The computer simulation of RNA folding involving pseudoknot formation

A.P. Gultyaev
All-Union Institute of Influenza, Laboratory of Genetic Engineering, prof. Popova str. 15/17, 197022 Leningrad, USSR

Received July 11, 1990; Revised and Accepted January 14, 1991

ABSTRACT

The algorithm and the program for the prediction of RNA secondary structure with pseudoknot formation have been proposed. The algorithm simulates stepwise folding by generating random structures using Monte Carlo method, followed by the selection of helices to final structure on the basis of both their probabilities of occurrence in a random structure and free energy parameters. The program versions have been tested on ribosomal RNA structures and on RNAs with pseudoknots evidenced by experimental data. It is shown that the simulation of folding during RNA synthesis improves the results. The introduction of pseudoknot formation permits to predict the pseudoknotted structures and to improve the prediction of long-range interactions. The computer program is rather fast and allows to predict the structures for long RNAs without using large memory volumes in usual personal computer.

INTRODUCTION

One of the essential structural properties of RNA is the formation of secondary structure consisting of stacking regions and loops. The RNA secondary structure plays an important role in many processes involving RNA (1). Due to the difficulties in the experimental determination of RNA secondary structure the methods of theoretical prediction on the known sequence are often used.

Many different algorithms for such predictions were developed (2-10). Nevertheless, this problem has not been solved. The algorithms which search for the state of minimal free energy using the estimated values for stacking and loop regions are the most used (4, 5). However, for long molecules these predictions are not consistent with the models obtained from phylogenetic comparisons which proved to be very useful, especially in the case of ribosomal RNAs (11). The second problem in the application of such algorithms is the large computer memory requirements and the time consumed for calculation. The methods based on the folding simulation (6-8) or using statistic analysis of sequence (9, 10) seem to be more reliable for long RNAs.

In addition to well-known structure elements such as stacking regions, hairpin, internal and bifurcation loops, the pseudoknot structures have been recently evidenced or proposed in the number of RNAs (12-19). It seems that these tertiary interactions must be considered in the predictions to obtain correct results. However the energetic parameters of pseudoknots are unknown and the majority of published algorithms do not account for these structures.

In the presented paper the algorithm and the program simulating RNA folding are proposed. The algorithm simulates the stepwise folding by generating random structures by Monte Carlo method followed by the selection of helices supposed to form on certain step of folding. This selection is based on both the probabilities of helices to occur in random structure and their free energies.

The possibility of folding during RNA synthesis is considered. The pseudoknot formation was also incorporated in the algorithm and this was shown to improve the results considerably. The program was tested on small subunit ribosomal RNAs and on RNAs with known or supposed pseudoknots.

METHODS

The RNA folding was simulated by stepwise selection of helices compatible with those selected on previous steps. On each step the helices compatible with the previously constructed structure and having stacking energies lower than some level are searched for. Then the set of random structures is created of these helices and the frequency of occurrence in the structures and the free energy gain for each helix after its adding to pre-existing structure. The product of this frequency and the free energy gain was chosen as the parameter for a helix to be selected to the predicted structure. Thus, the algorithm is divided into four main steps:

1. The calculation of all helices with stacking free energies less than some level (if some structure is already built, all these helices must be compatible with the previous structure).
2. The construction of the set of random structures and the calculation of the product of occurrence frequency in a random structure and the free energy gain for each helix after its adding to pre-existing structure.
3. The selection of helices to the predicted structure in the descending order of parameters from step 2.
4. The change of the level for step 1.

Then the procedure is repeated.

The random structures were generated by Monte Carlo method. The generation begins from some random helix of the set of helices, produced by the random number generator. The second
random one is taken from those compatible with the first and this procedure is repeated until no further helices could be added. The structure obtained is considered as 'random structure'. One hundred of such structures are generated and the frequencies of incorporation to these structures for all helices are calculated. It was checked that this number (100) was sufficient to avoid its influence on the result. Helices incorporated in at least half of the random structures (50) are considered in the selection to the final structure and the product of free energy gain and the probability to occur in a random structure are calculated for each helix in a given cycle of construction.

The values for stacking and loop energies from Freier et al. (20) were used. The bifurcation loop destabilizing energies of 0.02 kcal/mole per base were introduced. The energies for unpaired terminal nucleotides and terminal mismatches were not considered. The pseudoknot destabilizing energies were approximated by the sums of hairpin values for corresponding loops divided by two. The loops located in the helical grooves of the helices forming pseudoknots (12) were considered to be at least 2 bases long. The helices forming pseudoknots with more than two helices were not included in the structure.

The algorithm was also modified to account for the folding during RNA synthesis. In these versions first cycles of structure construction were performed on uncompleted RNA chains with increasing the chain after each cycle.

Other parameters of the program were changed in different versions as indicated in 'Results and discussion'. These parameters include the rule of change of the level in the step 1, the number of helices selected in each cycle and the rate of chain length increase in the RNA synthesis simulation.

The program was written in language 'turboC' and was run on Amstrad PC 1512 computer with 8 MHz 8088 processor without coprocessor.

The output for user contains the following information: coordinates of helices (positions of terminal paired nucleotides), the free energy gains of helices after their adding to the previous structure, the helix stacking energies, the current free energies of overall structure constructed at any moment of program run and the number of helices forming pseudoknot structure with each helix. The program works with the sequence files from both EMBL Nucleotide Sequence Data Library and GenBank database.

The program is available from the author upon written request.

RESULTS AND DISCUSSION
Algorithm background
The prediction of RNA secondary structures of large molecules has two main problems: the computer time and memory requirements in calculation and the problem of confirming the experimental and phylogenetic data. The search for free energy minimum is rather difficult for long RNAs and can lead to structures which are not consistent with phylogenetic data. One of the ways to overcome these difficulties is to simulate in some manner the folding process (6–8).

Such simulation may be based on the successive addition of helices to the constructed structure. This addition would use some criterion for the selection of added helices. Even the simple procedure of adding the longest or the most energetic of possible helices can be useful (6). Nevertheless the selection has to consider the competition between the helices: the choice of certain helix may prevent the formation of structure fragment with better energy parameters, consisting of shorter helices.

This difficulty can be, at least partially, overcome by including the competition between the helices based on their mutual compatibility rather than on energetic parameters. The competition can be simulated by generating a number of random structures consisting of some set of helices. The frequency of occurrence of given helix in these structures will be the measure for the helix to permit the formation of larger number of others. It can be used in some combination with the energetic parameters of the helix. In the presented algorithm the product of the free energy gain and the probability to be incorporated to a random structure is chosen as the parameter which maximum value determines the stepwise selection of helices. The helices which occur in less than half of random structures (incompatible with the great number of others) are not selected even when they have large free energy value.

In each step only the helices with stacking energy less than some level are considered and the level is increased (with the decrease in the absolute value) as the structure is constructed. This may resemble the folding process: the more stable helices fold on the early stages. We can proceed further and try to simulate the folding which may occur before the RNA synthesis completion.

The elements of tertiary interactions called pseudoknots were proposed for RNA secondary structure (12). The pseudoknots were evidenced in a number of RNAs and shown to play an important role in RNA function (12–19). It is evident that the correct prediction of the pseudoknots would improve any algorithm of RNA secondary structure prediction. The energetic parameters of pseudoknot structures are unknown. However, the presented algorithm is not very sensitive to energy scale and some approximations may be used for pseudoknot energies. Puglisi et al. (21) have shown that the synthetic pseudoknotted RNA oligonucleotide melts as single structure with the intermediate parameters between two hairpin values. So the first approximation for pseudoknot loop destabilizing energies can involve the intermediate values of hairpin-like energies of all loops formed in the pseudoknot. Each pseudoknot includes three loops. In the presented algorithm the destabilizing energy of these loops is equal to the half of the sum of destabilizing values for hairpin loops with the same sizes. In the case when one loop is absent and pseudoknot consists of two coaxial helices the destabilizing energy will be intermediate between two hairpin energies. The pseudoknots with pairing of distant regions are also possible and the loops may have their own structure (12). In this case the loop size may be considered excluding the folded part.

The algorithm and the program for RNA secondary structure prediction including pseudoknots has been published during the preparation of the manuscript (22). This program uses the constant destabilizing energy value 4.2 kcal/mole for pseudoknots with loops shorter than fifteen nucleotides. The approach described in (22) also simulates stepwise folding; the algorithm presented here differs from it mainly by the account of helix competition using random structure generation and by the consideration of folding during RNA synthesis.

The secondary structures of ribosomal RNAs are the most investigated. It is reasonable to test the program versions on these structures and on RNAs for which the pseudoknots have been evidenced. Three variants of program were developed: the version calculating on the full-length sequence without pseudoknots, the version simulating the folding during the RNA
synthesis, and the versions with pseudoknot formation. As it will be seen further, this order coincides with the rise in the predictive ability.

**Folding of full-size RNAs**

In this version the folding begins from the helices whose stacking energies are less than some level depending on RNA length. The selection of helices in each cycle is continued until the number of helices remaining in this set (helices incompatible with the selected ones are eliminated) is more than the half of the initial number. The energy level for calculating the set in a given cycle is determined by the number of helices selected on the previous steps. The different dependencies of initial energy level and of its change were probed. This version was tested on small subunit ribosomal RNAs from *E. coli* (1542 bases), *H. volcanii* (1472), *S. cerevisiae* (1798), *X. laevis* (1826) and on 23S-RNA from *E. coli* (2904).

Thus, such simulation provides the basis for predicting the secondary structure of any RNA chain and its change were probed. This version was tested on small subunit ribosomal RNAs from *E. coli* (1542 bases), *H. volcanii* (1472), *S. cerevisiae* (1798), *X. laevis* (1826) and on 23S-RNA from *E. coli* (2904).

In certain region of energy levels chosen for calculating the helices there is almost no dependence on these values. The majority of helices forming hairpins and their expansions forming internal loops are predicted correctly. These helices appear mainly in the first cycles of simulation. But long-range interactions evolved in the simulation are not consistent with RNA models and depend strongly on the energy limits in the cycles.

Thus, such simulation provides the basis for predicting the hairpins and stem-loop structures but predicted helices forming the bifurcation loops are unreliable. The program may be stopped before the full prediction and one can obtain about 25–30 helices and depend strongly on the energy limits in the cycles.

### Table 1. The predicted (+) and unpredicted (−) helices in the secondary structures of 16S-like RNAs from *E. coli* (E.c), *H. volcanii* (H.v.) and *S. cerevisiae* (S.c). The nucleotide numbers are given for *E. coli* sequence.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>9–13</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>655–672</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>17–20</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>677–684</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>27–37</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>769–775</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>39–47</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>783–786</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>51–58</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>821–827</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>73–82</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>829–840</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>113–115</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>888–891</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>122–128</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>893–902</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>136–142</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>926–933</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>144–150</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>938–943</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>153–158</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>946–955</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>198–207</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>960–963</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>240–245</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>984–990</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>252–259</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>997–1003</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>289–296</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1006–1012</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>339–342</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1046–1067</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>339–342</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1068–1071</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>368–371</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1072–1076</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>373–385</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1086–1089</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>406–409</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1113–1117</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>416–419</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1118–1124</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>437–446</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1128–1135</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>455–462</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1161–1165</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>500–504</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1241–1247</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>511–517</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1253–1255</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>576–580</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1308–1314</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>588–594</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1419–1430</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>597–606</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1435–1445</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>612–617</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1506–1515</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>664–672</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1508–1517</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>677–684</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1519–1529</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

*The eucaryotic structures in this region are not homologous to the procaryotic ones.*
This version is approximately 1.5 times faster than the previous one.

The folding with pseudoknot formation

The formation of pseudoknots was introduced in the folding simulation as described in 'Methods'. This modification in rather rough approximation showed its usefulness for the correct prediction of long-range interactions besides to pseudoknot prediction. The version was tested on small subunit ribosomal RNAs from *E. coli*, *H. volcanii*, *S. cerevisiae* and *X. laevis* and on alpha-operon mRNA from *E. coli* which formed pseudoknot structure evidenced by site-directed mutagenesis.

Several internal program parameters were tested. The results reported below refer to the consensus variant. In this variant the folding begins from chain length of 400 or 500 nucleotides (if RNA length is longer). This length is increased by 50 nucleotides in the subsequent cycles. The number of helices selected in each cycle corresponds to one helix per 400 bases of chain length. The energy level is constant for chains which are shorter than 1000 bases in the cycles with uncompleted RNA chain. The chain length is increased by 500 nucleotides after reaching the length of 1000 nucleotides with simultaneous decrease in the energy level by 2 kcal/mole. After reaching the full RNA length the energy level is increased by the value equal to the number of helices selected in the previous cycle, multiplying by 0.1 kcal/mole, and added to 0.5 kcal/mole. The folding is stopped when this level becomes positive.

The initial energy level may be changed and the results may be slightly different. It seems that the optimal value of this level must be determined by organism and RNA type. To avoid the dependence on this level choice several runs of program were used and five structures with the minimal free energies were chosen for the prediction. The helices were selected to the final prediction in the descending order of their occurrence in these five structures.

Figure 1 shows the degree of coincidence of such prediction with the secondary structure model for *E. coli* 16S-RNA deduced from comparative sequence analysis (23). It is seen that the majority of helices proved by phylogenetic comparisons are predicted correctly (about 75% of base-pairs). The similar results were obtained for other small subunit rRNAs. These predictions have approximately the same percentage of similarity with the published models as the prediction for *E. coli*. Table 1 summarizes the predictions for representatives of the three primary kingdoms, eubacterial (*E. coli*), archaeabacterial (*H. volcanii*) and eucaryotic (*S. cerevisiae*). The homologous helices are referred to their nucleotide numeration in *E. coli* 16S-RNA. These results can be compared with the predictions of Abrahams et al. (22) (the program involving pseudoknots) and of Jaeger et al. (24). The presented algorithm predicts correctly for *E. coli* 16S-RNA 43 of 65 phylogenetically deduced helices (66%) as they are defined in (24) (three or more consecutive basepairs with bulges or interior loops containing less than three nucleotides). This is better compared with the predictions of Abrahams et al. (40%) and of Jaeger et al. (55%). It should be emphasized that the presented simulation does not depend on the knowledge about the rRNA domains used in (22, 24) and the evolved long-range helices are in agreement with the partitioning of three-dimensional structure into the major domains (23).  

---

Figure 1. The comparison of the secondary structure prediction for *E. coli* 16S-rRNA with the model deduced by comparative sequence analysis (23). The predicted fragments of model are shown by solid lines, the dashed lines represent the fragments which are not predicted.

Figure 2. (A) The pseudoknot structure of *E. coli* alpha-operon mRNA evidenced by site-directed mutagenesis (17) and (B) the predicted structure. It should be noted that the predicted helices may be shorter due to the pseudoknot geometry constraints.
The results indicate significant improvements in predictions for *H. volcanii* (63%) also predicts the main long-range helices correctly and should be compared with the value 80% in the prediction (24) for four separate domains. The time necessary for single program run on computer described in 'Methods' varied from about 1.5 hours to 2.5 hours. This seems to be 2–3 times faster than the time which is necessary for the prediction of such long RNAs by the program described by Abrahams et al. (22) and much faster than the times consumed by programs based on the search of optimal or suboptimal energy state (24, 25).

Each of the predicted structures of rRNAs (from *E. coli*, *H. volcanii*, *S. cerevisiae* and *X. laevis*) contained 3–5 pseudoknots. It is difficult to estimate their relevance to the true structures without systematic studies. It may be noted that in all predicted rRNA structures the pseudoknots evolve in the region corresponding to the fragment 1080–1190 of *E. coli* RNA. Thus the introduction of pseudoknot formation in the RNA folding simulation does not produce a great number of pseudoknots but improves the prediction of long-range interactions.

The accuracy of predictions decreases for longer RNAs and the simulation of *E. coli* 23S-RNA folding (2904 nucleotides) produces 55% of helices. This is mainly determined by incorrect prediction of majority of long-range pairings. Nevertheless, some of them are predicted correctly in addition to hairpins whose prediction is much more reliable.

It is known that the synthesis of ribosomal proteins is autogenously regulated and this regulation is determined by ribosomal protein binding to specific mRNA structures which are similar to the binding sites in rRNA (26). The prediction of such specific structures may be useful for the studies on the regulatory functions of ribosomal proteins (27, 28). The translational repression of *E. coli* alpha operon mRNA by S4 protein requires the complex pseudoknot structure evidenced by the systematic studies on site-directed mutations that create compensatory base pair changes (17). Figure 2 shows this structure and the predicted folding. The part of alpha-mRNA between the start and the sequence coding for RNA polymerase requires the complex pseudoknot structure evidenced by the systematic studies on site-directed mutations that create compensatory base pair changes (17). Figure 2 shows this structure and the predicted folding. The part of alpha-mRNA between the start and the sequence coding for RNA polymerase requires the complex pseudoknot structure evidenced by the systematic studies on site-directed mutations that create compensatory base pair changes (17).

The relevance of predicted simulation to possible folding process

It is interesting to estimate the degree of approximation of folding by the proposed algorithm. Two main conclusions can be made. The results indicate to the significant improvements of predictions derived from the introduction of uncompleted chain folding. It is reasonable to suppose that such event really influences strongly on the final structure. Such effect can explain the differences between the predicted structures for the overlapping RNA segments which often appear in the prediction of large structures (29). It may be also concluded that the pseudoknot formation must be considered in the prediction to obtain reliable long-range interactions even in RNAs without considerable pseudoknotting.

However, the proposed algorithm is still a rough approximation of RNA self-organization. The hierarchy of helices based only on their stacking energies does not seem to reflect completely the real situation, mainly for helices which are the expansions of others with internal loop formation. The substantial improvements of the presented model may be expected from more precise consideration of pseudoknot free energies and the detailed geometry of pseudoknotting which is almost ignored here. For example, the total helical twist of RNA helix determined by its length may influence on the mutual arrangement of flanking fragments (30) and this would contribute to the free energy of pseudoknot involving the helix. The consideration of complex pseudoknots consisting of large number of long-range interactions seems to be necessary for intron structure predictions (31). The introduction of such structures also requires additional assumptions and makes the program more complicated and slower in operation but would improve the prediction ability.

**ACKNOWLEDGEMENT**

I thank Prof. D.E. Draper (Johns Hopkins University, Baltimore) for helpful discussion and sending manuscript before publication.

**REFERENCES**