccharomyces cerevisiae 5S ribosomal RNA molecules.

for example, keeping highly conserved sequences in single-stranded regions. We found that it was possible to propose a secondary structure for each 5S rRNA molecule we examined that was quite similar to the proposed structures for B. mori and D. discoideum 5S rRNA. Proposed structures for four additional published 5S rRNA sequences (20) are shown in Figure 8 and Figure 9.

Discussion

In this paper, we have discussed two approaches to determining potential secondary structures for RNA molecules. The first approach emphasizes the use of a sophisticated thermodynamic energy minimization algorithm to generate candidate structures with energetically stable base-paired regions, favoring relatively short range interactions over longer range interactions. The second approach utilizes an interactive computer graphic environment to allow the user to impose thermodynamic or experimentally determined structural criteria on a structure generated by a simple set of heuristics or defined by the user at the outset. The first approach, in its present implementation, does not enable the user to interactively modify the structure that the thermodynamic algorithm generates, while the second approach does not enable

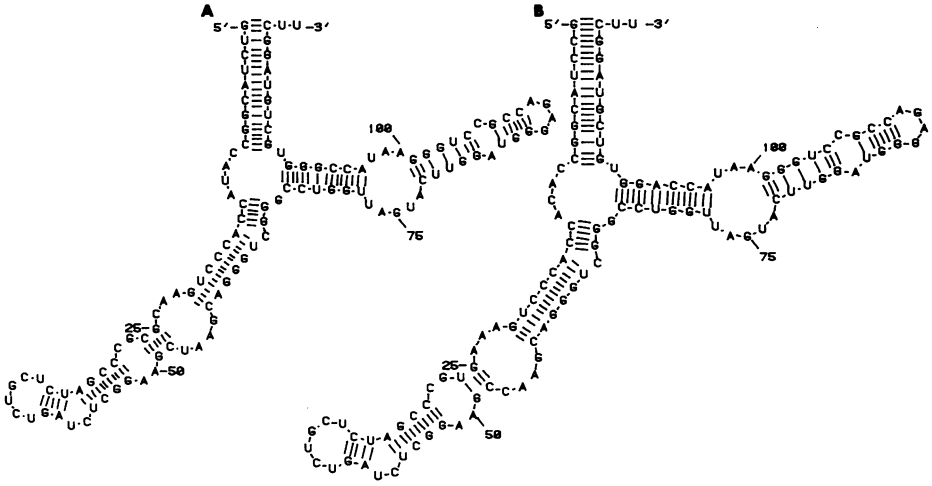

Figure 9. PROPHET generated secondary structures for (A) Homo sapiens (from Hela cell culture) and (B) Xenopus laevis 5S ribosomal RNA molecules.

the user to automatically generate local or global thermodynamically minimized structures.

Each of these two approaches has considerable merit in its own right. We have shown that the thermodynamic algorithm identifies a number of base paired regions that are consistent with available experimental data in many RNA molecules. One can, however, generate examples where it appears to find only a small portion of the purported base-paired regions, and it seldom if ever predicts a complete structure that is consistent with experimental data in every detail. The interactive approach, on the other hand, gives the user complete control over and plentiful feedback on the secondary structure it ultimately generates, using whatever criteria the user prefers, but it can become quite tedious if the user wishes to set up a secondary structure based on the output of a thermodynamic algorithm run on a separate system.

We believe that a comprehensive integration of these two approaches would provide a very effective tool for studying the potential secondary structures of RNA molecules. We are therefore proceeding with the implementation of a version of the thermodynamic algorithm described here on the NIH PROPHET computer system, and with modifying the PROPHET-based interactive algorithm to accept structural information generated by the thermodynamic algorithm. It will be possible for the user to start with a candidate structure generated by the thermodynamic algorithm, and at any point in his interactive refinement of

the structure he will be able to invoke the thermodynamic algorithm on any piece of the molecule that he selects. He will also be able to ask the system for theoretical energy measurements for any structure or structural detail that he would like to test.

There are several other additional features that could be added to PROPHET's interactive approach. For example, it may be desirable to be able to access more than one thermodynamic algorithm in this context. It would also be possible to incorporate heuristics that take into account the fact that the molecule folds as it is synthesized from the 5' to the 3' end as well as the relative instability of G-U pairs in generating candidate structures. Any heuristics or algorithms that are added to help the user generate candidate hairpin regions will always remain under the complete control of the the user, since the essence of this approach is the ability it gives the user to accept or reject proposed base-paired regions based on whatever evidence he may have available.

ACKNOWLEDGMENTS

We thank Alexander Rich, Gary Quigley, Regina Wurst, Lauren Rieser and George Pavlakis for advice and discussions, Alexander Rich for facilities and support, David Aperion for supplying purified RNA, Walter Goad for providing the Los Alamos sequence database, and James K. Prater and Harold Perry for expert computer technical assistance. Special thanks are offered to Michael Zuker for valuable discussions and for providing his thermodynamic RNA folding programs. This work was supported in part by grants from NIH (GM 22280) and NSF (TCM 8004620) to J.V. P.A. was supported by Fellowship DRG-425F of the Damon Runyon-Walter Winchell Cancer Fund.

+To whom reprint requests should be addressed.

REFERENCES

1. Pipas, J.M. and McMahon, J.E. (1975) Proc. Nat. Acad. Sci. USA 72, 2017-2021
2. Studnicka, G.M., Rahn, G.M., Cummings, I.W., Salser , W.A. (1978) Nucleic Acids Res. 5, 3265-3387
3. Zuker, M. and Stiegler, P. (1981) Nucleic Acids Res. 9, 133-148
4. Tinoco, I., Uhlenback, O.C. and Levine, M.D. (1971) Nature 230, 362-367
5. Salser, W. (1977) Cold Spring Harbor Symp. Quant. Biol. 42, 985-1002
6. Alden, C.J. and Kim, S.-H. (1980) in Nucleic Acid Geometry and Dynamics, Sarma, R.H. Ed., pp. 399-418, Pergamon Press, New York
7. Kallenbach, N.R., Mandal, C. and Englander, S.W. (1980) in Nucleic Acid Geometry and Dynamics, Sarma, R.H. Ed., pp. 233-251, Pergamon Press, New York
8. Fresco, J.R., Alberts, B.M. and Doty, P. (1960) Nature 188, 98-101
9. Woese, C.R., Magrum, L.J., Gupta, R., Siegel, R.B., Stahl, D.A., Kop, J., Crawford, N., Brosius, J., Gutell, R., Hogan, J.J. and Noller, H.F. (1980) Nucleic Acids Res. 8, 2275-2293
10. Noller, H.F. and Woese, C.R. (1981) Science 212, 403-411

11. Stiegler, P., Carbon, P., Zuker, M., Ebel, J.-P. and Ehresmann, C. (1981) Nucleic Acids Res. 9, 2153-2172
12. Spillman, S., Dohme, F. and Nienhaus, K.H. (1977) J. Mol. Biol. 115, 513-523
13. Raub, W.F. (1974) Federation Proc. 33, 2390-2392
14. Ghora, B.K. and Aperion, D. (1981) J. Biol. Chem. 254, 1951-1956
15. Vassilenko, S.K. and Rait, V.K. (1975) Biokhimiya 40, 578-583
16. Wurst, R. and Vournakis, J.N. (1977) J. Cell Biol. 75, 356a
17. Wurst, R., Vournakis, J.N. and Maxam, A. (1978) Biochemistry 17, 4493-4499
18. Pavlakis, G.N., Lockard, R.E., Vamvakopoulos, N., Rieser, L., RajBhandary, U.L. and Vournakis, J.N. (1980) Cell 19, 91-102
19. Troutt, A., Savin, T.J., Celentano, J., Curtiss, W.C. and Vournakis, J.N. (1981) to be submitted
20. Erdmann, V.A. (1981) Nucleic Acids Res. 9, r25-r42