

Method for Predicting RNA Secondary Structure

(hydrogen bonding/computer modeling/tRNA)

JAMES M. PIPAS AND JAMES E. McMAHON

Institute of Molecular Biophysics, Florida State University, Tallahassee, Fla. 32306

Communicated by Michael Kasha, March 19, 1975

ABSTRACT We report a method for predicting the most stable secondary structure of RNA from its primary sequence of nucleotides. The technique consists of a series of three computer programs interfaced to take the nucleotide sequence of any RNA and (a) list all possible helical regions, using modified Watson-Crick base-pairing rules; (b) create all possible secondary structures by forming permutations of compatible helical regions; and (c) evaluate each structure for total free energy of formation from a completely extended chain. A free energy distribution and the base-by-base bonding interactions of each possible structure are catalogued by the system and are readily available for examination. The method has been applied to 62 tRNA sequences. The total free-energy of the predicted most stable structures ranged from -19 to -41 kcal/mole (-22 to -49 kJ/mole.) The number of structures created was also highly sequence-dependent and ranged from 200 to 13,000. In nearly all cases the cloverleaf is predicted to be the structure with the lowest free energy of formation.

We have developed a technique for predicting the secondary structure of RNA from its primary sequence. The method uses thermodynamic and structural criteria to generate all possible secondary interactions in the molecule and evaluates each for free energy. The technique can be used in conjunction with experimental procedures to elucidate the most favorable conformation of a polyribonucleotide chain.

The secondary and tertiary structure of RNA is assumed to play an important role in determining the interactions of these macromolecules with proteins. In the most studied case, that of the tRNAs, it has been found that some form of structural integrity other than the primary sequence of nucleotides must be maintained in order to ensure the biological activity of these molecules (1, 2). This is not surprising, since the tRNAs must be able to interact specifically with a myriad of proteins, including those involved in (a) tRNA maturation (i.e., methylases, thiolases, nucleases, and other modifying enzymes); (b) amino-acid activation (the aminoacyl synthetases); and (c) other translational factors (ribosomal proteins, initiation factor 2, and probably others). In addition to their function in translation, some tRNAs have been shown to be involved in the autogenous regulation of certain cistrons and, thus, must also interact with certain transcriptional components (3).

Likewise the rRNAs are believed to possess a definite secondary and tertiary structure which serves to govern their interaction with ribosomal proteins (4). The genomes of certain RNA bacteriophages have also been shown to fold owing to secondary and tertiary interactions (5, 6). In these cases too, the evidence demonstrating the functional significance of specific structure is convincing.

Studies attempting to demonstrate structure-function relationships in the RNAs have been hindered by a lack of knowledge concerning the exact nature of the structure of

these molecules in solution. The use of common physical probes such as nuclear magnetic resonance, circular dichroism-optical rotatory dispersion, and x-ray diffraction is difficult and time consuming. With these techniques only one polynucleotide sequence can be examined at a time, where it is often useful to compare a number of similar sequences and search for common features.

It is generally assumed that the information governing three-dimensional folding is contained in the primary sequence of nucleotides that make up the polynucleotide chain. Therefore, it should be possible to deduce these interactions from the nucleotide sequence. Several attempts at predicting these interactions have been made (7–10).

In this paper we report a new method for predicting the secondary structure of RNA, given its primary sequence. The technique consists of a series of three computer programs which (a) creates a list of all possible helical regions that can be derived from a given sequence; (b) tests each of these regions against all other regions for structure compatibility; (c) lists all possible structures derivable from the primary sequence by creating all permutations of compatible helical regions; and (d) evaluates each possible structure for total free energy and several other parameters. The method is entirely general and can be applied to any polyribonucleotide chain up to 150 bases long. With some minor modifications in storage and the use of packing routines, RNAs up to 3500 nucleotides long can be handled.

We further show the utility of this technique by applying it to 62 known tRNA sequences. It is demonstrated that for most of these sequences, the cloverleaf structure or some close variant is the lowest free energy form when only secondary interactions are considered. Some interesting exceptions to this result are discussed. Finally, we show how this method can be used to create a free energy distribution for each sequence that takes into account all possible acceptable structures. Further applications of this system are discussed.

METHOD

tRNA Sequences. The nucleotide sequences for the 62 tRNAs used were those compiled by Sodd and Doctor (11). Each sequence was coded as to the hydrogen-bonding type of its bases. This is necessary, since the tRNAs contain modified bases that have altered hydrogen-bonding capabilities. All modified bases were examined and classified into one of five types: G, C, U, A, O. Bases coded as O are blocked by modification at key groups and cannot form hydrogen bonds.

Computer Methods. The method consists of a series of three computer programs which take as input the primary sequence of nucleotides and generate as output the base-by-base bonding that defines the structure with the most favorable

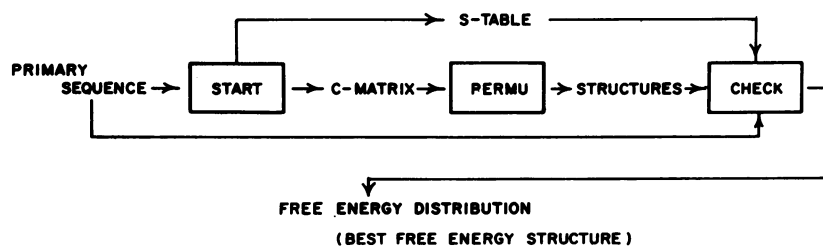


FIG. 1. Generalized flow diagram showing the input and output of information through the three computer programs in the system. START takes the primary nucleotide sequence as input and generates a list of all possible helical regions derivable from the sequence (S-Table) and a matrix indicating the compatibilities of these regions (C-matrix). The C-matrix is received by PERMU, which yields as output all possible structures by generating all permutations of nonzero elements in the matrix. CHECK receives the primary sequence, the S-Table, and the structures as input and evaluates each structure base-by-base for total free energy and several other parameters. The structure with the lowest free energy is predicted as the most stable.

free energy when only secondary interactions are considered. Fig. 1 is a generalized diagram showing the flow of input and output through the three programs, which are arbitrarily assigned the names START, PERMU, and CHECK.

START: The initial program, START, reads the nucleotide sequence and sets up a complete bonding matrix for the sequence, using the following rules: (1) If base i is able to form a classical Watson-Crick hydrogen-bonding pair with base j , then the matrix element $B_{i,j} = 1$. A classical bonding pair corresponds to a G·C or A·U base pair. (2) If base i forms a G·U pair with base j , then $B_{i,j} = 2$. The nonclassical G·U base pair is allowed in the special cases discussed below. (3) If base i and base j do not form one of the types of pairs discussed above, $B_{i,j} = 0$.

Next the program searches for possible stable regions of hydrogen bonding (helical regions) which might occur in a structure. A stable helical region is defined as three or more consecutive base pairs ordered such that the strands are antiparallel. On the bonding matrix this corresponds to finding a set, $\{B_{i,j}, B_{i+1, j-1}, \dots, B_{i+m, j-m}\}$ where all elements are nonzero. The additional restriction is added that G·U base pairs ($B_{i,j} = 2$) cannot occur at either end of a region.

The helical regions thus created are yielded as output by START in the form of an S-Table. Here the bonded bases in each region are stated explicitly by number; the base of the 5'-terminus is assigned the number one and each ensuing base is numbered sequentially (Fig. 2). At the time of creation of the S-Table, several previously accepted regions are eliminated as energetically or sterically unfavorable and some new regions are allowed. No region, for example, may close a hairpin (loop where the chain doubles back on itself) of less than three bases. If such occurs, the last pair is opened to form a three- or four-base hairpin and the region will contain one less bonded pair. If after such an operation, less than three base pairs remain in a region, the region is eliminated from the S-Table. In cases where a G·U base pair occurs second from the end of a region of five or more base pairs, a new region is created by opening these last two pairs.

After the program has compiled this list of all possible helical regions derivable from the given primary sequence it sets up a compatibility matrix (C-matrix) to indicate which regions can occur together in a given structure. The elements defining the diagonal of this matrix are set equal to one and the matrix is symmetric about this diagonal. If regions R_i and R_j are found to be compatible (i.e., they can exist together in a given structure) then the matrix element $C_{i,j} = 1$. If they are not compatible $C_{i,j} = 0$. Two criteria are used to

test compatibility. First, compatible regions cannot contain any of the same bases. This is defined mathematically as follows: let R_i be a set whose elements are the bases (indexed by numbers as explained above) contained in region i . Let R_j be a set whose elements are the bases contained in region j . Then the matrix element $C_{i,j} = 1$ if and only if $R_i \cap R_j = \emptyset$. In all other cases $C_{i,j} = 0$. The exception is those elements on the diagonal that are defined as being equal to one. The second test for compatibility means physically that all bases in a hairpin are restricted to bond only with other bases within that same hairpin. To express this let H_i be a set whose elements are the bases in the hairpin closed by region i . Then let A_j be a set whose elements are the bases in region j . Finally, let Q be a set whose elements are all bases not included in H_i , A_j , or i . The matrix element $C_{i,j} = 1$ if and only if $A_j \cap H_i = \emptyset$ or $A_j \cap Q = \emptyset$. In all other cases $C_{i,j} = 0$, again with the exception of the diagonal. The reasons for this restriction will be discussed below.

PERMU: The compatibility matrix created by START serves as the input for the second computer program in the system, which is called PERMU. The function of PERMU is to create all possible structures obtainable from the given polynucleotide sequence. It accomplishes this by generating all possible permutations of the nonzero (compatible) elements in the C-matrix. Here, a structure is defined as a set of three or more compatible helical regions. The output generated comprises a list of all secondary structures that are possible within our set of restrictions. By indexing against both the S-Table generated from START and the primary sequence of nucleotides a base-by-base bonding scheme can be obtained for each of the generated structures.

CHECK: The third program in the system, termed CHECK, evaluates each of the generated structures one at a time, base by base, assigning a total free energy to each one. The structures are then ordered by their free energies, and the best (most negative) free energy structure is printed out explicitly along with all other structures that occur within 5 kcal/mole (21 kJ/mole) of it. The best free energy structure for tRNA^{A1a} is shown in Fig. 2. CHECK also compiles several other parameters which we shall not discuss here.

Free Energy Assignments. Favorable free energy contributions are assumed to be made by stacking interactions within regions of hydrogen-bonded base pairs. The specific values assigned are empirical and were obtained by Gralla and Crothers in 1973 (12). However, we have excluded the contributions from terminal G·U base pairs.

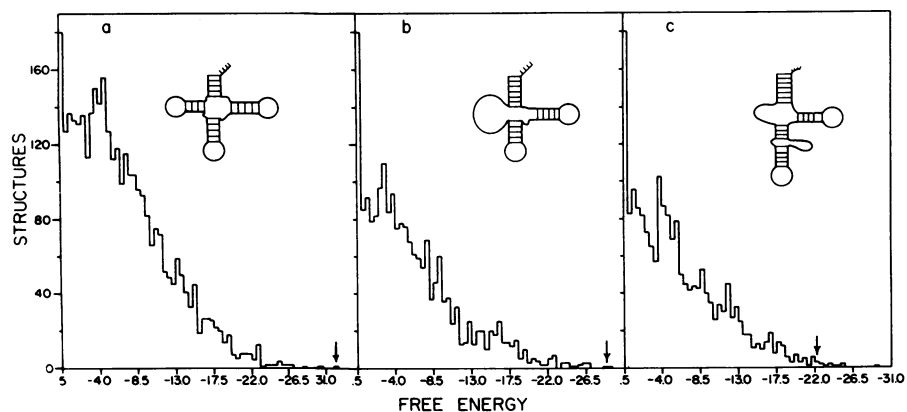


FIG. 4. Histogram showing the free energy distributions for three tRNA sequences. (a) tRNA^{Ala} (Class I), (b) tRNA^{Trp} (Class II), and (c) tRNA^{Leu} coded by T4 bacteriophage (Class III). The stick diagram in the upper corner of the graph shows the general form of the best free-energy structure for the given sequence. The arrow indicates where on the histogram the cloverleaf as proposed in the literature occurs. It should be pointed out that while the shapes of these graphs are fairly typical, the free energies of the favorable structures (ranging from -19 to -41 kcal/mole) and the total number of structures acceptable (ranging from 200 to 13,000) are highly sequence-dependent.

best free-energy form. In 13 cases a close variant of the cloverleaf was preferred. Out of these sequences eight had open dihydrouridine stems, two had open T ψ C stems, two had open extra stems, and one had an open anticodon stem. In these cases positive free energy contributed by the hairpin outweighed the stacking interactions of the region and the base-pairs opened. In 10 of these Class II sequences the cloverleaf was the second best free-energy structure. It has been reported that for at least one tRNA the dihydrouridine stem melts out with tertiary interactions and that the two events are inseparable (14). It is thus possible that tertiary structure serves to stabilize those helical regions we predict will open in Class II sequences. In most of the sequences, a favorable tertiary interaction of small magnitude neglected in our treatment would be sufficient to make the cloverleaf the best free-energy structure.

Class III sequences fall into two general categories. Those which predict a structure other than the cloverleaf because of a faulty assumption made in the system (Class IIIa), and those for which the cloverleaf is apparently not the best free energy structure (Class IIIb). Five sequences fall into the former classification and 12 in the latter. The reason for the failure in these cases will be discussed below. All Class III sequences produce some form of extended structure rather than the cloverleaf as having the best free energy. In all cases the cloverleaf or some close variant occurs within 5 kcal/mole of the predicted structure. Extended forms occur at the favorable end of the free energy distribution for nearly all sequences. The significance of these forms will be discussed in detail in a later communication (in preparation).

An examination of the cloverleaf structure for any of the five cases where the method has failed instantly yields the reason for the failure. The acceptor stem of these molecules contains two proposed regions of hydrogen-bonding separated by a two-base internal loop. Our method finds one of the helical regions but the second is discarded because it contains only two base pairs. If the two-base pair region is included it will contribute a favorable free energy to the structure. This free energy is enough to make the cloverleaf the best free-energy form. It should be pointed out that extended forms are still close competitors (within 1 kcal/mole) in these cases.

Since the computer programs generate all possible structures from a given sequence and assign a free energy to each structure, it is possible to plot a free energy distribution for the structures created from each tRNA. Fig. 4 shows the free energy distribution obtained for three different tRNAs. The free energy distributions obtained for all tRNAs examined are highly sequence-dependent. The shape of the distributions, number of acceptable structures, and free energy of the predicted structure vary greatly from one tRNA to another.

DISCUSSION

The cloverleaf is the generally accepted model for describing the secondary interactions of tRNA. Evidence for the biological importance of this structure comes from a variety of sources, none of which are totally convincing. First, all known tRNA primary sequences can be arranged in the form of a cloverleaf structure by using classical Watson-Crick base pairing rules with the occasional inclusion of G·U pairs. The cloverleaf model is consistent with a large body of spectral data, including circular dichroism-optical rotatory dispersion and nuclear magnetic resonance studies (15, 16). For at least one case, in which a specific tRNA was isolated and crystallized, yeast phenylalanine tRNA has been shown to be in the form of a cloverleaf by x-ray diffraction (17, 18). Of course, all of these data have been accumulated on the relatively few species of tRNAs that have been purified. The generality of the results obtained has been assumed.

Using some empirical assumptions and thermodynamic data, our method predicts that the cloverleaf will be the form of lowest free energy in at least 50 of 62 tRNA sequences examined. In all cases the cloverleaf, or some close variant, is at the favorable end of the free energy distribution. It must be realized that we are considering only secondary interactions here. There are strong indications that tertiary interactions stabilize certain secondary structures more than others. This may be the case in the 12 sequences for which an extended form is predicted to be more stable than the cloverleaf. In all of these cases the cloverleaf is within 5 kcal/mole of the predicted structure. The fact that tertiary structure will stabilize certain secondary interactions has been established for at least one species of tRNA (14).

On the basis of these results, we would assert that the cloverleaf is certainly an important structure for tRNA sequences and, in fact, in most cases is the best free-energy form of the macromolecule. Various types of extended structures are also important for certain sequences.

One of the major advantages the method presented here has over other techniques for predicting RNA structure is the fact that all possible structures are produced and evaluated under a set of restrictions. From the free energy distribution created for each sequence, it is possible to ascertain exactly how favorable the predicted structure is and exactly what its close competitors are. Thus, the results obtained from our method can be easily compared with experimental data. The manner in which the programs are set up also allows each of the assumptions built into the system to be analyzed in detail. Thus, it is easy to vary the value of certain free energy assignments, or lift certain steric requirements and observe the effect on the outcome. This allows the investigator great flexibility and critical discrimination in examining some of the more tenuous assumptions. It should be pointed out here that the cost of utilizing this system is low both in terms of computing time and core memory. All 62 tRNA sequences were evaluated in about 3000 seconds (octal) on the CDC6500 computer.

In order to reduce the number of possible structures for a given sequence, we have made several simplifying assumptions. Most of these are introduced in the initial program, START. Two of these will be discussed here, since they both grossly affect our results. The first major assumption is that we only consider helical regions consisting of three or more consecutive base pairs. While in certain cases shorter regions are favorable, we felt that their contribution to the total free energy of a given structure would be negligible. However, for at least five tRNAs the favorable free energy from a region of two G·C base pairs is the difference between predicting the cloverleaf or an extended form as the most probable structure.

A second major assumption made is that no bases occurring within a hairpin may bond with bases outside that region (see discussion of C-matrix). Such bonding has been proposed for the 5S RNA from *Escherichia coli* and the denatured form of tRNA_{3^{Leu}} from yeast (19, 20). While it would be easy to allow this type of bonding at the stage of the C-matrix (START), it would be difficult to evaluate the resulting stabilizing and destabilizing free energy contributions.

This study was undertaken to develop a general method for the prediction of the secondary interactions in RNA. While the technique was applied here only to tRNA sequences, the method is general and can be used to predict the secondary structure of any class of RNA.

Note Added in Preparation. M. Levine and I. Tinoco have developed a similar method for predicting RNA secondary structure (24).

This work was a collaborative effort between two research groups. J.P. is a graduate student in the laboratory of Robert H. Reeves. J.M. is a graduate student in the laboratory of William Rhodes. We thank Robert H. Reeves for help in coding the modified bases as to bonding type and for advice concerning the biology and chemistry of tRNA. We also wish to acknowledge Robert H. Reeves and William Rhodes for many helpful discussions and for encouraging us in this project. We thank Patrick Shannon for work on computer programming in the early stages of this project. We also thank Dennis Cravens for helpful discussions relating to computer techniques. This research was supported by a grant from the Division of Biomedical and Environmental Research, Atomic Energy Commission no. AT-(40-1)-2690 and funds from the Florida State University computing center.

1. Gartland, W. J. & Sueoka, N. (1966) *Proc. Nat. Acad. Sci. USA* **55**, 948-956.
2. Fresco, J., Adams, A., Ascione, R., Henley, D. & Lindahl, T. (1966) *Cold Spring Harbor Symp. Quant. Biol.* **36**, 527-537.
3. Goldberger, R. (1974) *Science* **182**, 810-816.
4. Pace, N. R. (1973) *Bacteriol. Rev.* **37**, 562-603.
5. Gralla, J., Steitz, J. & Crothers, D. (1974) *Nature* **248**, 204-208.
6. Weissmann, C., Billeter, M., Goodman, H. M., Hindley, J. & Weber, H. (1973) *Annu. Rev. Biochem.* **42**, 303-328.
7. Fresco, J. R., Alberts, B. M. & Doty, P. (1960) *Nature* **188**, 98-101.
8. Tumanyan, V., Sotnikova, L. & Kholopov, A. (1966) *Dokl. Akad. Nauk SSSR* **166**, 1465-1468.
9. Delisi, C. & Crothers, D. (1971) *Proc. Nat. Acad. Sci. USA* **68**, 2682-2685.
10. Tinoco, I., Uhlenbeck, O. & Levine, M. (1971) *Nature* **230**, 362-367.
11. Sodd, M. & Doctor, B. (1974) in *Methods in Enzymology*, eds. Grossman, L. & Moldave, K. (Academic Press, New York), Vol. 29, pp. 741-746.
12. Gralla, J. & Crothers, D. (1973) *J. Mol. Biol.* **73**, 497-511.
13. Cole, P., Yang, S. & Crothers, D. (1972) *Biochemistry* **12**, 4358-4368.
14. Crothers, D., Cole, P., Hilbers, C. & Shulman, R. (1974) *J. Mol. Biol.* **87**, 63-88.
15. Kearns, D. & Shulman, R. (1974) *Accounts of Chemical Research* **7**, 33-39.
16. Shulman, R., Hilbers, C., Wong, Y., Wong, K., Lightfoot, D., Reid, B. & Kearns, D. (1973) *Proc. Nat. Acad. Sci. USA* **70**, 2042-2045.
17. Robertus, J. D., Ladner, J., Finch, J., Rhodes, D., Brown, R., Clark, B. & Klug, A. (1974) *Nature* **250**, 546-551.
18. Kim, S., Suddath, F., Quigley, G., McPherson, A., Sussman, J., Wang, A., Seeman, N. & Rich, A. (1974) *Science* **185**, 435-440.
19. Kearns, D. & Wong, Y. (1974) *J. Mol. Biol.* **87**, 755-774.
20. Kearns, D., Wong, Y., Chang, S. & Hawkins, E. (1974) *Biochemistry* **13**, 4736-4746.
21. Uhlenbeck, O., Borer, P., Dengler, B. & Tinoco, I. (1973) *J. Mol. Biol.* **73**, 483-496.
22. Gralla, J. & Crothers, D. (1973) *J. Mol. Biol.* **78**, 301-319.
23. Jacobson, H. & Stockmayer, W. (1950) *J. Chem. Phys.* **18**, 1600-1606.
24. Levine, M. (1974) Ph.D. Dissertation, University of California, Berkeley.