Prediction of RNA secondary structure, including pseudoknotting, by computer simulation

Jan Pieter Abrahams[1,2], Mirjam van den Berg[2], Eke van Batenburg[2] and Cornelis Pleij[1]
[1]Department of Biochemistry, Gorlaeus Laboratories, University of Leiden, PO Box 9502, 2300 RA Leiden and [2]Institute for Theoretical Biology, University of Leiden, PO Box 9516, 2300 RA Leiden, The Netherlands

Received December 11, 1988; Revised and Accepted April 17, 1990

ABSTRACT

A computer program is presented which determines the secondary structure of linear RNA molecules by simulating a hypothetical process of folding. This process implies the concept of 'nucleation centres', regions in RNA which locally trigger the folding. During the simulation, the RNA is allowed to fold into pseudoknotted structures, unlike all other programs predicting RNA secondary structure. The simulation uses published, experimentally determined free energy values for nearest neighbour base pair stackings and loop regions, except for new extrapolated values for loops larger than seven nucleotides. The free energy value for a loop arising from pseudoknot formation is set to a single, estimated value of 4.2 kcal/mole. Especially in the case of long RNA sequences, our program appears superior to other secondary structure predicting programs described so far, as tests on tRNAs, the LSU intron of *Tetrahymena thermophila* and a number of plant viral RNAs show. In addition, pseudoknotted structures are often predicted successfully. The program is written in mainframe APL and is adapted to run on IBM compatible PCs, Atari ST and Macintosh personal computers. On an 8 MHz 8088 standard PC without coprocessor, using STSC APL, it folds a sequence of 700 nucleotides in one and a half hour.

INTRODUCTION

The folding of a single stranded RNA molecule is, at least for a great part, determined by its nucleotide sequence. The formation of base pairs like A-U, G-C and G-U gives rise to specific structural motifs like double helical or stem regions and single stranded regions like hairpin-, bulge-, multibranched- and interior loops (1–3). The ensemble of these structural elements in a planar presentation is called the secondary structure of RNA. As soon as a set of free energy values for the various structural elements was available (4), attempts were made to predict the most stable secondary structure for a given RNA sequence by searching for a minimum free energy. Several algorithms have been developed for this purpose and they have become an additional tool in the determination of the structure of RNA. For recent reviews see (3) and (5).

Most of the programs, including the most widely used one developed by Zuker and Stiegler (1), generate this most stable structure by an exhaustive search through all possible structures. Recently, extensions for some of these programs were published which produce a number of suboptimal foldings (6–9). An exception to this approach is the algorithm of sequential stem selection proposed by Martinez (10). It uses the simple rule that the next stem chosen to be added to a growing structure is the one having the largest association constant.

Studies on plant viral RNAs in our laboratory have led to the discovery of a new RNA folding principle which is based on so-called pseudoknotting (11). One specific type of pseudoknot was shown to occur often in a number of RNAs. It involves basepairing of a stem-adjacent region from a hairpin loop with a complementary sequence outside this hairpin (Fig. 1). For more details about this type of pseudoknot in viral, messenger or ribosomal RNAs, the reader is referred to (11) and (12).

Thus far, pseudoknots have not been allowed in programs for RNA structure prediction. The main reasons were that these structural elements were not known to occur in natural RNAs (13), and that the algorithm was thought to become too complicated (2). Another problem was, and still is, that no thermodynamic parameters are available for some essential parts of pseudoknotted structures; see also (3).

In this paper we describe a program which is able to predict pseudoknotted structures together with an otherwise orthodox secondary structure. This program has been used and tested for a number of years on a variety of natural RNAs (14–16,25) and has been gradually improved. It was applied successfully in the case of the 5′ noncoding region of a picornaviral RNA where it predicted pseudoknots not observed before (14).

Here we will present the results of a large number of foldings of various RNA molecules, some of which harbour experimentally proven pseudoknots or pseudoknots which are strongly supported by sequence comparisons. A great number of these pseudoknots are predicted with our program, despite the lack of experimentally determined energy values for the connecting loops involved. The performance of our program in terms of computer time needed and the number of correct base pairs predicted is superior to that of the most widely used one, described by Zuker and Stiegler (1), especially in the case of long RNAs.

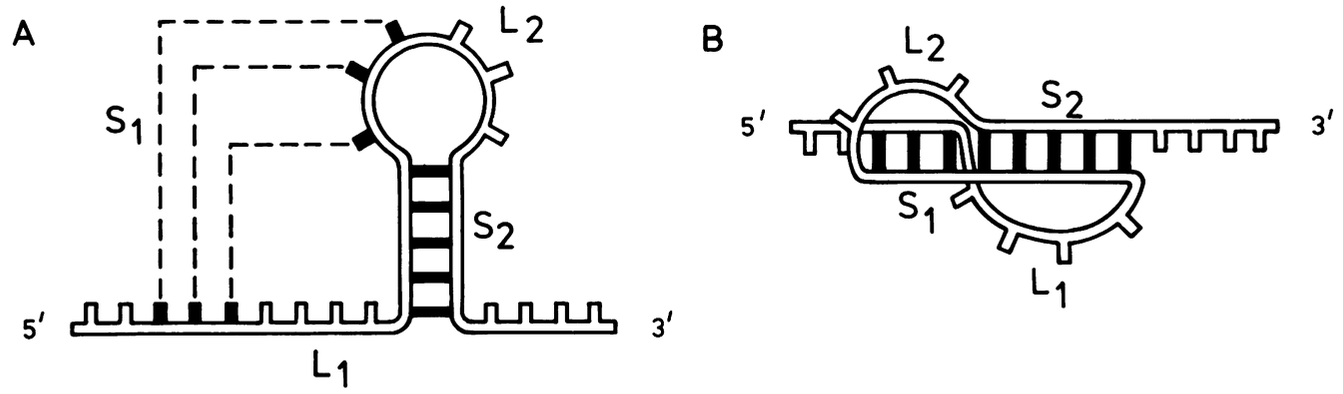

Fig. 1. The principle of pseudoknot formation in RNA. (A) Nucleotides from the hairpin loop basepair with a complementary sequence outside this loop. (B) Formation of a quasi-continuous double helix by coaxial stacking of the two double helical segments. The black bars represent base pairs. The nomenclature for the stems (S1 and S2) and connecting loops (L1 and L2) is derived from (11). For further details about this so-called H-type pseudoknot see (12).

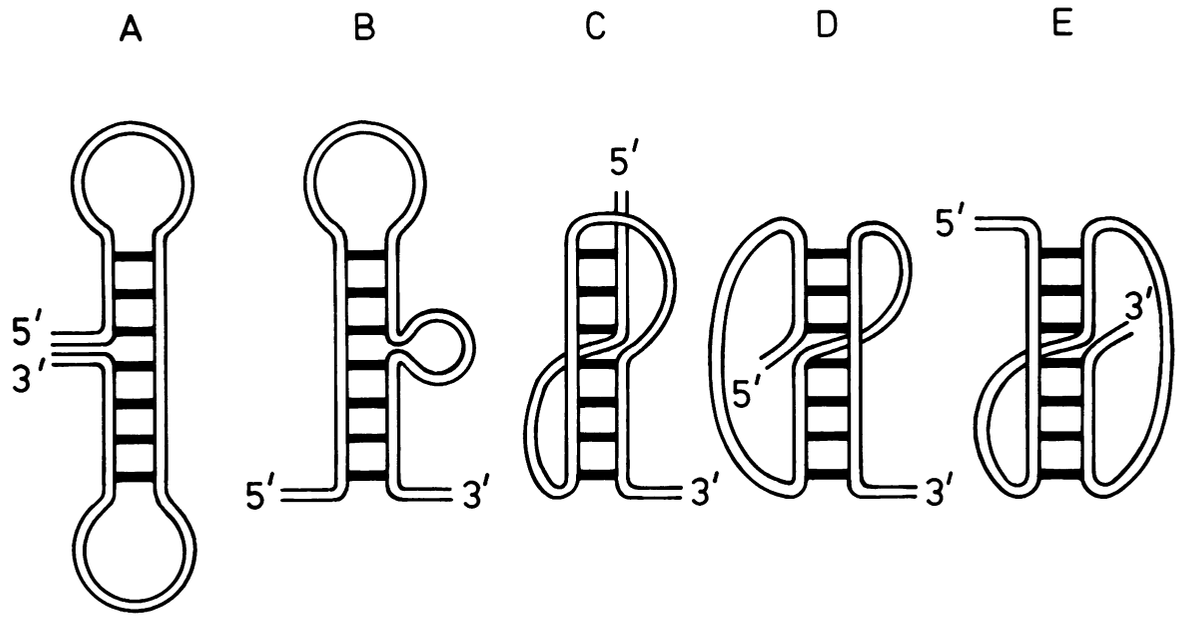

Fig. 2. Five possible ways to stack coaxially two different stem segments. (A) Bifurcated structure or fork. (B) Bulge loop. (C) H-type pseudoknot (see also Fig. 1). (D) and (E) Two novel pseudoknotted structures described in the text. The black bars represent base pairs.

SYSTEM AND METHODS

First, we will describe how we simulated the folding of RNA, next we will show how we included some types of tertiary interactions in the process and then we will describe how we dealt with long distance interactions. Finally, we will give specific details on the algorithm used, on the implementation and on the availability of the program.

Folding of RNA

The final shape of an RNA molecule emerges via intricate kinetic pathways; it is predetermined by the rates of formation and disruption of all of its conformational elements in all their combinations. Rather than base pairs, stems can be considered as the conformational elements of the secondary structure of RNA. This is allowed because the initiation is the rate limiting step of stem formation: once a few bases of a stem pair with each other, the rest quickly follows (17). We consider the formation of the secondary structure as a stepwise process, where intermediate structures evolve into the native one by subsequent addition of stems. Of all stems possible, only a few can be incorporated into the final structure. Those stems forming fastest and lasting longest are preferred.

The distance between two stemhalves can be measured by counting the number of nucleotides in the connecting loop which results upon the formation of a stem. This distance determines the amount of free energy needed to bring the two stemhalves in close proximity. Clearly this energy is an important determinant of the rate of stem formation.

Once the stemhalves are in close proximity, their speed of association is determined by the efficiency of nucleation (initial base pair formation). After the stem is formed, it should be stable by itself, or last long enough to be stabilized by the formation of another stem. The amount of free energy released by hydrogen

Table 1. Energy values of the connecting loops of pseudoknots of the H-type

	1	2	15	16	
2	99.9	99.9	99.9	99.9	Energy of the connecting loop crossing the deep groove of the RNA-helix (kcal/mole).
3	99.9	4.2	4.2	99.9	
4	4.2	4.2	4.2	99.9	
9	4.2	4.2	4.2	99.9	
10	99.9	99.9	99.9	99.9	

	1	2	3	15	16	
1	99.9	99.9	99.9	99.9	99.9	Idem for the shallow groove of the RNA-helix.
2	99.9	99.9	4.2	4.2	99.9	
3	99.9	4.2	4.2	4.2	99.9	
6	99.9	4.2	4.2	4.2	99.9	
7	99.9	99.9	99.9	99.9	99.9	

The first number of each row denotes the number of nucleotides of the loop crossing a groove, the first number of each column denotes the number of base pairs being crossed by the loop (see also Fig. 1). For economic reasons columns and rows are omitted, if all the energy rules are equal to those of neighbouring ones.

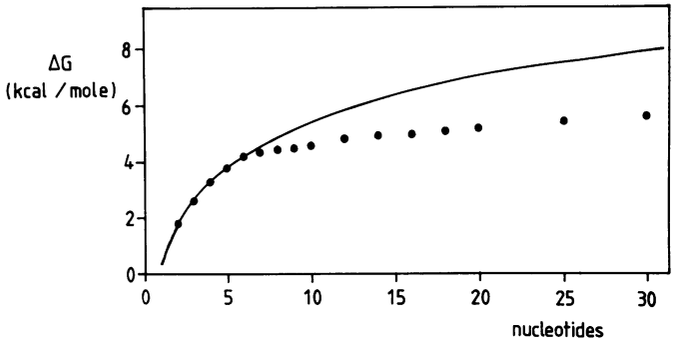

Fig. 3. The free energy of closing an internal loop with A-U/A-U base pairs is plotted versus its length in nucleotides. The values denoted by the filled symbols were taken from (18). The drawn line is a logarithmic extrapolation of the first five points.

bonding and base pair stacking resulting from the formation of a stem, determines the efficiency of nucleation and the stability of that stem.

The rate of formation of a stem thus depends on two types of free energy: first the energy-uptake for closing the loop and next the energy-release after pairing and stacking the bases. Energy values of various kinds of loops and of stacks of base pairs have been published and are used by almost all RNA secondary structure predicting programs (5, 18). It is not entirely clear how these free energies, whose sum is a measure of the overall stability of the stem, should be used to calculate the rate of stem formation. In practice we solved this problem by simply adding the free energy of closing a loop to that of stacking and pairing. As an option, the first can be multiplied with a weighting factor. We assume that the resulting value is more or less proportional to the rate of stem formation. Therefore, the stem with the lowest value is chosen to be incorporated into the nascent structure.

In some cases, we could raise the success of our program considerably by multiplying the free energy of loop formation with a weighting factor prior to adding it to the free energy of basepairing, thus penalizing long distance interactions. This weighting factor is associated with the speed of folding. Up to now, we have not carried out this procedure systematically, so that all the secondary structures presented in this paper are

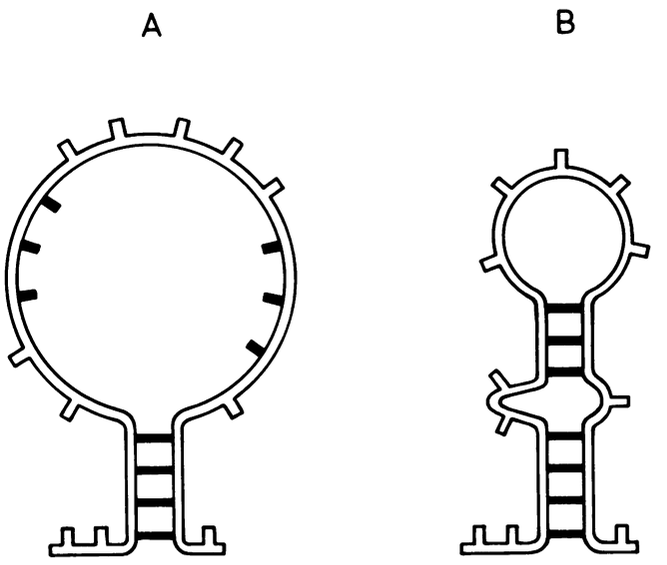

Fig. 4. The free energy contribution of a large loop to an RNA structure can be diminished by dividing the loop into two smaller ones plus a stacked region. Note that it is possible that the upper stem in B is not stable enough by itself and that it needs the bottom stem for its existence. Of each stem which is incorporated into the structure, the program calculates the free energy of the loop which it encloses, the free energy of the stacked base pairs, and the free energy of the loop enclosed by the stem in the structure beneath it. Of each stem in the structure, only the free energy of its stacked base pairs and its enclosed loop is stored. To calculate the free energy change of the structure resulting from the introduction of a new stem in an already existing loop, the energy of this loop is substituted by the energy of the loop beneath the newly formed stem. The energy of the loop enclosed by the new stem and the energy of its stacked base pairs are not changed.

calculated without this feature. In practice this means that each incorporated stem is the one adding most to the stability of the structure at that point.

Implicit to the procedure described above is that stem formation can be cooperative: formation of one stem can bring two distant stemhalves of another stem nearer to each other, thus facilitating their pairing.

Tertiary interactions

Our program is capable of dealing with interactions which are formed upon basepairing of a loop with a single stranded stretch elsewhere. Because the energy of stacking stems upon each other is important for the formation of the final structure, we tabulated all possible ways in which this stacking can be reached. We found five different ways, represented in Fig. 2: fork, bulge, pseudoknot of the H-type and two new types of pseudoknots. Other programs deal with such stacking interactions only in the case of bulges (Fig. 2B). The pseudoknot in Fig. 2C is the type of interaction which is reviewed elsewhere (11, 12). Note the two new types of pseudoknots depicted in Fig. 2D and 2E, which have not yet been reported to exist in natural RNA at the moment of writing. In these two types of pseudoknots three different domains can be distinguished: the quasi-continuous helical region, one single stranded stretch which does not cross any groove, and one crossing the shallow groove in the case of the knot in Fig. 2D, or crossing the deep groove in the other case (Fig. 2E). Details of these new types of pseudoknots will be published elsewhere (Mans *et al.*, manuscript in preparation).

It is important to realise that in these knot-like structures, three

different types of single stranded stretches can be distinguished. The energy values which the program assigns to these stretches depends on (a) the number of bases crossing a groove, (b) the number of base pairs to be crossed, and (c) the type of groove being crossed (deep, shallow or none). These values are listed in Table 1. They were based on an analysis of experimentally proven pseudoknots. For instance, we found a pseudoknot which can only exist, if the energy-values for its connecting loops are 4.2 kcal/mole or less. We therefore introduced this value for all H-type pseudoknots, whose connecting loops are shorter than fifteen nucleotides. Other pseudoknots were prohibited by assigning a value of 99.9 kcal/mole to the energy of the single stranded stretches crossing both grooves. We realise that this length limit and single energy value are oversimplifications, but we nevertheless got good results with these provisional values. Note that these values can be changed at will.

Because only very little is known about other types of tertiary interactions and pseudoknots in RNA, the program assigns one standard positive free energy value to their single stranded regions. In the examples given in this paper, this value was chosen so large that this meant in practice that tertiary interactions other than the above-mentioned pseudoknots were excluded.

Stems giving rise to pseudoknots are treated exactly as other stems: for each stem the amount of stability added to the structure, should the stem be incorporated, is determined. Of course when doing this, the kind of tertiary interaction introduced is taken into account.

Long distance interactions

When we used higher free energy values than in (4,18,20) for loops longer than seven nucleotides, we could raise the success of our predictions, especially when long RNA sequences were involved. No systematic experimental studies have been performed on RNA fragments with loops of this size, except for a recent study of hairpin loops with nine nucleotides as the upper limit (21).

Thus for these free energy values were estimated by extrapolation using the theory of Jacobson and Stockmayer (22). Once the free energy needed to join the ends of an unrestricted, zero-volume polymer is known, this theory predicts the free energy needed to form a similar, but larger loop. Because of lack of data, RNA loops consisting of only seven nucleotides were used as a starting point of this extrapolation (4, 5, 20). The apparent differences in the multiplication factor of the Jacobson-Stockmayer equation used by the various authors, undoubtedly reflects their unease in assuming that such a short piece of RNA behaves like an unrestricted, zero-volume polymer. The number of nucleotides an RNA molecule should count before it starts behaving as such a polymer still needs to be determined.

When a plot is made of the experimentally determined loop energies of small internal loops, extrapolation using the Jacobson-Stockmayer relation and starting with the value of a loop of seven nucleotides yields a clear break in the curve (Fig. 3). Since we do not see any physical reason for a sudden change in the relation between the length of a loop and energy content, we used a normal logarithmic extrapolation of the experimentally determined values. The pre-logarithmic multiplication factor we found, was also used to extrapolate the energies of all other loops. As a result, the energy values we used for loops do not deviate much from those published by Turner and co-workers (5).

Thermodynamic parameters

Throughout this paper the loop destabilizing energy values implemented in our program have been estimated (for pseudoknot connecting loops) or extrapolated (hairpin, bulge and interior loops) as indicated above. All other values, in particular those for nearest neighbour stacking energies, were derived from Jacobson *et al.* (18).

We point out that the user of our program can easily change these values at will.

The algorithm

The stems which can be formed by a given primary stucture are stored in two datasets with a comparable design: one includes those stems which are actually incorporated into the (intermediate) structure, the other includes the remaining stems. We will refer to the first dataset as 'structure' and to the second as 'list'.

To fold a primary structure into a secondary structure, we used the following algorithm:

1. Find all possible stems of the given sequence and determine the free energies of their loops and base pairs.
2. Add stems to the structure until the free energy of the structure can not be decreased anymore.
   - 2.1 Determine the stem most likely to be formed next.
     - 2.1.1 Multiply the free energies of the loops created upon formation of each of the stems in the list with a constant factor.
     - 2.1.2 Add these values to the free energies of the corresponding basepaired regions.
     - 2.1.3 Order the list ascendingly.
     - 2.1.4 Return the stem highest on the list which decreases the free energy of the structure.
   - 2.2 If this stem decreases the free energy of the structure: add this stem to the structure and update the stems in the list.
     - 2.2.1 Transfer the selected stem from the list to the structure.
     - 2.2.2 Remove those base pairs from the stems on the list, where one of the bases in the pair is also present in the selected stem and recalculate the stacking energy of the changed stems.
     - 2.2.3 Add the appropriate energy to the stacking energy of those stems in the list which stack on the selected stem.
     - 2.3.4 Recalculate the loop energy of those stems in the list whose loop-length changed because of the inclusion of the selected stem in the structure.
3. Print the structure.

Instead of discussing all the details of the algorithm, we will only point out some of its general features.

- Almost always the number of stems which can be formed in theory rises quadratically with the number of nucleotides in the sequence. When the program analyses long sequences, it drops the least stable stems. The number of stems retained depends on the amount of free computer memory. In the examples given in this paper this number varied between 1000 and 3000 stems.
- All base pair stackings are taken into account, including those between different stems (see Fig. 2).
- Odd pairs are included if the user provides the appropriate stacking energies. The examples presented in this paper used a free energy value of +2.0 kcal/mole for all stacks in which an odd pair other than G-U is involved.
- If two different base pairs could both stack onto a third base pair (for example in a branching loop of length zero), the energetically most favourable stack is chosen.

– When a stem is formed inside the loop of another stem, first the free energy associated with this loop is subtracted from the overall free energy, then the free energy associated with the resulting loops is added to the overall free energy. See also Fig. 4.
– When a pseudoknot is formed, in principle the same procedure is followed: the free energy associated with the existing loop is subtracted from the overall free energy, then the free energies belonging to the newly formed loops are added.

Implementation

The program was developed on an Amdahl V7B mainframe-computer with an APL SV 4.0 interpreter system. After the program was tested, it was transferred to IBM compatible PCs, Atari ST and Macintosh personal computers.

RNA sequences can be entered manually, but the program accepts input from ASCII files as well. Fig. 5 shows a typical

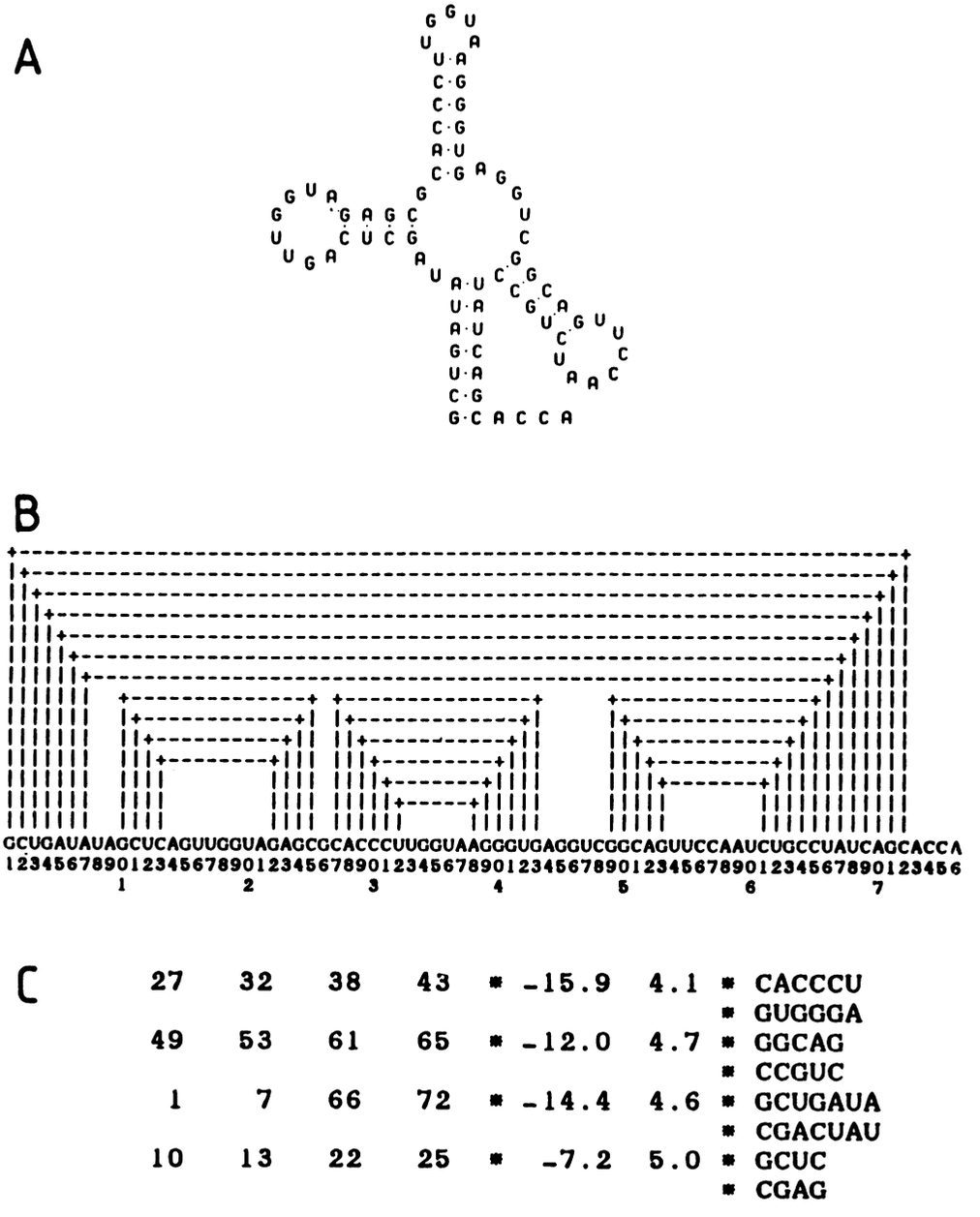

Fig. 5. Representation of RNA secondary structures. (A) The structure of the transcript of the thrT tRNA gene from *E. coli* (see also Fig. 6). (B) Indication of base pairs in a primary structure by connecting complementary regions with lines. (C) Enumeration of stems. The first four columns define the location of a stem, the next two columns show the energy content of the stem and finally the base composition of the stem is given:
– The first column contains the base number of the most 5′ nucleotide of the stem.
– The second column contains the base number of the 3′ end of the 5′ stemhalf.
– The third column contains the base number of the 5′ end of the 3′ stemhalf.
– The fourth column contains the base number of the most 3′ nucleotide of the stem.
– The fifth column contains the Gibbs free energy of the base pairs.
– The sixth column contains the Gibbs free energy of the single stranded stretch of RNA located directly between the corresponding bases of the second and third column. If between these bases another stem can be found, the free energy of the single stranded bases enclosed by this stem is not included.
The order in which the stems are incorporated into the structure, is the same one as in this list.

output of a structure as given by the program. The first type of output (Fig. 5B) helps in visualizing the overall structure, the second type of output (Fig. 5C) contains information on the base-composition, the energy-content and the exact location of the stems and the order in which they are generated.

Availability

A package including the program, an easy users interface and documentation is available from the third author. Academic institutions are entitled to a substantial discount. At the moment versions for Atari ST and MacIntosh computers are available, while a similar package is developed for IBM compatible PC's.

Table 2. Comparison of the secondary structures of tRNA molecules as predicted by various programs

	% correctly predicted		
stem	Zuker&Stiegler	Le *et al.*	our program
aminoacyl acceptor	75	92	71
dihydro U	37	58	40
anticodon	49	73	98
T	68	94	91

For the entries of the Zuker- and Le-programs: see (8). It is unclear from the latter which set of tRNAs was used for the comparison. The tRNAs we folded were selected from (24): we selected every tenth entry of the list, starting with sequence number 10 (A412), giving a total of 45 sequences. We solved the problem of base modification by either changing the modified nucleotide by the normal pairing one or by defining it as a non-pairing nucleotide if it was modified on N_1, N_2 or O_6 of guanosine, on O_3, N_3 or N_4 of cytidine, on N_1 or N_6 of adenosine, or on O_4 or N_3 of uridine. Exceptions to this procedure were: N_4-acetylcytidine (entered as cytidine), N_2-methylguanosine (entered as guanosine) and N_2,N_2-dimethyl-guanosine (entered as adenosine, since it only seems to pair with uridine).

RESULTS

tRNA and 16S rRNA

Almost all tRNA molecules adopt the well-known cloverleaf secondary structure. This group of RNAs is therefore very suitable to test any program predicting secondary or tertiary structures (8,18,23). We have randomly selected 45 tRNA sequences from a published list (24) and have folded them with our program. We found the perfect cloverleaf in 17 cases (38%) as compared to 24% obtained with the Zuker program (18). 26% of our predicted structures harboured three stems of the cloverleaf. When our results are compared with a recent calculation (8), it turns out that our program scores better than the Zuker program (Table 2). For this comparison, the program of Le *et al.* only predicted double stranded regions if they contained at least three base pairs (according to (8)), thereby allowing only the type of stems present in the cloverleaf model of tRNA. This feature could be a (partial) explanation of the good results obtained with this program. In the case of the Zuker and Stiegler algorithm better results could be obtained, but in that case also suboptimal foldings were taken into account or an extension of the algorithm together with improved free energy parameters were used (7,8,25).

The fact that our program can also handle longer sequences successfully is illustrated by the folding of the primary transcript

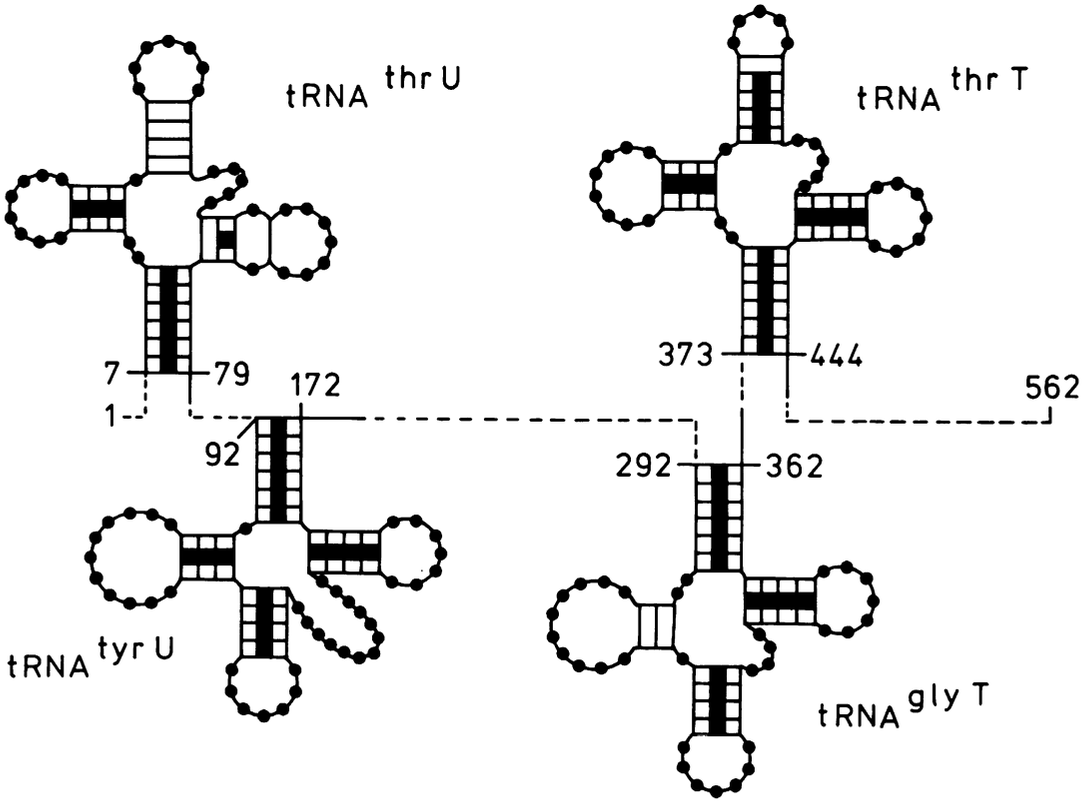

Fig. 6. Secondary structure prediction of the four tRNA genes in the transcript of the tufB operon of *E. coli*. The four tRNAs are presented in the cloverleaf form as derived from a published list (24). The thick black bars represent the stem segments predicted by our program upon folding 562 nucleotides from the 5' noncoding region in one single run. Other parts of the predicted structure, indicated by the dashed lines, have been omitted for the sake of clarity. The black dots indicate base residues in single stranded regions.

of the tufB operon from *E. coli*. This RNA transcript contains four successive tRNA genes upstream of the EF-Tu coding region. The folding of 562 nucleotides from the 5′ noncoding part predicted two perfect cloverleaves and two imperfect ones having three and two stems respectively (Fig. 6). This result is even somewhat better than the one presented in Table 2. We note that one of the three tRNA stems not predicted (the T-stem of $tRNA^{ThrU}$) harbours a C-A mismatch. For further details about the structure of this transcript see (26).

Jaeger *et al.* (25) recently presented results for the separate folding of the four domains of the 16S ribosomal RNA from *E. coli*. Folding of the same four domains with our program produces 29 out of the 97 canonical stems (including single base pairs as a stem). When the same definition for a helix is used as in Jaeger *et al.* (25), our program predicts 26 of the 65 phylogenetically deduced helices (40%). These figures should be compared with the 20% of the helices found in the optimal structure when Salser's rules are used and with the 55% when using the improved free energy parameters of Turner (see Table 2 in Ref. 25). Note that in our program non-base-paired interactions were neglected so far and that the nearest neighbour base pair stacking energies of Salser were used (see System and Methods).

The *Tetrahymena thermophila* LSU intron

The 414 nucleotides long intervening sequence in the ribosomal RNA precursor of *Tetrahymena thermophila* has been used by many authors to test their folding algorithm or free energy parameters (5-7,10,27,28). Here we present the result obtained with our program for the sake of comparison. We note that our predicted structure was obtained without any constraints as deduced from experimental results (27) or by making a choice from a set of suboptimal structures (6,7). The following stems present in the current model (29) were predicted correctly in this order: P2, P9.1, P9 (variant), P6b, P2.1 (bottom part), P8, P9.1a, P5b, P1. Then two stems which are not present in the current model were added, obstructing the formation of P3 and P7. The program continued with stems P5a (upper part), P5c (variant), P5a (bottom part), P5, P9.2, P9.2a, P6a, a base pair not present in the model, P2.1 (top part, variant), P2.1 (middle part), a short stem not present in the model, P5a, three short stems not present in the model and finally one base pair of stem P4. All together, this means that the stems P7, the majority of P4 and the pseudoknot forming P3, all occurring in the core of the structure, were not predicted. Out of the 127 established ones, 95 base pairs were predicted correctly, giving a score of 75%. If the variant stems P 2.1, P5c and P9 are included, the percentage even rises to 83%. These figures should be compared to the 79% mentioned by Turner *et al.* (5) or the 74% reported by Zuker (7). However, the latter two results were obtained using improved energy parameters or including suboptimal foldings (5, 7). A comparison of our result with two other predicted structures presented in the literature can not be made because of the poor presentation of the resulting secondary structures (10, 28).

An interesting aspect of our algorithm is that it is exactly known at which stage the predicted structure starts to deviate from the established one. In this connection it is worth mentioning that after the bottom part of P2.1(see above) a stem was added which consists of stretches from both P3 and P4, thereby preventing the formation of the latter two stems later in the folding process. This sort of observations may be very helpful in detecting shortcomings and in developing improvements of our program in the future.

TYMV RNA

Differences between our program and others become more apparent upon folding longer sequences (>300 nucleotides). A typical result is presented here for the 694 nucleotides long coat protein messenger RNA of turnip yellow mosaic virus (TYMV) (30). This viral RNA possesses a tRNA-like structure at its 3′ terminus for which a detailed secondary and tertiary structure is available based on a number of experimental methods (31−33). In a schematic way Fig. 7 shows a comparison between the results obtained with our program and FOLD (the one described by Zuker and Stiegler (1), GCG version 6.1). Our program folds this viral RNA in a series of relatively small RNA domains without many long distance interactions (Fig. 7B), in contrast to the more compact structure predicted using the program FOLD. The structure predicted by the Zuker program (Fig. 7A) also shows that the 3′ and 5′ end are brought closely together, a result often encountered when this program folds long sequences. In fact, only a few common stems are found. However, our program almost perfectly predicts the tRNA-like structure at the 3′ end, comprising stem I to IV and including the pseudoknot in the aminoacyl acceptor arm at the very 3′ end. It is worth mentioning here that all seven valine-accepting tRNA-like structures known to date (34) were completely predicted with our program except for the anticodon stem in Andean potato latent virus (APLV) RNA.

Furthermore, stem V and VI, which were proposed earlier based upon structure mapping (33), are both present in our predicted structure. Finally, the single stranded region between stem V and VI is in agreement with its sensitivity to nucleases and degradation by RNase H after hybridisation of short deoxyoligonucleotides (35). Based upon these results for the 3′ terminal 150 nucleotides we assume that a large part of the rest of the structure will turn out to be correct.

3′ Noncoding region of TMV RNA

One of the best RNA sequences to test our program for the correct prediction of pseudoknotted structures is the 200 nucleotides long 3′ terminal noncoding region of tobacco mosaic virus (TMV) RNA. We have shown previously, using structure mapping, sequence comparisons and model building, that this RNA stretch harbours five pseudoknots (19). Four of them are of the type illustrated in Fig. 1 and one includes a large bulge loop in the tRNA-like moiety (see Fig. 8).

The stems predicted with our program upon folding this 3′ noncoding region are indicated in Fig. 8. They all correspond to stem regions as proposed in the earlier model. The tRNA-like structure at the 3′ end was folded correctly except for the pseudoknot involving the large bulge loop in the putative anticodon arm. The latter is due to the fact that the program is not able to predict this type of pseudoknot with the present energy rules. Taken together, 11 out of the 13 constitutive stems were found, while also one particular base pair (G_{111}-C_{122}) was absent. A folding of the 5′ noncoding region of the related U2-strain of TMV RNA, for which a similar model was proposed recently (36), even produced 12 out of the 13 proposed stem segments. Only the upper segment of the equivalent of the anticodon arm was absent.

FOLD, the program of Zuker, gives a largely different solution for these RNA fragments which is not surprising in view of the many pseudoknotted structures present (results not shown).

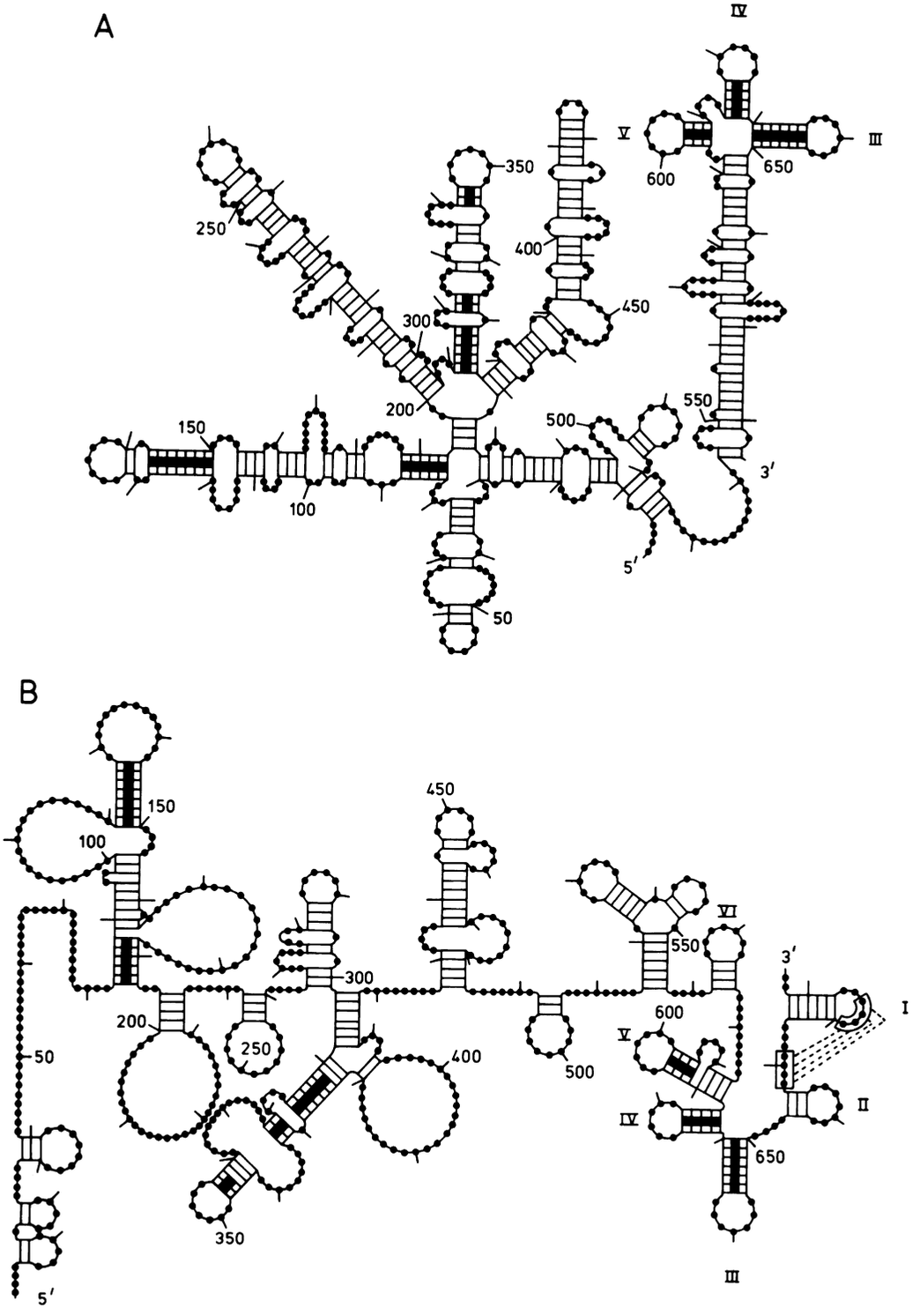

Fig. 7. Schematic presentation of the predicted secondary structure of the coat protein messenger RNA of turnip yellow mosaic virus (TYMV). (A) Secondary structure produced by the RNAFOLD program of Zuker using the energy parameters as reviewed by Turner and coworkers (5). (B) Secondary structure predicted by our program using the energy values given in (18). Base pairs are given as thin lines and unpaired nucleotides as black dots. Every tenth nucleotide is indicated by a tick. The thick black bars represent stem regions which are identical in both structures. The Roman numbers indicate individual hairpins as defined previously (31,33). The dashed lines indicate the pseudoknot interaction in the aminoacyl acceptor arm of the tRNA-like structure.

Application on RNA with an unknown secondary structure

Our program also detected some pseudoknotted structures not described before. An interesting example is the 5′ noncoding region of some picornaviral RNAs. Analysis of this region of three different strains of foot and mouse disease virus (FMDV) RNA revealed three or four consecutive pseudoknots just downstream of the polyC tract (14). All these pseudoknots had the same characteristics, because each was part of a sequence repeated three or four times. Six out of the eleven possible

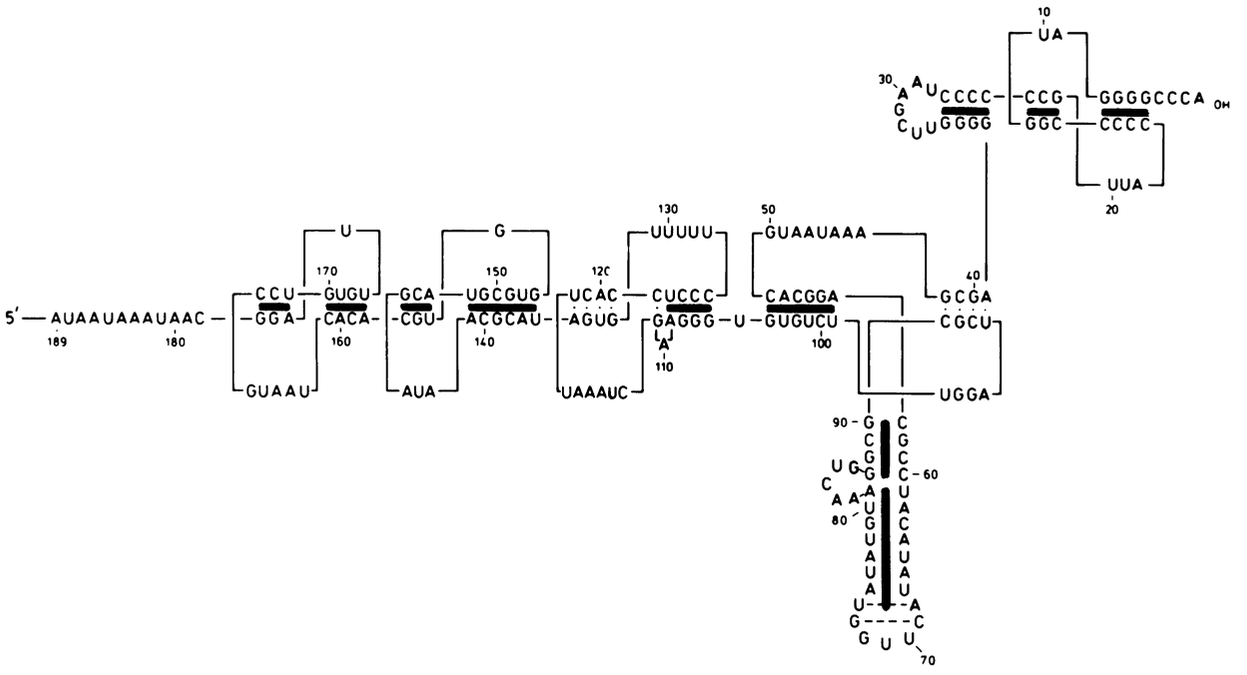

Fig. 8. Secondary structure of the 3′ noncoding region of tobacco mosaic virus (TMV) RNA as proposed previously (19). The numbering of the nucleotides is from the 3′ end. The black bars indicate the stem regions predicted with our program.

pseudoknots were predicted directly upon folding each of the three 700 nucleotides long stretches in a single cycle. The other five were located easily by an intra- and interstrain comparison of the repeated sequences. For further details see Ref. 14. Folding of other picornaviral 5′ noncoding regions brought to light a number of pseudoknots of the type as illustrated in Fig. 1, as will be published elsewhere (C.P., manuscript in preparation).

Pseudoknotted structures have also been detected in the coding region of many retroviral, coronaviral and luteoviral RNAs just downstream of a sequence in an overlap of two open reading frames where frameshifting into the −1 direction occurs (37). The search for pseudoknots in these regions was inspired by the suggestion of others that a pseudoknotted structure could be present in the *gag-pol* overlap of Rous sarcoma virus RNA (38). Recently, experimental evidence for a pseudoknot downstream of the frameshift site in avian infectious bronchitis virus (IBV) RNA has been reported (39). These authors also proposed similar pseudoknotted structures in a number of retroviral RNAs. Some of these proposed structures were predicted directly by our program.

DISCUSSION

The computer program presented in this paper is the first one able to predict certain features of RNA tertiary structure, together with an otherwise orthodox secondary structure. Apart from its usefulness for the prediction of pseudoknotted structures, it performs on a better level than the Zuker program, especially when long RNA sequences are involved. We have regularly used our program for almost four years. It has been thoroughly tested and improved.

A great advantage of the presented algorithm, in combination with the use of the APL language, is that the speed is high and that it can be used on personal computers. The computation time needed with our program is only proportional to N^2 (N being the number of nucleotides in the RNA chain) and the program therefore can handle some 1500 nucleotides in a single overnight run on a standard 640 kbyte IBM-compatible 8088 PC. Other programs, like the one presented by Zuker and Stiegler (1) require computation times proportional to N^3 to N^4 (for a discussion: see (3)).

The algorithm described here is similar to the one developed by Martinez (10) and later by Stüber (28), in which also one stem after the other is added to a nascent structure. As already noted by Martinez (40) such an algorithm may be the only existing one suitable for introducing structural motifs like pseudoknots. We strongly believe that the value of this type of algorithm has been underestimated. This may partially be due to the low number of examples given in the original papers (10, 28) (see also (3)) or found in the literature.

Many programs calculate the secondary structure with the lowest free energy. The fact that the predicted structure often deviates from the established one may be ascribed to imperfect energy parameters available, but it is questionable if this is the conclusive answer. Recently, programs which can also generate suboptimal structures have been published (6−9). Although our program does not return the structure with the lowest free energy *per se*, the free energy content of the resulting structure often does not deviate much from the global energy minimum. Instead, the program returns the structure which probably forms easiest.

The algorithm developed here is based on a simulation of RNA folding, starting from a completely denatured RNA chain. Though RNA will seldom or never occur as such *in vivo*, it is not unrealistic to assume that it can fold into the native conformation upon renaturation. This at least is implicit in the numerous studies performed on RNA after it has been purified on denaturing columns or polyacrylamide gels. It is also interesting to mention that the structure at the 3′ noncoding region of TMV RNA (see Fig. 8) is absent in the virion but apparently folds reproducibly after phenol extraction of the viral RNA. These

are facts which are seldom realised or appreciated by many workers in the field of nucleic acids.

We assume that the folding process, starting from a fully denatured RNA will strongly resemble the one *in vivo* where a so-called sequential folding will prevail (41). In this case, however, the RNA starts folding at the 5′ end irrespective of the nature of the 3′ end, since this has not been formed yet. It does so by a stepwise process in which intermediate structures are formed of which only the more labile parts can or will undergo refolding (see also (42)). We have implemented this concept in our program by giving stems a penalty proportional to their distance from the 5′ side. Since we have not yet systematically tested what the magnitude of the penalty should be for proper folding, the examples presented in this paper were calculated without this feature.

Our program starts folding the most stable stem regions present, assuming that these are kinetically favoured and therefore act as nucleation centres for further folding. However, there is the problem that during these first steps the early formation of a stable long distance interaction can trap a large part of the chain in a wrong structure. This problem has been partially circumvented by introducing a higher penalty for such long range interactions (see also (5)). This adjustment is justified as long as no experimentally determined energy parameters are known for these long range interactions. Also, it is not completely clear how the interplay of the energy of loop formation and the energy of basepairing governs the rate of stem formation. Since it can be expected that the largely enthalpy driven energy of basepairing is of less importance for this rate than the entropy driven energy of loop formation, we provided for a means to emphasize the latter in selecting a stem.

One of the main problems in predicting pseudoknots is the complete lack of thermodynamic parameters for the loops created by these structural motifs. There is only one publication dealing with a thermodynamic study of an oligonucleotide harbouring a pseudoknot, but no rules can be derived from it as yet (43). It is important to point out that the value of 4.2 kcal/mole we have used so far for all loops is an estimation only, based on our experimental experience with some particular pseudoknots occurring in plant viral RNAs (19, 31) and on the assumption that the values for small connecting loops will not deviate that much from those of medium-sized hairpin loops. This value is probably close to the real one in view of the success we had in using it. On the other hand, there is strong reason to suspect that the two connecting loops of the relatively simple pseudoknot as shown in Fig.1 will have different values (see (12)). Anyhow, it is clear that collecting thermodynamic data for pseudoknotted structures is of great importance for their computer prediction.

Another problem is that some of the predicted pseudoknotted structures are sterically just impossible. Solution of this problem requires more insight in the three-dimensional consequences of all types of possible pseudoknots. Such an analysis using model building with computer graphics is under way and may help to improve the program further.

Finally we want to point out that all foldings reported in this paper were carried out without any constraints from experimental data or whatsoever. Such constraints can be easily introduced, as well as other options for introducing the concept of sequential folding (41) or the search for conserved structural motifs or domains among related RNAs (7, 9). Moreover, we expect an improvement of our predictions by implementing the free energy parameters of Turner and coworkers (25).

ACKNOWLEDGEMENTS

We thank Krijn Rietveld for his contributions in the early stages of this work.

REFERENCES

1. Zuker, M. and Stiegler, P. (1981) Nucleic Acids Res., **9**, 133–148.
2. Zuker, M. and Sankoff, D. (1984) Bull. Math. Biol., **46**, 591–621.
3. Gouy, M. (1987) In Nucleic Acid and Protein Sequence Analysis: A Practical Approach, Bishop, M.J., Rawlings, C.J. (eds), IRL Press, Oxford, pp. 259–284.
4. Tinoco, I. Jr., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. and Gralla, J. (1973) Nature New Biol., **246**, 40–41.
5. Turner, D.H., Sugimoto, N. and Freier, S.M. (1988) Ann. Rev. Biophys. Biophys. Chem., **17**, 167–192.
6. Williams, A.L. and Tinoco, I. Jr. (1986) Nucleic Acids Res., **14**, 299–315.
7. Zuker, M. (1989) Science, **244**, 48–52.
8. Le, S., Chen, J.-H., Nussinov, R. and Maizel, J.V. Jr. (1988) CABIOS, **4**, 337–344.
9. Konings, D.A.M. and Hogeweg, P. (1989) J. Mol. Biol., **207**, 597–614.
10. Martinez, H.M. (1984) Nucleic Acids Res., **12**, 323–334.
11. Pleij, C.W.A., Rietveld, K. and Bosch, L. (1985) Nucleic Acids Res., **13**, 1717–1731.
12. Pleij, C.W.A. and Bosch, L. (1989) In Dahlberg, J.E. and Abelson, J.N. (eds), RNA Processing, Part A: General Methods. Methods in Enzymology, Vol. **180**, 289–303.
13. Studnicka, G.M., Rahn, G.M., Cummings, I.W. and Salser, W.A. (1978) Nucleic Acids Res., **5**, 3365–3387.
14. Clarke, B.E., Brown, A.L., Currey, K.M., Newton, S.E., Rowlands, D.J. and Carroll, A.R. (1987) Nucleic Acids Res., **15**, 7067–7079.
15. Van der Veen, R., Arnberg, A.C. and Grivell, L.A. (1987) EMBO J., **6**, 1079–1084.
16. Agsteribbe, E. and Hartog, M. (1987) Nucleic Acids Res., **15**, 7249–7263.
17. Saenger, W. (1984) Principles of Nucleic Acid Structure. Springer-Verlag, New York, pp. 141–149.
18. Jacobson, A.B., Good, L., Simonetti, M. and Zuker, M. (1984) Nucleic Acids Res., **12**, 45–52.
19. Van Belkum, A., Abrahams, J.P., Pleij, C.W.A. and Bosch, L. (1985) Nucleic Acids Res., **13**, 7673–7686.
20. Salser, W. (1977) Cold Spring Harbor Symp. Quant. Biol., **42**, 985–1002.
21. Groebe, D.R. and Uhlenbeck, O.C. (1988) Nucleic Acids Res., **16**, 11725–11735.
22. Jacobson, H., Stockmayer, W.H. (1950) J. Chem. Phys., **18**, 1600–1606.
23. Papanicolau, C.M., Gouy, M. and Ninio, J. (1984) Nucleic Acids Res., **12**, 31–44.
24. Sprinzl, M., Hartman, T., Weber, J., Blank, J. and Zeidler, R. (1989) Nucleic Acids Res., **17**, r1–r61.
25. Jaeger, J.A., Turner, D.H. and Zuker, M. (1989) Proc. Natl. Acad. Sci. USA **86**, 7706–7710.
26. Van Delft, J.H.M., Marinon, B., Schmidt, D.S. and Bosch, L. (1987) Nucleic Acids Res., **15**, 9515–9530.
27. Cech, T.R., Tanner, N.K., Tinoco, I. Jr.,Weir, B.R., Zuker, M. and Perlman, P.S. (1983) Proc. Natl. Acad. Sci. USA, **80**, 3903–3907.
28. Stüber, K. (1985) CABIOS, **1**, 35–42.
29. Cech, T.R. (1988) Gene, **73**, 259–271.
30. Guilley, H. and Briand, J.P. (1978) Cell, **15**, 113–122.
31. Van Belkum, A., Verlaan, P., Jiang, B.K., Pleij, C. and Bosch, L. (1988) Nucleic Acids Res., **16**, 1931–1950.
32. Dumas, P., Moras, D., Florentz, C., Giegé, R., Verlaan, P., Van Belkum, A. and Pleij, C.W.A. (1987) J. Biomol. Struct. Dynamics, **4**, 707–728.
33. Florentz, C., Briand, J.P., Romby, P., Hirth, L., Ebel, J.P. and Giegé, R. (1982) EMBO J., **1**, 269–276.
34. Van Belkum, A., Jiang, B.K., Rietveld, K., Pleij, C.W.A. and Bosch, L. (1987) Biochemistry, **26**, 1144–1151.
35. Rietveld, K. (1984) Ph.D. Thesis, University of Leiden.
36. Garcia Arenal, F. (1988) Virology, **167**, 201–206.
37. Ten Dam, E.B., Pleij, C.W.A. and Bosch, L. (1990) Virus Genes, in press.
38. Jacks, T., Madhani, H.D., Masiarz, F.R. and Varmus, H.E. (1988) Cell, **55**, 447–458.
39. Brierley, I., Digard, P. and Inglis, S.C. (1989) Cell, **57**, 537–547.
40. Martinez, H.M. (1988) Nucleic Acids Res., **16**, 1789–1798.
41. Nussinov, R. and Tinoco, I. Jr. (1981) J. Mol. Biol., **151**, 519–533.
42. Mironov, A.A., Dyakonova, L.P. and Kister, A.E. (1985) J. Biomol. Struct. Dynamics, **2**, 953–962.
43. Puglisi, J.D., Wyatt, J.R. and Tinoco, I. Jr. (1988) Nature, **331**, 283–286.