Volume 9 Number 1 1981 Nucleic Acids Research

Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information

Michael Zuker and Patrick Stiegler+

Division of Biological Sciences, National Research Council of Canada, Ottawa K1A 0R6, Canada

Received 5 November 1980

ABSTRACT

This paper presents a new computer method for folding an RNA molecule that finds a conformation of minimum free energy using published values of stacking and destabilizing energies. It is based on a dynamic programming algorithm from applied mathematics, and is much more efficient, faster, and can fold larger molecules than procedures which have appeared up to now in the biological literature. Its power is demonstrated in the folding of a 459 nucleotide immunoglobulin γ1 heavy chain messenger RNA fragment. We go beyond the basic method to show how to incorporate additional information into the algorithm. This includes data on chemical reactivity and enzyme susceptibility. We illustrate this with the folding of two large fragments from the 16S ribosomal RNA of *Escherichia coli*.

INTRODUCTION

The sequence of nucleotides of an RNA molecule carries the information required for its actual conformation in three dimensions. Investigating the secondary structure, or folding, of the nucleotide chain may lead to a first sketch of the organization of the molecule. Up to the present, RNA secondary structures have been predicted by applying various topological and thermodynamic rules for finding the energetically most favourable structure for a given sequence. Tinoco *et al.* (1) devised a basic method to estimate RNA secondary structures. Stability numbers were assigned to the predicted structures according to melting temperature and thermodynamic data on double stranded oligoribonucleotides and polyribonucleotides. Estimates of free energy contributions were then experimentally refined by several authors (2-4). The stability numbers gave way to more precise stacking and destabilizing energies, allowing improvements in the method for predicting RNA secondary structures (5). More recently, thermodynamic data available from the literature have been compiled by Salser (6).

Prediction of RNA secondary structures using computer methods that attempt to minimize free energy is not new. An early attempt to predict

secondary structure in RNA using thermodynamics was published by Pipas and McMahon (7). This early technique can easily handle relatively short sequences such as transfer RNA's, but is inefficient for folding long RNA sequences since it requires computation time proportional to 2^N, where N is the number of nucleotides in the sequence. A vast improvement of this technique was published by Studnicka *et al.* They define precise topological rules for folding an RNA molecule and make full use of available thermodynamic data. However, their algorithm is unnecessarily complicated and requires several stages involving human intervention in order to arrive at a solution. In addition, it is slow and inefficient when compared with optimization algorithms from applied mathematics. Finally, because of various shortcuts and compromises which are made, the solutions are in general suboptimal, especially for longer chains.

On the mathematical side, there are two relevant papers that are worth mentioning. Nussinov *et al.* (9) have formulated an excellent dynamic programming algorithm which obtains maximum base pairing in a folded molecule. It can be easily modified to assign different weights to the various kinds of base pairings, but it cannot be used to compute optimal structures according to the thermodynamic criteria compiled by Salser (6). Stacking and destabilizing energies cannot be incorporated into this algorithm. Waterman and Smith (10) have define an algorithm that is less efficient, but takes into account stacking and destabilizing energies, although their energy calculations are not those used by Studnicka *et al.* (8). Furthermore, their algorithm is limited to sequences of two hundred nucleotides or less.

This paper presents what is in essence a fusion of the two approaches mentioned above. We use the identical folding rules and energy computations as do Studnicka *et al.* (8). The improvement is the creation of a dynamic programming algorithm specially designed for this problem. Thus it differs in subtle but important ways from the algorithms of Nussinov *et al.* (9) and Waterman and Smith (10). The result is an algorithm that has already been used successfully to fold sequences up to 600 nucleotides long and could probably deal with sequences of up to 800 nucleotides. It computes an optimal structure for a sequence of N nucleotides in time proportional to N^3, a real improvement compared with N^5 for the program of Studnicka *et al.* (8). Computation time is virtually negligible for sequences up to 200 nucleotides long. The program requires no human intervention and works in a single computation cycle. Finally, because no compromises are made to save time or whatever, the structure produced has the minimum possible free energy according to the rules used.

We go beyond the usual folding rules by incorporating additional information into our algorithm. To be specific, when there is information on the reactivity of certain nucleotides to chemical modification , or when enzymatic studies reveal which phosphodiester bonds are most likely to be cut under conditions of partial hydrolysis, we can build this information directly into the algorithm and predict an optimal structure compatible with these data. Even phylogenetic data on secondary structure conservation or evidence of specific long range interactions from the examination of RNA digests can be used by the algorithm. Indeed, our basic premise is that the use of such additional information is not only desirable, but essential. The basic folding rules and the thermodynamic information available today are simply inadequate to predict a correct secondary structure with much confidence, no matter how powerful a computer program is used. The more additional information used, the better the predicted folding will be. This will be illustrated later in the discussion of the folding of *E.coli* 16S ribosomal RNA.

METHODS

A. Definitions

In this section we define the group of structures from which an optimal one will be chosen. Although our folding rules are the same as those used by Studnicka *et al.* (8), our approach is to lay down very simple rules which even allow some impossible structures. Such invalid structures are eliminated by assigning high energies to them. Schematic representations of structures, known as graphs in mathematical theory, are introduced. These representations are virtually identical to those used by Nussinov *et al.* (9). Although not essential to the theory, they help clarify precisely what structures are being considered and are a good way to show how energies of structures are computed.

By convention, we number the nucleotides of an RNA molecule from the 5'end of the molecule, denoting by S_i the i^{th} nucleotide for $1 \leq i \leq N$, where N is the total number of nucleotides. The letter S alone denotes the entire molecule, and S_{ij} denotes the nucleotides from S_i to S_j inclusive, assuming $1 \leq i < j \leq N$. Figure 1 shows the unspecified nucleotides of an RNA molecule laid out equally spaced on a semicircle. The N nucleotides of a molecule are referred to as vertices in this geometric picture. The N-1 arcs of the semicircle between the bases are called exterior edges, and they represent the phosphodiester bonds between consecutive nucleotides. Base pairing is represented by line segments between nucleotides, that is, by chords on the semicircle between two vertices. A chord is called admissible if it connects

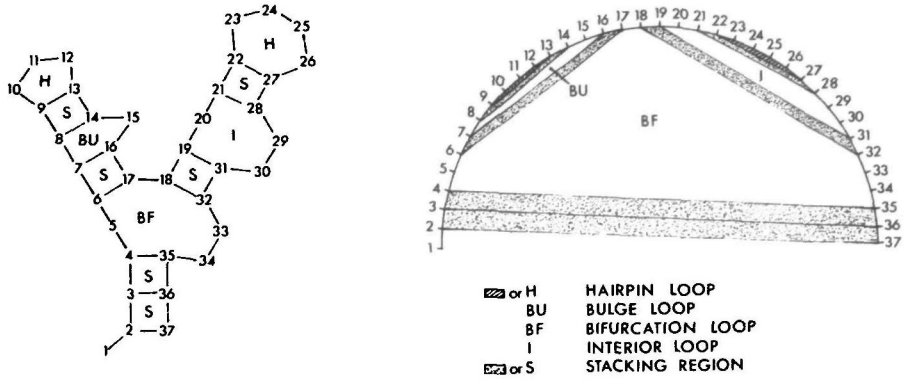

Figure 1 : Two schematic representations of a simple RNA secondary structure. The diagram on the left is conventional while the one on the right is a more abstract representation in terms of a mathematical graph.

two nucleotides which can base pair : G-C, A-U or G-U. These chords are referred to as interior edges. The entire collection of edges and vertices is called a graph. An admissible structure is defined to be a structure whose graph contains only admissible chords which never touch or intersect one another. Not allowing chords to touch is equivalent to saying that a nucleotide can base pair with at most one other nucleotide. Disallowing intersection of two chords is more profound. This condition is precisely what is needed to rule out all knotted structures. Using the nomenclature of Studnicka *et al.* (8), our admissible structures are all 'orthodox', and include all of their 'complex hyperstructures'.

The description is completed with the definition of the free energy of a structure, or, equivalently, of its graph. A face of a graph is defined to be any planar region bounded on all sides by edges. The free energy of a structure is associated not with the bonds, as is done in effect by Nussinov *et al.* (9), but with the regions between bonds. In terms of the graphical representation, the energy depends on the faces of the graph. A face with a single interior edge is called a hairpin loop. Faces with two interior edges are classified into three groups. If the interior edges are separated by single exterior edges on both sides, the face is called a stacking region. If they are separated by a single exterior edge on one side, but by more than one exterior edge on the other side, the face is called a bulge loop.

Otherwise, the face is referred to as an interior loop. Finally, a face with three or more interior edges is called a bifurcation loop. The word hairpin refers to a structure or substructure whose faces are a sequence of consecutive stacking regions, bulge loops, or interior loops, ending with a hairpin loop. Figure 1 gives an example of a structure and its associated graph. All the different types of regions occur. The substructures formed from nucleotides 6 to 17 and from nucleotides 18 to 32 are both hairpins.

If F is a face, we let E(F) denote its associated free energy. For example, if F is a stacking region between two A-U pairs, E(F) = -1.8 kcal/mole (6). The energy of a structure is the sum of the energies of its faces, and the problem at hand is to find a structure with minimum free energy. The energy function is a powerful tool. Its proper use can rule out invalid or undesired structures. For example, if F is a hairpin loop with fewer than four exterior edges, set $E(F)=\infty$. This will exclude the selection of stearically impossible structures containing hairpin loops of fewer than three bases. Even the more involved rule allowing G-U base pairings only in the interior of a succession of stacking regions can be accomplished through correct use of the energy function. On the other hand, the energy function can be used to favour certain types of structures. This will be discussed later. We use the energies compiled by Salser (6). To date, this is the best information available on the stabilizing effect of stacking regions and the destabilizing effects of various loops. There is still virtually no knowledge on the destabilizing effects of bifurcation loops. We treat them in two different ways. In one version of the algorithm, they all have zero energy, while in another, they are treated as interior loops. The latter algorithm is of course more complex. All the results shown in this paper are produced from a computer program using the second method.

B. The algorithm

We shall now describe in detail how the minimal free energy of a secondary structure is obtained when bifurcation loops are given zero energy. The algorithm is simple and yet extremely powerful. No compromises are made. Not a single possibility is overlooked, and yet the algorithm selects a structure of minimum energy out of a number of structures that can be immense even for a molecule that is as small as a 5S ribosomal RNA. The main mathematical technique is to compute two possibly different energies for each subsequence S_{ij} of a given RNA sequence. For all pairs i,j satisfying $1<i<j<N$, let W(i,j) be the minimum free energy of all possible admissible structures formed from

the subsequence S_{ij}. In addition, let V(i,j) be the minimum free energy of all possible admissible structures formed from S_{ij} in which S_i and S_j base pair with each other. If S_i and S_j cannot base pair, then $V(i,j) = \infty$. The numbers V(i,j) and W(i,j) are computed recursively, first for all pentanucleotide subsequences, and then for all successively larger and larger subsequences of S. Pentanucleotide sequences are very easy to deal with. They form no stable structures, so W(i,j) = 0 if j-i = 4. If S_i and S_j are a G-C or A-U pair in this case of a 3 nucleotide hairpin loop, V(i,j) = + 8.4 or + 8.0 kcal/mole respectively (6). If j-i = d>4, V(i,j) and W(i,j) can be computed in terms of V(i',j') and W(i',j') for various pairs i',j' satisfying j'-i'<d. These numbers will already have been computed. Imagine an admissible structure on S_{ij} with energy V(i,j), assuming that S_i and S_j can base pair. We denote by FH(i,j) the hairpin loop containing the interior edge between S_i and S_j, and by FL(i,j,i',j') the face containing exactly two interior edges, one between S_i and S_j,and the other between $S_{i'}$ and $S_{j'}$ (assuming i<i'<j'<j). The faces denoted by FL are either stacking regions or else bulge or interior loops. The face adjacent to the edge between S_i and S_j is one of three possible types, as illustrated in figure 2A. It has either one, two or more than two interior edges. In the first case, V(i,j) = E(FH(i,j)). In the second case, V(i,j) = E(FL(i,j,i',j')) + V(i',j') for some pair i',j' satisfying i<i'<j'<j. In the last case, V(i,j) = W(i+1,i') + W(i'+1,j-1) for some i' satisfying i + 1<i'<j-2. Here the energy splits into the sum of the energies of two substructures ; hence the word 'bifurcation'. Thus V(i,j) is the minimum of the energies which can be obtained in these three ways, so that we can write V(i,j) = min {E1,E2,E3}, where E1 = E(FH(i,j)),

$$E2 = \min_{i<i'<j'<j} \{E(FL(i,j,i',j')) + V(i',j')\},$$

and

$$E3 = \min_{i+1<i'<j-2} \{W(i+1,i') + W(i'+1,j-1)\}.$$

Now imagine an admissible structure on S_{ij} with energy W(i,j). Again there are three possibilities. As illustrated in figure 2B, either S_i or S_j (or both) do not participate in the structure, or they base pair with each other, or else they both base pair, but not with each other. The first case is trivial. The structure has at least one dangling end, and W(i,j) = W(i+1,j) or W(i,j-1). In the second case, W(i,j) = V(i,j), which has already been computed. The last case is referred to as an open bifurcation because the structure splits into two separate parts with no connection between S_i and S_j. If S_i base pairs with $S_{i'}$ and S_j base pairs with $S_{j'}$, where i<i'<j'<j,

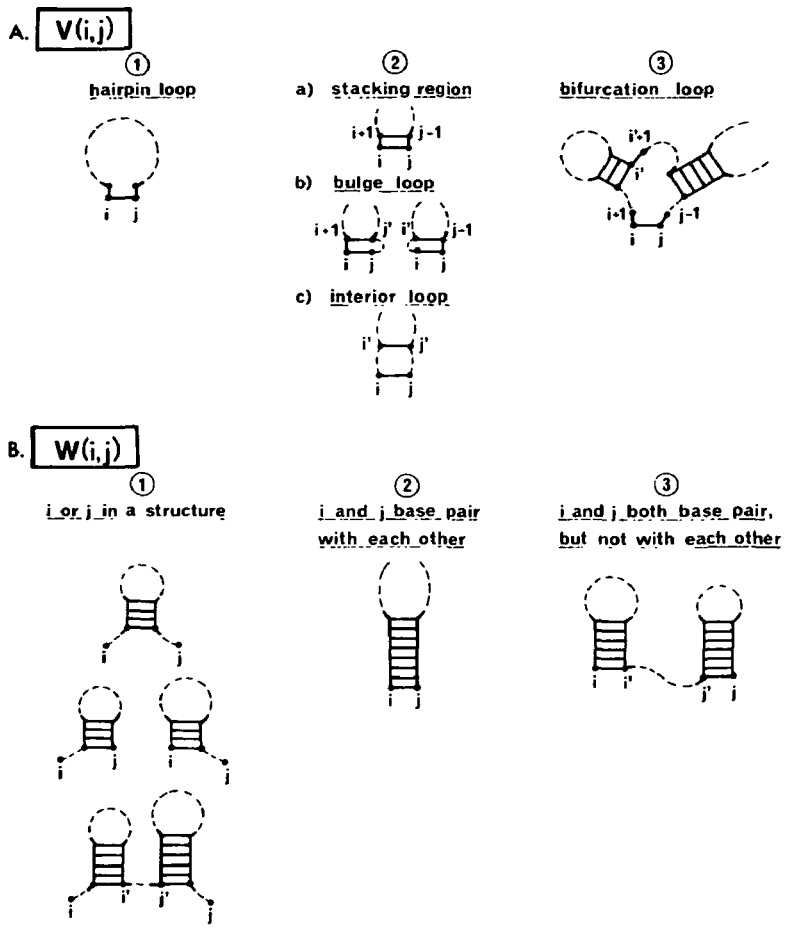

Figure 2 : A : V(i,j) is the minimum free energy of an admissible structure on the subsequence S_{ij} where S_i and S_j base pair with each other.
B : W(i,j) is the minimum free energy of an admissible structure on the subsequence S_{ij}.

then W(i,j) = W(i,i') + W(i'+1,j) = W(i,j'-1) + W(j',j). Thus W(i,j), the minimum energy obtainable from these three cases, is given by :

$$W(i,j) = \min\{W(i+1,j), W(i,j-1), V(i,j), E4\}, \text{ where}$$

$$E4 = \min_{i<i'<j-1} \{W(i,i') + W(i'+1,j)\}.$$

Heuristically, this recursive algorithm works by adding one nucleotide at

a time to a sequence, and observing what the best structure is at each step. The last number to be computed, W(1,n), is the desired answer. It is the minimum energy of an admissible structure on S. However, the labour expended to compute W(1,n) has in fact produced much more, for the minimum energy of an admissible structure on every subsequence of S is also known. All that remains is the construction of the structure, which is equivalent to identifying the interior edges of the associated graph. This is achieved by a traceback through the matrices W and V and is straightforward.

C. Computation techniques

The algorithm described has been programmed in Fortran. Its implementation is on an IBM 3032 processor with an IBM TSS/370 time sharing operating system. The most efficient version of the program stores the energies of the V and W matrices as half integers in the same square array to save space. The energy computation is done in a small subroutine which facilitates changing the rules and creating special versions of the program.

The energy function plays a wider role than simply defining energies from thermodynamic studies. We have already indicated that it is used to enforce certain topological and folding rules such as the minimum size of hairpin loops and the allowable positions of G-U base pairings. Prohibited base pairings are ruled out by giving very large positive energies to adjacent faces. Similarly, some base pairings can be forced by assigning very large negative energies to adjacent faces. In between these two extremes, some base pairings can be favoured by the use of small 'bonus' energies. An example of this is given in the discussion of the folding of *E.coli* 16S ribosomal RNA, where we show how certain phylogenetic data were worked into the algorithm. The energy function can also be used to incorporate additional information on chemical modification and enzyme accessibility into the algorithm. In the case of *E.coli* 16S ribosomal RNA, information on chemical reactivity came from kethoxal modification of single stranded guanylic residues (12,13). Enzyme accessibility data were compiled from topographical studies on 16S ribosomal RNA in either protein-RNA complexes (14,16) or 30S ribosomal subunits (P. Stiegler, P. Carbon and C. Ehresmann, unpublished results). Any guanylic residue which could be kethoxal modified was not allowed to base pair at all. Enzymatic data were not so simple to use. T_1 ribonuclease cuts the phosphodiester bond at the 3'end of a guanylic residue. Pancreatic ribonuclease acts similarly on cytidylic and uridylic residues. In either case, under conditions of partial hydrolysis, only some of these

bonds are cut, indicating single stranded regions. The instructions given to the computer program are that the recognized nucleotide and the next one (toward the 3'end) cannot base pair simultaneously. In practice, stacking regions containing both these nucleotides are given very large positive energies. We thus allow for an accessible bond at the end of a base pairing region. This information is built right into the energy function and no time is wasted computing energies of undesired structures.

Several special versions of the program have been written. One does not allow bifurcation loops and computes an optimal structure which is a succession of hairpins. Such structures are called open, and they are useful in identifying local structures. This program is extremely fast. Another version allows for designated subsequences to be excised from the sequence and the remaining nucleotides to be folded. This is used when we are already confident of the structure of certain subsequences and are searching for long range interactions.

There are two kinds of output. One kind is an actual computer drawing of the structure. An example is given later on in figure 5. The program also produces a detailed output listing all collections of consecutive stacking regions and all destabilizing loops, with associated negative and positive energies.

RESULTS AND DISCUSSION

A. Folding 459 nucleotides from an immunoglobulin γ1 heavy chain messenger RNA fragment.

To illustrate the power of our algorithm and its associated computer program, we folded the same immunoglobulin γ1 heavy chain messenger RNA fragment as did Rogers *et al.* (11) using a slightly modified version of the program of Studnicka *et al.* (8). The folding of this 459 nucleotide fragment is illustrated in figure 3. Computing was performed taking into account the destabilizing effect of the bifurcation loops (see methods). Our computed structure has a free energy of -181.4 kcal/mole, a 15 % improvement over the minimum free energy found by Rogers *et al.* (11) (-158.5 kcal/mole) using the same energy calculations. In addition, our procedure is fast and easy to use. It takes only 134 seconds to compute the most stable secondary structure for this 459 residue fragment. There is a single computing cycle that requires only the nucleotide sequence as input. Thus, many terms required by Studnicka *et al.*, such as 'group number', 'branch migration', and 'primary region', do

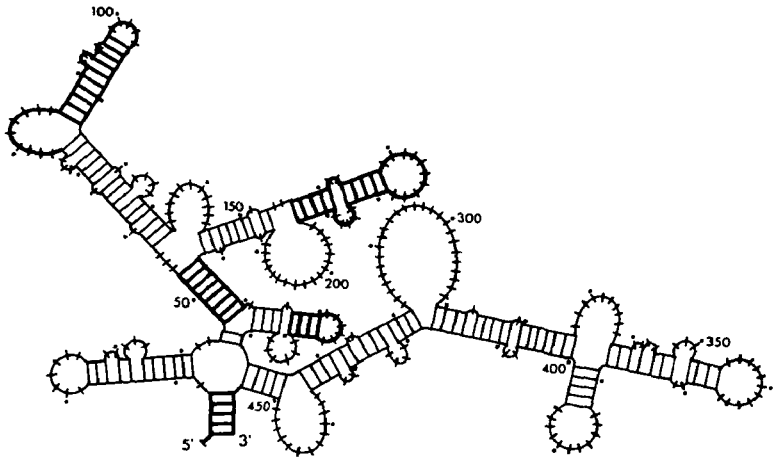

Figure 3 : Folding of an immunoglobulin γ1 heavy chain mRNA fragment. The heavy lines indicate base pairings which coincide with those found by Rogers *et al.* (11). Our folding of this 459 nucleotide fragment found a structure with a 15 % energy improvement (-181.4 kcal/mole, vs. -158.5 kcal/mole for Rogers *et al.*) in 134 seconds.

not occur in our simpler yet more powerful algorithm. No compromises are made to save time ; they are not needed. The 15 % improvement in minimum free energy is therefore not surprising, since we solved the problem precisely as stated. Most noteworthy is the almost complete lack of similarity between our computed structure and the best folding proposed by Rogers *et al.* (11). In figure 3, the base pairing regions which occur in both proposed foldings are shown by heavier lines. Only 29 % of the base pairings in our structure can be found in the corresponding one by Rogers *et al.* We do not claim that the structure we have produced is 'better' than theirs. This example was given mainly to illustrate the power of our method. The deeper question raised by it is how to choose one proposed folding over an other.

B. Suboptimality and additional information.

The words 'optimal structure' refer to a folding of minimum free energy. Such a structure is not necessarily unique. With molecules the size of 5S ribosomal RNA's or greater, and with the rather intricate energy function used, computing energies to the tenth of a kcal/mole, it is unlikely to have two or more optimal structures of exactly the same energy, but it can happen.

It is possible to design a traceback procedure to look for all solutions of minimum energy, but we have not done so. It would be futile. For a sequence of 200 nucleotides, there could be many structures within five or ten percent of the minimum free energy, even when trivial substructures are eliminated. The energy calculations, extrapolated from studies done on oligoribonucleotides and polyribonucleotides, cannot be considered precise enough to discriminate between various structures close to the minimum energy. It is not meaningful to exhibit a list of several suboptimal structures that are produced arbitrarily by the use of a particular algorithm. Why show these structures and not others ? Additional information must be supplied before meaningful alternate structures can be computed. In our studies on *E.coli* 16S ribosomal RNA, such additional information is derived from data on nucleotides that are reactive to a specific chemical reagent or susceptible to ribonuclease attack (see methods). These nucleotides are therefore not involved in base pairing or are located at the end of base pairing regions. Our strategy is to incorporate such information directly into the algorithm by proper use of the energy function (see methods). Improved structures are produced by folding a molecule taking into account auxiliary information. The predicted structures are then carefully examined, and biological evaluation is needed to select those features which will be included in a final model.

In the case of the immunoglobulin heavy chain messenger RNA fragment mentioned above, we have no information other than the nucleotide sequence, and present our structure as a tentative improvement over the folding proposed by Rogers *et al.* However, it is clear that there are many more different structures for this sequence with energies close to -181.4 kcal/mole which might be biologically more meaningful.

C. Application to *E.coli* 16S ribosomal RNA

The main application of the computer program described in this paper has been to facilitate studies on the secondary structure of the *E.coli* 16S ribosomal RNA. Experimental data have accumulated on the folding of this RNA chain and on its relevant structural organization within the small subunit of the ribosome (12-18 and P. Stiegler, P. Carbon, C. Ehresmann, unpublished results). These topographical results provide biological information that is needed for correct model building.

Folding all 1542 nucleotides of the 16S ribosomal RNA at once is beyond the capability of our computer program which can fold a molecule with up to 800 nucleotides. For this reason, we chose to subdivide the 16S ribosomal RNA

sequence, not arbitrarily, but according to topographical studies that delineate distinct structural domains in the molecule. As an example, we present here the computer folding of the first 574 nucleotides. This section forms the RNA interaction site for the ribosomal protein S4, which is able to bind individually to the 16S ribosomal RNA (14,15). Topographical investigation using enzymes as a probe has also shown that this domain preexists in the RNA molecule even in the absence of the protein (14,16).

The 574 nucleotide RNA fragment was folded using our most sophisticated program that takes into account all chemical and enzymatic information available for this sequence (12,13,16). The result is displayed schematically in figure 4. The arrows point to accessibilities indicated by either chemical modification or enzymatic data. This is the most stable structure that can be generated that is consistent with the auxiliary information. Nevertheless, the computed structure lacks 'biological reality'. It does not completely

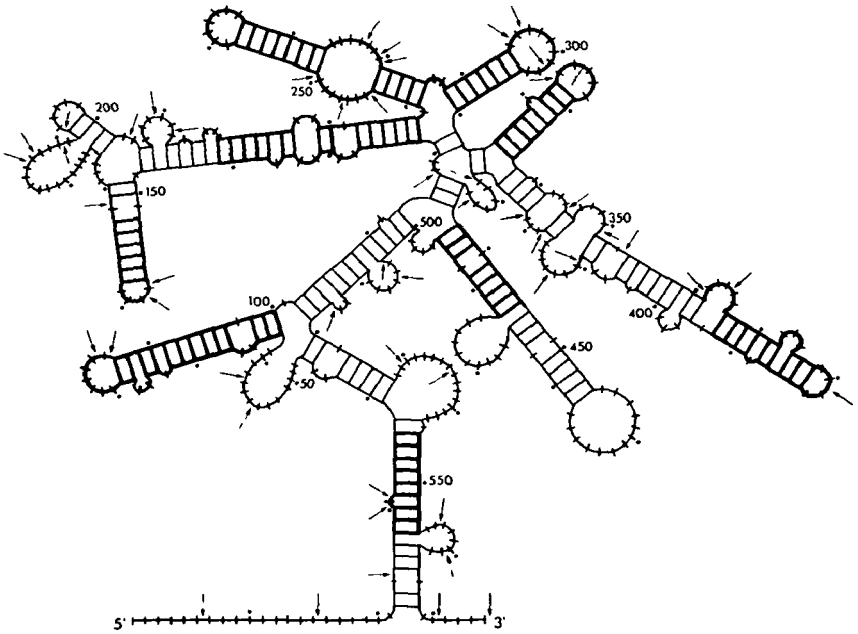

Figure 4 : This is an optimal folding of the first 574 nucleotides of *E.coli* 16S rRNA(19) taking into account chemical and enzymatic evidence. It has a free energy of -203.7 kcal/mole. The arrows point to the 3'end of nucleotides indicated as accessible either by kethoxal modification or enzymatic accessibility studies. Heavier lines indicate the base pairing regions essentially preserved in the final model for the 16S rRNA secondary structure.

satisfy experimental observations derived from detailed structural studies on this RNA domain that could not be sufficiently quantified to be used by the program directly. Even so, there are many structural motifs that can be selected from this rather complicated structure. Indeed, about half of all the indicated base pairings in figure 4 have been preserved in a tentative model for the 16S ribosomal RNA secondary structure (P. Stiegler, P. Carbon, M. Zuker, J.P. Ebel, C. Ehresmann, manuscript in preparation). Refinements were performed using specialized programs derived from the basic one, as described in the methods section. One task was to search for local structures. Such an investigation is justified if one considers that a ribonucleic acid molecule is synthesized sequentially from the 5'terminus to the 3'end *in vivo*. Sequential synthesis may therefore ensure correct folding of the molecule as suggested by a recent study on transfer RNA (20). The search for local structures on the first 574 nucleotides or on subsequences was accomplished by finding an optimal open structure. The size of each of the hairpins was arbitrarily limited to force the computer to display a succession of relatively short hairpins. Another version of the program was used to evaluate long range interaction possibilities after 'excising' local structures already selected with high confidence. In figure 4, the areas drawn by heavier lines are those local structures and long range interactions that were selected for the tentative model of the 16S ribosomal RNA secondary structure. Some of these selected structures can also be found in two recently proposed secondary structures for *E.coli* 16S ribosomal RNA (21,22). Other base paired regions displayed by the computer model were rejected because of specific evidence from topographical studies on the protein S4-RNA binding site. For example, nucleotides 567 to 574 are thought to interact with sequences located in the centre of 16S ribosomal RNA (nucleotides 818 to 897) and can therefore not base pair with nucleotides 22 to 27.

One important criterion for deciding whether or not an indicated base pairing region is valid is to check if the secondary structure of the region is preserved in other 16S ribosomal RNA molecules of related or unrelated species or organisms (21,23). Often, compensatory base changes occur in pairs so that base pairing is preserved, thereby permitting conservation of an identical or similar secondary structure motif. For example, the long range base pairing region including nucleotides 122 to 141 and 220 to 237 (see figure 4) is preserved through compensatory base changes in *Proteus vulgaris* (P. Carbon and C. Ehresmann, personal communication). Therefore this structure of very low free energy (-30.5 kcal/mole) was included in the final model of

16S ribosomal RNA. An attempt was made to incorporate some phylogenetic information directly into the program. We refolded the first 574 nucleotides of *E.coli* 16S ribosomal RNA giving an energy bonus of -3.0 kcal/mole to each base pairing which was preserved in the closely related 16S ribosomal RNA sequence of *Zea mays* chloroplast (23) by a compensatory change. The resulting computer-generated model exhibited some additional coherent structures but other regions from figure 4 were lost. Clearly a more sophisticated approach is needed here. Nevertheless, this example underlines how easily the algorithm adapts to specialized searches.

Figure 5 shows the predicted structure for nucleotides 571 to 765 of *E.coli* 16S ribosomal RNA. About 80 % of the indicated base pairings survived close scrutiny and were preserved for the final model. This figure shows the form of the computer output and contains examples of base pairings of nucleotides designated as 'accessible' by enzymatic studies.

D. Conclusions

On the purely technical side, there is not much room for improvement. The computation time of our algorithm is proportional to the cube of the number of nucleotides, and this performance cannot be improved further except perhaps for a reduction in the proportionality constant. Available computer storage limits our algorithm to folding at most 800 nucleotides. The program that ignores the destabilizing effect of bifurcation loops could be

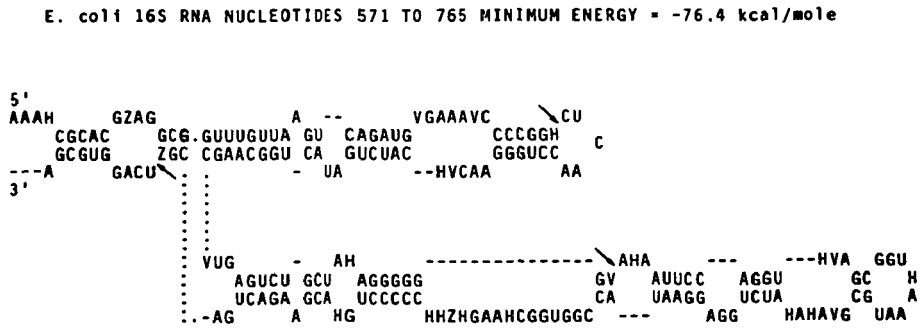

Figure 5 : This is a typewritten reproduction of the actual computer output. The letters 'Z', 'H', and 'V' designate nucleotides 'C', 'G' and 'U' respectively that are known to be accessible from chemical or enzymatic evidence. Arrows point to those phosphodiester bonds that are susceptible to enzymatic hydrolysis, although adjacent to base pairing regions. Only the dots indicating the bifurcation loop have been added to the computer output.

rewritten more efficiently to fold up to 1000 nucleotides. However, as we have just seen, there are problems with folding just 600 nucleotides. The folding of the first 574 nucleotides of *E.coli* 16S ribosomal RNA is no problem technically. It is even affordable, taking only 244 seconds on our computer. Nevertheless, the resulting structure is not fully acceptable to us despite the fact that it has a minimum free energy and conforms to all evidence on chemical reactivity and enzymatic susceptibility. We do not agree with those who would abandon all such complicated folding algorithms and use only the most rudimentary computer searches for base pairing in their investigations. Our method has proved to be a valuable and flexible tool in investigating the secondary structure of *E.coli* 16S ribosomal RNA. Much tedious work was eliminated as we quickly and easily tested numerous folding possibilities for rather long fragments or relevant subsequences.

The first area where improvement is needed is in the thermodynamic computations. Recently, Ninio (24) has taken a step towards more meaningful energy computations. His thermodynamic rules have been designed to make the cloverleaf model the most energetically favourable for transfer RNA's. This was done because the rules compiled by Salser (6) are inadequate to predict the cloverleaf in about half of all known tRNA's. Whether or not Ninio's rules are more meaningful when applied to larger RNA sequences is an open question. In any event, we do not consider the thermodynamic rules we use to be wrong, but merely inadequate. They form a necessary basis for folding a molecule, but cannot be expected to be sufficient in themselves. A program based solely on conformational rules and thermodynamics will not yield a biologically meaningful folding of a molecule on its own. There are too many different structures with similar energies. More and different kinds of additional information must be incorporated into the algorithm as well. With regard to 16S ribosomal RNA's, it would be useful to find an effective way to bring phylogenetic comparisons with related RNA's into the optimization procedure.

ACKNOWLEDGEMENTS

We wish to thank Dr. C. Ehresmann for stimulating discussion on RNA folding and for communicating unpublished results. P.S. thanks NATO for a long term fellowship.

†NRCC number 18755

Technical details on the computer program will be made available from the first author to whom reprints should also be requested.

+Present address :
Laboratoire de Biochimie, Institut de Biologie Moléculaire et Cellulaire du CNRS, 15, rue René Descartes, 67084 STRASBOURG Cédex - FRANCE.

REFERENCES

1. TINOCO, I., UHLENBECK, O.C. and LEVINE, M.D. (1971) Nature 230, 362-367
2. UHLENBECK, O.C., BORER, P.N., DENGLER, B. and TINOCO, I. (1973) J. Mol. Biol. 73, 483-496
3. GRALLA, J. and CROTHERS, D.M. (1973) J. Mol. Biol. 73, 497-511
4. GRALLA, J. and CROTHERS, D.M. (1973) J. Mol. Biol. 78, 301-319
5. TINOCO, I. BORER, P.N., DENGLER, B., LEVINE, M.D., UHLENBECK, O.C., CROTHERS, D.M. and GRALLA, J. (1973) Nature New Biol. 246, 40-41
6. SALSER, W. (1977) Cold Spring Harbor Symp. Quant. Biol. 42, 985-1002
7. PIPAS, J.M. and McMAHON, J.E. (1975) Proc. Nat. Acad. Sci. USA 72, 2017-2021
8. STUDNICKA, G.M., RAHN, G.M., CUMMINGS, I.W. and SALSER, W.A. (1978) Nucleic Acids Res. 5, 3365-3387
9. NUSSINOV, R., PIECZENIK, G., GRIGGS, J.R. and KLEITMAN, D.J. (1978) SIAM J. Appl. Math. 35, 68-82
10. WATERMAN, M.S. and SMITH, T.F. (1978) Mathematical Biosciences 42, 257-266
11. ROGERS, J., CLARKE, P. and SALSER, W. (1979) Nucleic Acids Res. 6, 3305-3321
12. CHAPMAN, N.M. and NOLLER, H.F. (1977) J. Mol. Biol. 109, 131-149
13. HERR, W., CHAPMAN, N.M. and NOLLER, H. (1979) J. Mol. Biol. 130, 433-449
14. EHRESMANN, C., STIEGLER, P., CARBON, P., UNGEWICKELL, E. and GARRET, R.A. (1977) FEBS Lett. 81, 188-192
15. UNGEWICKELL, E., GARRETT, R.A., EHRESMANN, C., STIEGLER, P. and CARBON, P. (1977) FEBS Lett. 81, 193-198
16. EHRESMANN, C., STIEGLER, P., CARBON, P., UNGEWICKELL, E. and GARRETT, R.A. (1980) Eur. J. Biochem. 103, 439-446
17. UNGEWICKELL, E., EHRESMANN, C., STIEGLER, P. and GARRETT, R.A. (1975) Nucleic Acids Res. 2, 1867-1888
18. RINKE, J., YUKI, A. and BRIMACOMBE, R. (1976) Eur. J. Biochem. 64, 77-89
19. CARBON, P., EHRESMANN, C., EHRESMANN, B. and EBEL, J.P. (1979) Eur. J. Biochem. 100, 399-410
20. BOYLE, J., ROBILLARD, G.T. and KIM, S-H. (1980) J. Mol. Biol. 139, 601-625
21. WOESE, C.R., MAGRUM, L.J., GUPTA, R., SIEGEL, R.B., STAHL, D.A., KOP, J., CRAWFORD, N., BROSIUS, J., GUTTEL, R., HOGAN, J.J. and NOLLER, H.F. (1980) Nucleic Acids Res. 8, 2275-2293
22. GLOTZ, C. and BRIMACOMBE, R. (1980) Nucleic Acids Res. 8, 2377-2395
23. SCHWARZ, Zs and KÖSSEL, H. (1980) Nature 283, 739-742
24. NINIO, J. (1979) Biochimie 61, 1133-1150.