Volume 12 Number 1 1984 Nucleic Acids Research

An RNA folding rule

Hugo M.Martinez

Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143, USA

Received 25 August 1983

ABSTRACT

The folding of single-stranded RNA into its secondary structure is postulated to be equivalent to the simple rule that the next double-helical region (stem) to form is the one with the largest equilibrium constant. The rule is tested and shown to give results consistent with the enzyme cleavage data of several sequences. Computational time complexity is of order NxN for a sequence of N bases. A modification of the rule provides for the probabilistic choice of the next stem among those having an equilibrium constant within a specified range of the largest. Populations of competing structures are thus generated for detecting common characteristics and for assessing the applicability of the simple rule.

INTRODUCTION

The computational time complexity of current methods for determining secondary structure based on global energy minimization is at least of order NxNxN for a sequence of N bases (1,2,6). This relative inefficiency has prompted us to ask: Assuming post-transcription, free-folding conditions, how does an RNA molecule fold, and can it be efficiently simulated?

To explore various possibilities we chose the strategy of first determining the potential double-helical regions (stems), as is common with the "stem-oriented" approach to secondary structure (1,2). We postulated that the folding process is equivalent to adding one stem at a time to a growing structure and then experimented with various rules for determining the sequence of stems added. A rule which seems promising is simply: Of all the remaining unformed stems which are compatible with those constituting the current structure, choose the one with the largest equilibrium constant. Compatibility here means that for a stem to form it must not have any bases in common or form a

knot with any of the stems of the current structure. Two stems are said to form a knot if their loops cannot be separated.

Examples can be devised to show that the proposed rule cannot always result in a structure having the lowest free energy. One must therefore approach the question of applicability by testing the rule on RNAs in which the secondary structure is known to be essential to function. Appealing to evolutionary and cooperativity arguments as given below, we predict that for such RNAs the rule should reflect the uniqueness of the folding, while for RNAs in which secondary structure is not important the lack of a strongly favored structure will render the rule inappropriate.

RATIONALE FOR AND TESTS OF THE RULE

The rationale guiding the choice of our simple rule is the consideration that existing RNA molecules must surely have undergone evolution and that any secondary structure pertinent to their function is likely to reflect this historical element. In particular, perhaps the folding process occurs in stages, each of which is stable by itself or contributes most strongly to the stability of the previous one. Also, the emergence of a new stage very likely accompanies an increased length of the molecule; so if one appeals to the idea that what had been achieved thus far must be largely conserved, then the new stage must indeed be highly dependent on what has already evolved. There might, therefore, be a sequential cooperativity in the folding and it is natural to associate it with one which is cooperative in the free energy sense, that is, the free energy of formation of the next stage is dependent on what has already formed and this cooperativity helps select out the particular next one.

The specific form of the cooperativity is envisaged to be in the entropy part of the free energy of formation. This part is concerned with the entropy decrease which occurs when two halves of a stem come together to form base pairs. Thus, if a stem forms in the single strand joining the two halves of an unformed stem, then the entropy decrease when the latter forms will be less than if the intermediate one had not formed. Hence the

potential cooperativity.

Our choice of 'largest equilibrium constant' as the criterion for selecting the next stem is based on the consideration that it gives an approximation to relative amounts of formation of the competing stems. For instance, let S denote the current structure at some stage of the folding and P the population of stems competing to form next. For stem s in P, the formation of s is governed by the reaction S <-> Ss in which Ss is taken to mean structure S with stem s just formed. If we were to isolate this reaction in the sense that S is not allowed to break down and no Ss is allowed to add on another stem, then letting K(s) be the equilibrium constant for the reaction S <-> Ss the concentration of Ss normalized by the sum of the concentrations of Ss' for all s' in P will be given by K(s)/Q in which Q is the sum of the K(s') for all s' in P. The equilibrium constant K(s) is calculated from the relation $K(s) = \exp(-\Delta G(s)/RT)$, in which $\Delta G(s)$ is the corresponding standard free energy of formation of stem s relative to the current structure S. It is thus seen that a dominant K(s) will imply a biasing of the competing reactions toward a preponderance of the structure Ss as compared to Ss' for all other s' in P. And because of the relationship between K(s) and $\Delta G(s)$, any cooperative effect on $\Delta G(s)$ will be exponentially magnified in the corresponding value of K(s).

The assumption of an isolated set of competing reactions necessarily makes our estimate of relative concentrations of the competing structures approximate. Nevertheless, if there is a large disparity between equilibrium constants due to cooperativity or otherwise, the approximation should tend to be a good one.

Our first two tests of the rule were to the structure of two transfer RNA precursors. Their sequences and cleavage data were as reported in (3). Figures 1a and 1b show the resulting structures and how they compare with the cleavage data. The consistency is nearly perfect. But especially noteworthy is the fact that some of the stems which formed underwent a significant change (factor of 10) in their potential free energy of formation as compared to when there was no structure at all. This is interpreted as the cooperative effect.

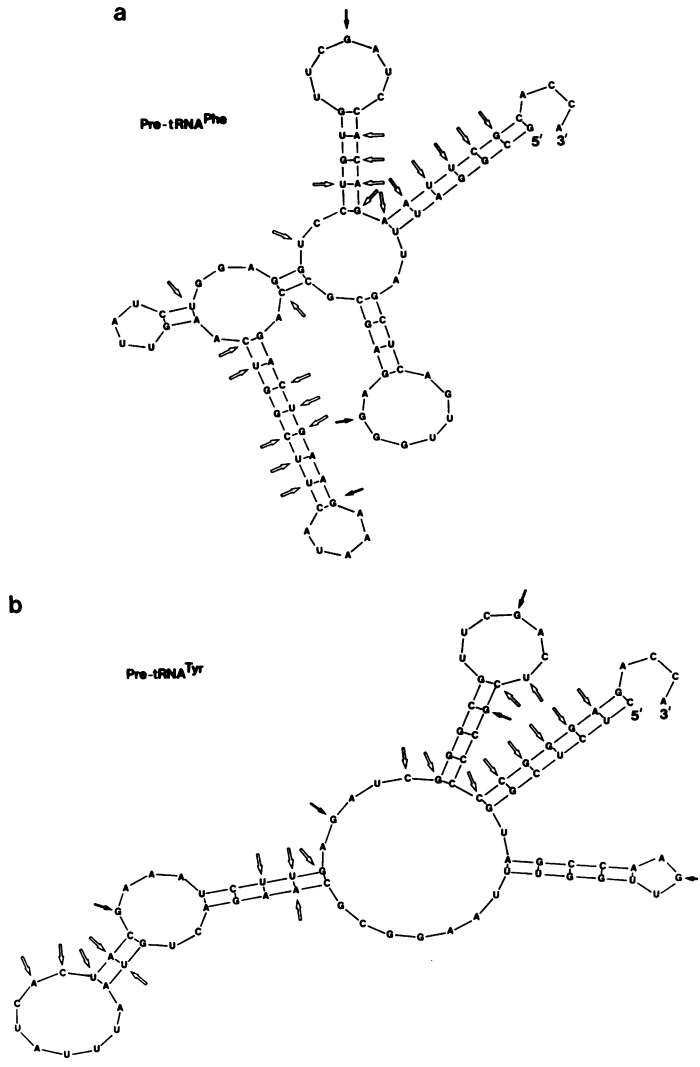

Figure 1
Secondary structures of two trna precursors: a) phenylalanine, and b) tyrosine. The free energies of formation are, respectively, -26.2 and -30.9 kilocalories. The T1 cleavage sites are indicated by solid arrows and the V1 sites by open arrows. Both structures were obtained by the simple rule. In 100 Monte Carlo foldings of each sequence in which there was allowed 100% competition, structure (a) occurred 78 times and structure (b) occurred 80 times. Illustrative of cooperative stem formation is the stem in the lower left corner of Figure 1b. It had a free energy of formation of -0.12 K cal before and -1.6 K cal after the formation of the 5-base pair stem above it.

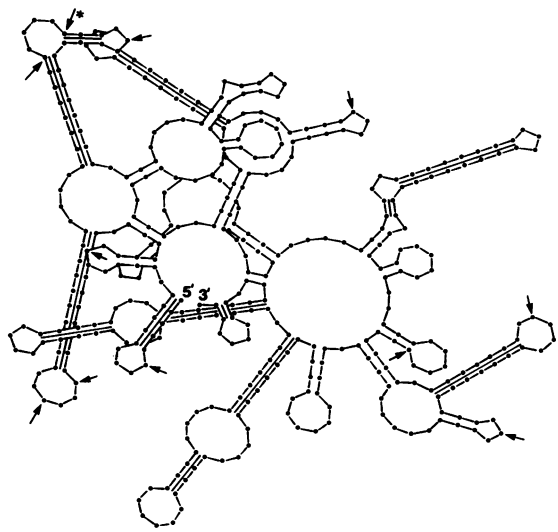

Figure 2
The secondary structure of the self-splicing ribosomal RNA intron sequence reported in (4), as obtained with the simple rule. It has a free energy of formation of -128.9 kilocalories, but does not occur in 20 Monte Carlo foldings in which there was allowed 100% competition. The darkened stems of the structure always occurred in these foldings. The T1 cleavage sites are indicated by the solid arrows. The site highlighted with an asterisk is the only one not consistent with the data (see text for explanation).

The next test was to the self-splicing ribosomal RNA intervening sequence reported in (4). The result is shown in Fig. 2, relative to which four points are notable. First, the structure was obtained without any constraints on what should or should not base pair. This is in contrast to the published structure (loc cit) that was obtained with a global energy minimization method constrained to conform with the T1 cleavage data. The second point is that our energy value of -128.4 K cal is very close to that of the reported structure, -131.4 K cal in (4). Therefore, our structure is a legitimate competitive one. The third point concerns how near the 5' and 3' ends of the molecule have been brought together. A particularly appealing aspect of the published structure is that the 5' and 3' ends are but 20 bases apart in comparison to 433 bases when there is no secondary

structure. If this is the only way by which the 5' and 3' ends can be brought into proximity with one another, then such a structure is certainly consistent with the self-splicing feature of this intron. In comparison, the corresponding distance in our structure is 17 bases; hence, our structure shows the same kind of consistency.

The fourth point concerns the discrepancy with the T1 cleavage data. There is but one discrepancy and is highlighted by an asterisk in Figure 2. It is our contention that this discrepancy can be reconciled by noting that at the location where there should be no base-pairing, the involved stem is rather easily broken. Thus, if in the original selection of the stem list there is imposed the constraint that no stem will be accepted unless its pairing and stacking energy is above 10 RTs (= 5.19 Kcal), then the stem in question does indeed disappear from the structure. This is to say that such a stem is very likely opening and closing (breathing) and thus permits the entry of T1 nuclease as required for cleavage.

A feature of any stable structure must be that no stem can be replaced by one which would further lower the total free energy of formation. An optional test for this feature is incorporated into the program. Thus, once a structure has formed according to the simple rule of sequential stem selection, the structure is subjected to a perturbation regime: each stem is broken and the structure allowed to reconstitute under the simple rule. The regime is stopped when the structure always reconstitutes into the same one no matter which stem is broken. Interestingly, both the precursor tRNAs and the intron studied have been insensitive to such perturbations. A necessary criterion of optimal stability is therefore satisfied in these cases.

GENERATING COMPETING STRUCTURES

The perturbation technique noted above is an attempt to improve the simple rule. It adds but little to the computation time and insures satisfying one optimality criterion.

A further possible improvement is to break stems two at a time from the formed structure and allow it to reconstitute under

the simple rule. But now a potentially heavy penalty is visible because even though the pair selection would be limited to stems whose bases are separated by single strands alone, the number of pairs could become rather large. Instead, it is deemed more practical to generate competing structures by the following Monte Carlo method.

Under the simple rule the next stem chosen to form is the one with the largest equilibrium constant. This rule can be relaxed by stipulating that the next stem to form be chosen from all those which have an equilibrium constant within a specified range of the largest one and that the choice of such a stem be made with a probability proportional to its equilibrium constant relative to this range. For instance, a 100% range would allow all the unformed stems at each step to compete, while a 50% range would allow a stem to compete only if its equilibrium constant is larger than 1/2 the maximum.

Applicability of the rule seems best assessed by using the Monte Carlo method with a 100% range. Applicability then corresponds to the condition that the structure generated by the simple rule should occur with a frequency significantly larger than others. In Figures 1a and 1b the structures shown are noted as occurring respectively, 78 and 80 times out of 100 Monte Carlo runs. These results for the precursor tRNAs favor applicability of the simple rule, while for the other sequence the lack of a dominating structure suggests, from our point of view, that if secondary structure is important in this molecule then it must reside in the features common to all the competing structures. Notable also is the fact that none of the structures generated exhibited significant cooperativity of stem formation. Hence the diversity. In the corresponding Figure 2 for this structure, the stems which always occur in a Monte Carlo run are highlighted by a filled-in stem.

RELATION TO OTHER METHODS

Suggestions have previously been made to the effect that short range interactions dominate the RNA folding process (2,5,6). In (6), for instance, there has purposely been implemented a "kinetic parameter" which limits attention to potential

double helical regions for which the number of bases in the corresponding hairpin loop is less than a prescribed amount. There is then implemented a global energy minimization scheme subject to such a restriction. In contrast, the kinetics approach advocated here makes explicit use of hairpin loops to calculate the relative advantage of a stem to form and uses this calculation to select a sequence of stems. It recognizes the possibility that relative advantage is dependent on what has already formed and hypothesizes that a strong dependency is what renders the folding process unique.

PROGRAM DESCRIPTION AND OPERATION

The program devised to carry out the proposed rule and its elaboration relies on first finding all the potential stems. The option is given to specify minimum stability for a stem relative to its base-pairing and stacking energy. This minimum is specified in RT units for which the default is one RT (= 0.57 kcal at 25 degrees C). One may also optionally specify the temperature, for which the default is 25 degrees C. A stem is interpreted to mean a double-helical region without bulges or internal loops and its base-pairing and stacking energy is calculated according to the table given in (7). The tabulation has been modified to allow for temperature dependence. Bulges and loops are allowed for by the joining of two stems with appropriate consideration for increased stacking energy in the case of bulges. This joining occurs in the addition of a stem to the current structure.

The free energy of formation of a stem consists of its BPS (base-pairing plus stacking) energy plus the entropy decrease resulting from the loss of potential configurations when the two halves of the stem are joined together. As discussed above, it is the entropy decrease which can depend on what has already formed and which we feel should be important in giving direction to the folding process. This decrease is calculated as follows.

To each stem there are associated four base position numbers corresponding to its four ends: top5', lower5', top3' and lower3'. The part of the sequence delimited by its top5' and top3' ends corresponds to its loop which may or may not contain other stems. Before a stem x forms its complementary halves may

or may not be part of the unpaired bases in the loop of an already formed stem, say y. If not, then the entropy decrease upon the formation of stem x is given by klog(U) in which k is a temperature dependent constant and U is the number of currently unpaired bases in its loop. Otherwise, the entropy decrease is given by the quantity [klog(U1) + klog(U2) - klog(U)] in which U is the number of unpaired bases in the loop of stem y prior to the formation of stem x, U1 is the number of unpaired bases in the loop of stem x, and U2 is the number of unpaired bases in the sequence portion extending from top5' of stem y to low5' of stem x and from top3' of stem y to low3' of stem x.

Once the temperature and minimum BPS energy has been specified, the option is given to perturb the folded structure(s). Another option is specification of the percent range relative to the maximum equilibrium constant in the event that one wishes to test for competing structures. Single and multiple foldings are then selectable relative to these parameter choices. The single folding option results in a record of the actual order in which the stems forming the folded structure are added. The multiple folding option, on the other hand, only records the final structure.

A structure, whether nascent or complete, is described in two ways: tabular and graphically. Files of these are automatically created during a run, whether corresponding to a single or multiple folding. Either a tabular or graphical file may be displayed according to an energy or frequency ordering. Thus, if twenty foldings are specified for a multiple folding run, then the resulting distinct final structures are recorded in tabular and graphical form and can be viewed in decreasing order of energy or frequency of occurrence. The graphical output is designed to be displayed on a Tektronix 4010 or 4014 terminal. In order to accommodate very complex figures, the option is given to omit drawing in of the bases.

In the case of multiple runs when doing a Monte Carlo simulation, there is created a file for recording some relevant statistics. Recorded is the number of times a stem occurs in the structures and the frequency of each of the structures. Additionally, for each of the different structures generated there is

given a string encoding for subsequent computerized comparisons of the structures. Such comparisons, which can be based on a variety of cluster analysis techniques, have not been been implemented yet.

Another option which should be mentioned concerns comparison of a generated structure with cleavage data. Scores are given relative to how well a generated structure satisfies such data. No option is given, however, to constrain a structure to fold so as to perfectly satisfy given cleavage data. This option, though relatively easy to implement, has not seemed appropriate to the aims of the method up to this point.

The program is coded in the C language. The potential stems are maintained in an array of records. Each of the records has two sets of pointers to positions in the array. One set is used to describe the nascent structure in the form of a doubly-linked list. The other is used to describe the current set of competing stems, also in the form of a doubly-linked list.

The time complexity of the program is determined by the number of stems that will be allowed to compete. This number is proportional to NxN, where N is the number of bases in the sequence. The proportionality constant is in turn determined by the minimum BPS energy allowed for a stem. The larger the minimum, the smaller the proportionality constant. Given the number of stems, determination of the folded structure by means of the simple rule appears to be proportional to this number. Typical times relative to the intron sequence (4), which is about 400 bases long are: 205 seconds for 4178 stems, 66 seconds for 1448 stems and 14 seconds for 347 stems. These times were obtained on a VAX 11/750 computer.

SUMMARY and DISCUSSION

A method has been presented for predicting the unconstrained folding of RNA into a secondary structure configuration. It has been argued that relevance of the structure should correlate highly with strongly favored folding pathways. Such a pathway is interpreted as being equivalent to the sequential formation of stems in a cooperative manner. The sequence is specified by cooperatively determined equilibrium constants.

The method has been tested on three sequences for which structural data were available and shown to be consistent with such data. Elaboration of the method allows for the Monte Carlo generation of a population of competing structures, relevant statistics of which can be used to detect common characteristics and assess the relevance of secondary structure. In the case of the precursor tRNAs tested, the population statistics are consistent with the established relevance of their secondary structure and argue for a strongly favored folding pathway as predicted by the simple folding rule posed. In the case of the third sequence, the population statistics do not favor a dominant structure. It can therefore be argued that if secondary structure is important then it must reside in the common characteristics of the population. These common characteristics are contained in the structure predicted by the simple rule, but it requires the Monte Carlo elaboration to reveal them.

A program has been described for implementing the method. It is written in the C language and intended primarily for use within a Unix operating system environment. It is available from the author as a separate program or as part of the Sequence Analysis Package of the Biomathematics Computation Laboratory, Dept. of Biochemistry and Biophysics, UCSF.

ACKNOWLEDGEMENTS

We are grateful to Christine Guthrie and Harold Swerdlow for making their data, referenced in (3), available to us prior to its publication. We would also like to acknowledge the helpful discussions with Leonard Peller of UC San Francisco and David Lipman of NIH.

This research was in part supported by NSF Grant PCM 802206.

REFERENCES

1. Studnicka, G.M., Rahn, G.M., Cummings, I.W., Salser, W.A. (1978) Nucleic Acids Res. 5, 3265-3387.
2. Dumas, J-P. and Ninio, J. (1982) Nucleic Acids Res. 10, 197-206.
3. Swerdlow, H. and Guthrie, C. (1983) "Structure of Intron-containing tRNA Precursors: Analysis of Solution Conformation Using Chemical and Enzymatic Probes" submitted to J. Mol. Biol.

4. Cech, T.R., Tanner, K.N., Tinoco, I. Jr., Weir, B.R., Zuker, M. and Perlman, P.S. (1983) "Secondary Structure of the Tetrahymena Ribosomal RNA Intervening Sequence: Structural Homology with Fungal Mitochondrial Intervening Sequences" P.N.A.S., in press.
5. Fresco, J., Alberts, B. and Doty P. (1960) Nature 198, 98-101.
6. Stucker, M. and Stiegler, P. (1981) Nucleic Acids Res. 9, 133-148.
7. Salser, W. (1977) Cold Spring Harbor Symp. Quant. Biol. 42, 985-1002.