A new method to find a set of energetically optimal RNA secondary structures

Giorgio Benedetti, Pasquale De Santis and Stefano Morosetti

Department of Chemistry, University of Rome I, P.le A.Moro 5, 00185 Rome, Italy

Received February 13, 1989; Revised and Accepted June 1, 1989

ABSTRACT
We present a computer method to determine nucleic acid secondary structures. It is based on three steps: 1) the search for all possible helical regions relied on a mathematical approach derived from the convolution theorem; it uses a tetradimensional complex vector representation of the bases along the sequence; 2) a ‘tree’ search for a set of minimum free energy structures, by the aid of an approximate energy evaluation to reduce the computer time requirements; 3) the exact calculation and refinement of the energies. A method to introduce the experimental data and reach an arrangement between them and the free energy minimization criterion is shown. In order to demonstrate the confidence of the program a test on four RNA sequences is performed. The method has computer time requirement proportional to $N^2$, where $N$ is the length of the sequence and retrieves a set of optimal free energy structures.

INTRODUCTION
The importance of the role of secondary and tertiary RNA structure in biological processes is grown in the last years (1−3). It is now generally assumed that primary sequence carries the information required for its actual three-dimensional folding. Predicting secondary structure first and then proceeding on to tertiary structure can be supposed to be a fruitful, if not infallible, approach. With the increase in length and number of the determined nucleic acid sequences, there has been a growing need for algorithms that can efficiently search for the more probable secondary structures. Several methods have been developed to this aim by minimizing the free energy (4−17). Some of these methods predict only one optimal free energy structure for each nucleotide sequence. Alternate equivalent and suboptimal free energy structures are not identified despite their possible biological significance. The methods capable of identifying more than one optimal secondary structure generally require user intervention and/or external constraints (7,11,13).

The computer programs based on all these methods generally require a computational time proportional to $N^3$, and memory requirements of the order of $N^2$, where $N$ is the number of nucleotides in the sequence. The experimental (enzymatic, chemical) data, if considered are generally introduced in the programs as constraints.

Here we present an algorithm able to select a set of optimal free energy structures, with computer time requirement proportional to $N^2$, and with the possibility of introducing a gradual competition between the experimental data and the free energy content of the structure.
Figure 1. a) Physical folding of the sequence around a nucleotide; b) the same folding obtained by shifting the sequence along itself reversed. 1) shift x=0; 2) generic shift x. The helical region shown has initial nucleotide position u≤x and then final nucleotide position t=x−u+1; 3) the same shift x, but in this case a helical region with u>x and then t=x−u+1+N is shown.

METHODS

The main problem to be solved is the search for optimal free energy secondary structures. The most stable secondary structure might not correspond to the biological meaningful one because the tertiary and environmental effects (17), or in some cases, other secondary structures should be invoked in order to completely describe the biological phenomenon (1−3). Therefore our algorithm is designed to predict a set of optimal free energy structures.

Additive free energy contributions considered are the favorable stacking interactions and the destabilizing loop terms as obtained by Salser (18) but other free energy parameters may be employed (19). In particular, we include the contributions from multistem loops, assumed as internal loops. Free energies are assigned also in the case the loops are formed by more than 30 nucleotides, according to Jacobson and Stockmayer (20). The main steps of our method are presented below.

Search for all possible helical regions.

We developed an original method to obtain a list of all possible helical regions that can be derived from a given sequence. These regions can be imagined as obtained by folding the sequence around a nucleotide or comparing the sequence with itself reversed and suitably shifted (see figure 1). This appearance suggested us to use an original mathematical process rising from the convolution theorem. If f(x) is a generic function, its self-convolution is (21):
Table 1. Tetradimensional complex vector assignements

|   | \( f(u) \) | \( \bar{f}(u) \) |
|---|---|---|
| G | \((1, i, 1, i) \ast \frac{1}{2}\) | \((2, 2i, 2, 2i) \ast \frac{1}{2}\) |
| C | \((1, -i, 1, -i) \ast \frac{1}{2}\) | \((1, -i, 1, -i) \ast \frac{1}{2}\) |
| A | \((1, -i, -1, i) \ast \frac{1}{2}\) | \((2, 2i, -2, -2i) \ast \frac{1}{2}\) |
| U | \((1, i, -1, -i) \ast \frac{1}{2}\) | \((2, -2i, 0, 0) \ast \frac{1}{2}\) |
| S | \((0, 0, 0, 0) \ast \frac{1}{2}\) | \((0, 0, 0, 0) \ast \frac{1}{2}\) |

a) Assignment to find Crick-Watson base pairings. b) Assignment to find also GU base pairing. Bases coded as S are blocked by modifications at key groups and they cannot form hydrogen bonds.

\[
h(x) = \int_{-\infty}^{+\infty} f(u)f(x-u) du
\]

In our case the \( f(u) \) has to be a discrete function, and \( u \) is the position number of the nucleotide in the sequence. Then \( h(x) \) becomes:

\[
h(x) = \sum_{u=-\infty}^{+\infty} f(u)f(x-u)
\] (1)

where \( u \) and \( x \) are integral numbers.

In the case of a RNA sequence, the physically meaningful \( u \) and \( x \) values are in the range between 1 and \( N \). \( N \) is the total number of nucleotides in the sequence. As consequence (1) assumes the form:

\[
h(x) = \sum_{u=1}^{N} f(u)f(x-u)
\] (2)

At this point we define \( f(u) \) so that the term \( f(u)f(x-u) \neq 0 \) only if the corresponding nucleotides are hydrogen bonded. Therefore we use tetradimensional complex vector assignment so that (2) becomes:

\[
h(x) = \sum_{u=1}^{N} \bar{f}(u) \cdot \bar{f}(x-u)
\] (3)

and the vectors are as in table 1. \( h(x) \) represents the number of coupled nucleotides, after the function \( f \) is normalized.

We are rather interested in finding helical regions and then consecutive base pairings than their total number for a given folding. Therefore, we modify (3) into:

\[
H_i(x) = \sum_{u=1}^{N} \prod_{k=0}^{l-1} [\bar{f}(u+k) \cdot \bar{f}(x-u-k)]
\]
Figure 2. Schematic representation of possible search mechanisms along the tree of structures. The numbers indicate hypothetical helical regions setting up the structures. Taking into account a new helical region, creates a new level in the growing tree. Refer to the text for a more complete description of this figure.

In this case \( H_l(x) \) represents the number of helical regions of \( l \) nucleotides length present in RNA sequence for the shift under examination. \( l \) ranges between 2 (the minimum length in order to use the free energy stacking parameters) and a maximum value \( l_{\text{max}} \) above that \( H_l(x) \) becomes 0 for all shifts.

The position of the helical regions in the sequence is an essential required information. To obtain that we separately evaluate each term:

\[
P(u,k,l) = \prod_{k=0}^{l-1} [f(u+k) \cdot f(x-u-k)]
\]
Figure 3. Example of a tree of structures in the case of a generic sequence. On the top are reported the sequence and the table of the helical regions used. a) the thin circles represent the structures excluded by the energetic criterion (x) or by removing of the duplicates ones (=) and the thick circles represent the remainder structures; b) the same situation, but only the remainder structures and the multiple pathways to reach them are represented.
Figure 4. Regression line, correlation coefficient (r) and standard deviation (sd) for a set of structures of Tetrahymena thermophila IVS sequence between approximate energies and those obtained by Salser's parameters.

In fact if $P(u,k,l) \neq 0$ and $f(u+k+1) \cdot f(x-u-k-1) = 0$, there is a helical region of length 1, with initial nucleotide at position $u$ for the 5' half region, coupled with a terminal nucleotide at position $t$ for the 3' half region, where:

$$t = x - u + 1 \quad \text{(for } u < x + 1\text{)}$$

or

$$t = x - u + 1 + N \quad \text{(for } u \geq x + 1\text{)}$$

Finally we store the helical regions as a list of $l$, $u$, and $t$ values.

The memory occupancy requirements for this method is of the order of $4N$, opposite to the matrix algorithms of the order of $N^2/2$.

Assembly algorithm.

The approach above described gives us a complete set of the possible helical regions afterwards used in the searching algorithm for a set of optimal free energy secondary structures. Any searching algorithm must include the topological relationships among the helical regions. If $l_i$, $u_i$, $t_i$ and $l_j$, $u_j$, $t_j$ are respectively the length, the initial and the final nucleotide positions of the regions $i$ and $j$, named so that $u_i \leq u_j$, there are the following possibilities (7):

1) $i$ includes $j$ if $u_i + l_i - 1 < u_j < t_j < t_i - l_i + 1$
2) $i$ excludes $j$ if $t_i < u_j$
Figure 5. Plot of CPU time vs. RNA sequence length N. Solid line is a quadratic fit, of equation: 
\[ \text{Time} = -0.46N + 0.0011N^2, \] 
with correlation coefficient 0.98. Different length natural as well as random generated RNA sequences are considered.

3) \( i \) crosses \( j \) if \( u_i \leq u_j \leq t_i \) and \( t_i < t_j \)
4) \( i \) overlaps \( j \) if \( u_i \leq u_j \leq u_i + l_i - 1 \) and \( t_j \leq t_i \) or \( t_i - l_i + 1 \leq t_j \leq t_i \)

In the assembly algorithm the helical regions with relationships 1 or 2 can coexist in the same structure but the cases 3 or 4 generate alternative possibilities.

Designing a searching algorithm to obtain the optimal free energy secondary structure there are two extreme possibilities: a) analysis of all the possible arrangements for the helical regions. The growth of the number of possible structures caused by the increase of the number of considered helical regions, is reported in the scheme in figure 2a. b) reaching the best free energy secondary structure by an unique assembly pathway. This case is shown schematically in figure 2b. Of course, in this case the problem is to find a suitable criteria able to locate this unique assembly pathway. This hypothetical solution would consume less computer time, if the aim of research is to obtain just the best free energy secondary structure.

The simplest criterion to obtain the case b), could be to choose the better free energy substructure at any level of the ‘growing tree’. But the pathway so obtained is dependent on the arrangement of \( n \) helical regions. Then the problem will be to find a suitable arrangement taking them into account (among \( n! \) possibilities) to reach the better structure. Even if we keep all the pathways which improve energetically at any level of the growing tree, we are not sure to reach the better structure. In fact local maxima along the pathways for the choosen arrangement of helical regions may be present. To overcome these difficulties we found effective an arrangement based on the increasing stacking free energy contribution and keeping of an up-to-date list of substructures based on the following criteria: 1) if the considered new helical region satisfies only the 1) and 2) topological conditions
Figure 6. Mean percentage of experimental data satisfied (□) and mean total energy (•) of a set of 50 structures vs. the experimental coefficient variation. The sequence considered is the Serratia marcescens trp operon (23), chosen for its abundance in enzymatic data.

For its characteristics, this method represents a compromise between the two extreme possibilities seen above (see figure 2c).

The number of structures so obtained can be very large for long RNA sequences. Then storing a fixed number of the better free energy secondary structures may be useful.

We introduced an approximate but more rapid estimate of the free energy terms to reduce the computational time requirement of the process. The criteria used are: 1) the stacking energy terms of all the helical regions are obtained by Salser’s parameters and stored before the searching algorithm. These values are used when the total energy of a structure is calculated. Then the total stacking term is only defective of the stackings arising from the eventual bulge formations. 2) the loops contribution is calculated using two terms,
Figure 7. The lowest free energy secondary structures of yeast tRNA\textsuperscript{Phe}.

one for the hairpins and the other for interior and bulge loops. They were obtained averaging the suitable terms of Salser's parameters, then neglecting the closing bases effect.

We verified the correctness of these criteria calculating the regression line between the approximate energy values and the exactly evaluated ones. An example of this check is reported in figure 4.

The assembly process is the most expensive CPU time in our algorithm. Some cases are reported in figure 5 in order to show the time vs. length N dependence. A quadratic law is found with a correlation coefficient greater than 0.98.
Figure 8. Secondary structure of the Tetrahymena thermophila IVS rRNA. *splicing sites. ○ addition site at 3' end forming circular IVS with release of 1–15 fragment. The base blocks phylogenetically identified are indicate following the convention used by Cech. The arrows indicate possible tertiary interactions.

*Introducing the experimental data in the algorithm.*

Experimental data, as enzymatic cuts or chemical modifications, may be introduced in our algorithm. However, there is not a complete reliability on these data. In fact, the initial
attack may lead to some unfolding or modification of the RNA structure. This is especially true in large molecules where the number of potential attack sites is very large (22). Moreover the experimental data may be in contrast with the formation of the more free energy favorable structures. Therefore we did not use an ‘a priori’ modification of the list of the helical regions, but we preferred respecting the majority of the experimental data in an adjustable arrangement with the free energy minimization criterion. It was carried out assigning a value \( s(m) \) depending on the experimental evidence of single (s positive) or double strand (s negative) to any base. Intermediate values may be assigned to \( s(m) \) if the experimental evidences present a different degree of confidence. We add this term to the free energy contributions. It can be weighted by the user and its reference value is the main free energy stacking for the sequence considered. It allows us to gradually reach the preferred compromise between the experimental data and the minimal free energy requirement. This method allows also some structural proposals about hypothesis of molecular mechanisms, if these can be explicitated as single or double strand evidences (for example a tract of RNA sequence not forming secondary structure because engaged with a protein interaction).

It seemed suitable to us to separate the experimental contribution \( S_{\text{tot}} \) in two terms, one regarding the double stranded \( (S_{\text{btot}}) \) and the other the single stranded regions of the structure \( (S_{\text{ltot}}) \):

\[
S_{\text{tot}} = S_{\text{btot}} + S_{\text{ltot}}
\]

and we can write:

\[
S_{\text{btot}} = \sum_{i=1}^{N_h} S_b(i)
\]

where \( N_h \) is the total number of the helical regions in the structure and \( S_b(i) \) is the experimental contribution for the region \( i \). We obtain by substitution of \( s \):

\[
S_{\text{btot}} = \sum_{i=1}^{N_h} \sum_{j=0}^{l(i)-1} S[j+k(i)]
\]

where \( l(i) \) and \( k(i) \) are respectively the length and the starting position of the region \( i \). The \( S_{\text{ltot}} \) can be developed in a likewise fashion:

\[
S_{\text{ltot}} = \sum_{i=1}^{N_l} S_l(i)
\]

where \( N_l \) is the total number of loops. \( S_{\text{ltot}} \) can be obtained subtracting from the sequence considered as a unique loop, the total stacking contribution \( S_{\text{btot}} \):

\[
S_{\text{tot}} = -\sum_{m=1}^{N} s(m) + S_{\text{btot}} = -\sum_{m=1}^{N} s(m) + \sum_{i=1}^{N_h} \sum_{j=0}^{l(i)-1} s[j+k(i)]
\]

The signs are adjusted so to have a negative contribution (in agreement with the free energy) if the experimental data are satisfied. Finally:
\[ S_{\text{hit}} = - \sum_{m=1}^{N} s(m) + 2 \sum_{i=1}^{N_b} \sum_{j=0}^{l(i)-1} s[j+k(i)] \]

This approach allows us to rapidly calculate the experimental contribution. In fact the first addendum depends only from the sequence and can be calculated once for all. The second term is a sum of contributions of the helical regions and these can be calculated before the search too.

We evaluated the effect of weighting the experimental contribution, studying the trend of the percentage of satisfied experimental data vs. the weight coefficient. Gradually forcing the coefficient results in rapidly reaching a high degree of satisfied data, as can be seen in figure 6.

**Final energy calculation and refinement.**

The energy used in the searching algorithm represents only an approximation and then the calculation of the exact free energy of formation of the select structures is the last step of the program. The program disrupts the terminal base paired of each helical region, and it checks the free energy. If there is a decrease of the total free energy of the structure, the process is continued, eventually until the complete break of the helix.

**RESULTS**

We checked the reliability of the algorithm by its application on some RNA sequences of different length. They were chosen because already studied and then useful for a comparison. In particular tRNA\textsuperscript{Phe} was chosen because its tertiary and secondary structure is experimentally known (24–26).

The tRNA\textsuperscript{Phe} four structures at minimum free energy are shown in figure 7. They are different in the folding. As can be seen the standard cloverleaf structure has not the lowest free energy. It may be expected because the tertiary interactions are neglected. In this case obtaining a set of different structures avoids the loss of biologically significant ones. This is a significative advantage of our method in respect to those which found only the lowest free energy structure.

We also folded the sequence of 55 bases from R17 viral RNA studied by Tinoco et al. (4) and the sequence of 108 bases of the primary transcript of yeast tRNA\textsuperscript{Tyr}, studied by De Robertis et al. (27). We retrieved the structures already proposed.

Finally, we studied the 414 nucleotides intervening sequence (IVS) of Tetrahymena thermophila rRNA. In the set of structures we obtained, it is also present that proposed by Cech et al. (28), characterized by an energy of $-119.89$ Kcal/mole. The authors obtained a different value using another way to calculate the energy. Structures with better free energy values are present in our set. The optimal structure has an energy of $-125.04$ Kcal/mole (not shown). Should be pointed out that one structure ($\Delta G = -120.67$ Kcal/mole) is also present, characterized by splicing sites closer than in the proposal of Cech et al. (see figure 8).

**CONCLUSIONS**

The computer time and memory occupancy requirements of our method depend on the amount of information saved. For example using a sequence 2000 bases long and saving the 100 better minimum free energy structures, require 45 CPU time using VAX 8530, running under VMS v. 4.7 and a computer program written in FORTRAN 77. Consequently
It is possible to search for optimal structures without using artificial constraints as the fragmentation (29), and then long range interactions can also be found.

It is possible to introduce with minor changes in our program the search for structures with pseudoknots. In this case the difficulties are mainly: a) the evaluation of the energy because of the lack of a complete set of experimental data; b) the increase of CPU time because of the rapidly growth of possible structures.

ACKNOWLEDGEMENTS
This work was supported by Ministero Pubblica Istruzione Grant 86 20902 225.

REFERENCES
1. Zaug, A.J., Been, M.D. and Cech T.R. (1986) Nature 324, 429–433
2. Eperon, L.P., Estibeiro, J.P. and Eperon, I.C. (1986) Nature 324, 280–282
3. Landick, R. and Yanofsky, C. (1984) J. Biol. Chem. 259, 11550–11555
4. Tinoco, I.Jr., Uhelenbeck, O.C. and Levine, M.D. (1971) Nature 230, 362–367
5. Pipas, J.M. and McMahon, J.E. (1975) Proc. Natl. Acad. Sci. 72, 2017–2021
6. Waterman, M.S. and Smith, T.F. (1978) Mathematical Biosciences 42, 257–266
7. Studnicka, G.M., Rohn, G.M., Cummings, I.W. and Salser, W.A. (1978) Nucleic Acid Research 5, 3365–3387
8. Nussinov, R. and Jacobson, A. (1980) Proc. Natl. Acad. Sci. 77, 6309–6313
9. Zuker, M. and Stiegler, P. (1981) Nucleic Acids Research 9, 133–148
10. Dumas, J.P. and Ninio, J. (1982) Nucleic Acids Research 10, 197–206
11. Quigley, G.J., Gehrke, L., Roth, D.A. and Auron, P.E. (1984) Nucleic Acids Research 12, 347–366
12. Jacobson, A.B., Good, L., Simonetti, J. and Zuker, M. (1984) Nucleic Acids Research 12, 45–52
13. Martinez, H. (1984) Nucleic Acids Research 12, 323–334
14. Jungck, J.R. and Friedman, R.M. (1984) Bull. Math. Biol. 46, 699–744
15. Gouy, M., Marliere, P., Papanicolaou, C. and Ninio, J. (1985) Biochimie 67, 523–531
16. Gouy, M. (1987) In Bishop, M.J. and Rawlings, C.J. (eds). Nucleic Acid and Protein Sequence Analysis. IRL PRESS, Oxford, pp. 259–284.
17. Williams, A.L.Jr. and Tinoco I.Jr. (1986) Nucleic Acids Research 14, 299–315
18. Salser, W. (1977) Cold Spring Harbor Symp. Quant. Biol. XLII, 987–1004
19. Freier, S.M., Kierzek, R., Jaeger, J.A., Sugimoto, N., Caruthers, M.H., Neilton, T. and Turner, D.H. (1986) Proc. Natl. Acad. Sci. USA 83, 9373–9377
20. Jacobson, H. and Stockmayer, W.H. (1950) J. Chem. Phys. 18, 1600–1606
21. Bracewell, R. (1965) The Fourier Transform and Its Applications. McGraw-Hill. New York. pp. 24–50
22. Noller, H.F. and Woese, C.R. (1981) Science 212, 403–411
23. Koruda, M.I. and Yanofsky C. (1984) J. Biol. Chem. 259, 12838–12843
24. Robertus, J.D., Ladner, J.E., Finch, J.T., Rhodes, D., Brown, R.S., Clark, B.F.C. and Klug, A. (1974) Nature 250, 546–551
25. Kim, S.-H., Suddath, F.L., Quigley, G.J., McPherson, A., Sussman, J.L., Wang, A.H.J Seeman N.C. and Rich, A. (1974) Science 185, 435–440
26. Quigley, G.J. and Rich, A. (1976) Science 194, 796–805
27. De Robertis, E.M. and Olson, M.V. (1979) Nature 278, 137–143
28. 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
29. Nussinov, R., Tinoco, I.Jr. and Jacobson, A.B. (1982) Nucleic Acids Research 10, 351–363