RNA Secondary Structure: A Complete Mathematical Analysis

M. S. WATERMAN AND T. F. SMITH
Los Alamos Scientific Laboratory of the University of California, Los Alamos, New Mexico 87545

Received 9 August 1978; revised 25 August 1978

ABSTRACT

Using a rigorous mathematical analysis, the prediction of RNA secondary structure as a function of free energy is obtained. The iterative method effectively allows a search over the entire configuration space of the RNA molecule not possible by earlier methods. The approach also allows for the direct inclusion of the nearest neighbor or stacking energies.

INTRODUCTION

The prediction of RNA secondary structure has been carried out by numerous authors [1–6]. The general approach has been to search for configurations of maximum base pairing or of minimum free energy. There are two basic problems encountered in these approaches. First, the entire RNA configuration space over which the search is to be performed is extremely large, and, until recently [7], no systematic method of searching the entire space has been proposed. The second problem is the assignment of proper free energies to the various substructural components. While considerable theoretical work has been carried out [8, 9], the most useful free energy data have been extrapolated from experiments with various oligoribonucleotides [10–13]. Such extrapolations, as noted by DeLisi and Crothers [3], often ignore such essentials as the nearest neighbor or stacking energies and lead to the imposition of somewhat *ad hoc* assignments of free energy to the configurations examined. There has recently been considerable work on the tertiary structure of some nucleic acids, in particular in comparisons with the x-ray data on various tRNAs. However, it should be noted that constraints arising from the most probable secondary structure base pairing are normally imposed on the tertiary structure considerations. This is analogous to the methods of predicting protein tertiary structure by starting with the statistics of forming helical and nonhelical regions.

 0025-5564/78/100257+9$2.25

In the present study the first problem is solved. This is accomplished through an iterative definition of all secondary structures and the extension of the sequence metric algorithms of Sellers [14]. The initial steps are based on the work of Needleman and Wunsch [15] and Tinoco et al. [2]. These ideas lead to the calculation of a minimum "distance" between segments of a RNA sequence, where "distance" is measured in free energy. The most probable secondary structure is then assumed to be the configuration having the minimum sum of all such aligned "distances."

SECONDARY STRUCTURE ENUMERATION

Modifying the approach of Tinoco et al. [2], define the base pairing matrix $P=(p_{ij})$, for a given RNA sequence $\mathbf{s}=s_1s_2\ldots s_n$ (and the reversed order sequence $\mathbf{s}'=s_ns_{n-1}\ldots s_1$) by $p_{ij}=1$ if s_i and s_j can form a bond and $p_{ij}=0$ otherwise. (the bonds are A—U, G—C, and sometimes G—U.)

A secondary structure for $\mathbf{s}$ is a configuration of the sequence $s_1s_2\ldots s_n$ with two properties: (i) Each point can be bonded to at most one other point. (ii) If s_i and s_j are bonded, then any bonding of s_k $(i<k<j)$ must be with points between i and j. It has been shown [6] that this definition includes all possible substructures (such as hairpins, helicies, bulges, tails, and interior loops). This definition does not include the B_{III} structure[1] proposed by Richards [6]. This is due in part to the standard definition of secondary structure [2, 9] as the primary folding conformation induced by Watson-Crick base pairing. While there is evidence of non-Watson-Crick base pairs [16] in such nucleic acids as the tRNAs, these are considered part of the tertiary structure resulting from the folding of, and consequential interactions between, secondary structural components.

The total number of structures having $i+1$ bonded pairs for a sequence $n+1$ long is given by a recursion relation. Let $N_{l,n}^{i}$ be the number of secondary structures containing exactly i bonded pairs formed on the subsequence $s_ls_{l+1}\ldots s_n$. Then

$$N_{l,n+1}^{i+1}=N_{l,n}^{i+1}+\sum_{j=l}^{n-m}\sum_{k=0}^{i}N_{l,j-1}^{k}N_{j+1,n}^{i-k}p_{j,n+1}, \tag{1}$$

where all hairpin loops have at least m bases. The equation follows from the fact that s_{n+1} is either bonded or not bonded. If s_{n+1} is not bonded, then

[1]There is no known recent experimental evidence for such structure involving the standard Watson-Crick pairing; this may be due to the fact that such structures would, for paired sequences of five or six, result in thermodynamically improbable knots.

there are $N_{l,n}^{i+1}$ structures of interest. Otherwise, $n+1$ is bonded to some j, $l \leqslant j \leqslant n-m$, and if k bonds are formed in $s_l \ldots s_{j-1}$, then $i-k$ must be formed in $s_{j+1} \ldots s_n$. The definition of secondary structure implies that any combination of a k bonded structure with an $i-k$ bonded structure gives a secondary structure. Thus $N_{l,n+1}^{i+1}$ satisfies Eq. (1).

The only sterical constraint in Eq. (1) is that the hairpin loop size must be at least m. It is possible to modify Eq. (1) so that no helices of length one are allowed. This has been done and the recursion applied to real RNA sequences. For many real RNA sequences of length forty there are over 10^6 structures, hundreds of which may have maximal base pairing.

MINIMUM FREE ENERGY STRUCTURES

Tinoco et al. [2] noted that their base pairing matrix contains the pairings for all structures, and that some set of non-overlapping antidiagonal strings of 1's must represent the optimal structure. For short sequences it appeared that the maximal pairing configuration could be deduced by direct observation [2] or from some counting schemes [17, 18]. There have been a number of attempts to construct an algorithm to search for the optimal combination of such antidiagonal strings [4, 17, 18, 23], none of which search the entire configuration space.

The problem is analogous to finding the optimal matching alignment between two evolutionary sequences. The solution to that evolutionary distance problem was proposed by Sellers [14] and generalized by Waterman et al. [19]. To help clarify the relationship between the two problems, it is useful to note that regions of homology between different sequences are analogous to complementary helical regions, nonhomologous regions are analogous to noncomplementary internal loop regions, and deletions/insertions are analogous to bulges. It is also helpful to recall that finding the maximum homology between evolutionary sequences is equivalent to finding the minimum mutational distance between them. As noted above, in this work a minimum "distance" measured in free energy is calculated between all subsequences.

The very large number of possible structures makes the RNA secondary structure problem considerably more difficult than the evolutionary sequence problem. However, an iterative algorithm has been constructed [7] which builds up complex structures from simpler ones. Before describing this algorithm, the free energy functions associated with various secondary structural components must be defined. (Recall the indexing on $\mathbf{s}' = s_n s_{n-1} \ldots s_1$.) We define

$\alpha_{ij} = \Delta G$ (free energy change) of binding of the ith element of the sequence $\mathbf{s}$ with the jth of $\mathbf{s}'$;

$\eta_{ij} = \Delta G$ resulting from nearest neighbor interaction between base pairs $i-1,j-1$ and i,j;
$\beta_j = \Delta G$ of a bulge j bases long;
$\gamma_{ij} = \Delta G$ of an interior loop of lengths i and j;
$\xi_{ij} = \Delta G$ of an end loop $n-i-j$ bases long due to the pairing of bases i and j;
$\tau_i = \Delta G$ of a free end or tail of length i.

The total free energy change of a secondary structure is defined to be the sum of the ΔG's associated with these substructures. This can be accomplished [7] by constructing an f matrix such that each element f_{ij} represents the free energy of formation of the i,j bound pair plus the free energy of that secondary structure having the minimum free energy among all substructures formed from the $i-1$ subsequence of s and the $j-1$ subsequence of s'. The elements of f_{ij} are undefined (plus infinity) for all i,j such that the ith base in s cannot form a Watson-Crick pair with the jth base in s' ($p_{ij} = 0$).

For the case when $p_{ij} = 1$, f_{ij} is defined as

$$f_{ij} = \alpha_{ij} + \min\Big\{ f_{i-1,j-1} + \eta_{ij},\ \min_{k>0}\{f_{i-k-1,j-1} + \beta_k\},\ \min_{k>0}\{f_{i-1,j-k-1} + \beta_k\},\ \min_{\substack{k>0\\ l>0}}\{f_{i-k-1,j-l-1} + \gamma_{k,l}\},\ 0\Big\}. \tag{2}$$

The free energy change of the best single loop secondary structure is calculated by

$$F_{1,n} = \min_{\substack{1<j<n\\ 1<j<n}}\{f_{ij} + \zeta_{ij}\}, \tag{3}$$

which includes the additional free energy associated with the end loops. Figure 1 shows the values of f_{ij} for a simple illustrative example using the component ΔG's given in Table 1, Column A. The insert in Fig. 1 shows the spatial relationship between previous elements of f_{ij} and a given element for a finite value of α. A complete, mathematical proof that this procedure obtains the minimum is given by Waterman [7].

To calculate more complex minimum free energy secondary structures, the single loop F_{ij} must be obtained for all viable subsequences $1 < i < j < n$. Then the bulges, interior loops, and tails must be examined for the possibil-

TABLE 1
Substructural Component Free Energies
in kcal at 23°C

A[a]	B[b]
$\alpha_{ij} = -1.0$[c]	$\alpha_{AU} = -0.25$
$\eta_{ij} = -1.0$[c]	$\alpha_{GC} = -1.4$
	$\alpha_{GU} = +1.9$[d]
$\beta_l = 1.0 + 0.5l$	$\eta_{ij} = -1.0$[c]
$\gamma_{kl} = 1.0 + 0.5(l+k)$	$\beta_l = 1.0 + 0.3l$
$\xi_{kl} = 1.0 + 0.5(n-k-l)$	$\gamma_{lk} = 1.5 + 0.2(l+k)$
$\Delta G_{join} = 0.0$	$\xi_{lk} = 3.05 + 0.1(n-k-l)$
	$\Delta G_{join} = 0.5 + 0.3l$

[a]Values for investigative use only, in the construction of Fig. 1. Such values allow the illustration all the major properties of the proposed algorithm.

[b]Values extrapolated from experimental values. The values for α_{AU}, α_{GC}, and η were chosen to given the standard values of -1.25 and -2.4 kcal in the limit of long bound chains.

[c]For all $i,j = A, C, G, U$.

[d]The value of α_{GU} was set equal to the interior loop value. This results in a ΔG for a G—U pair in the interior of a helical region of only -0.1 kcal. The values for the bulge, interior loop, and end loop contributions are linearizations of those due to DeLisi and Crothers [3], and as such have a limited argument range.

ity that these subsequences may form bonded single loop structures. It is not entirely easy to calculate the proper free energy changes for the addition of these structures. This is because f_{ij} gives tails weight zero, when they could become bulges or joins in the new composite structures. It is useful to calculate F_{ij} only for substructures such that s_i and s'_j are bonded. Waterman [7], using these restricted F_{ij}, was able to iterate and calculate minimum free energy structures of arbitrary complexity.

On the computational side, the elements of the various matrices fortunately have to be calculated only when p_{ij} is nonzero. Yet, even on large modern computers, one is limited in secondary structure calculations using the above algorithm to sequences of two hundred or less.

The secondary structure having the calculated minimum free energy change is obtained from a traceback procedure. Having found the f_{ij} which gives the single loop minimum F_{1n} in Eq. (3), one must trace back to find which terms in Eq. (2) and thus which structural component contributed at

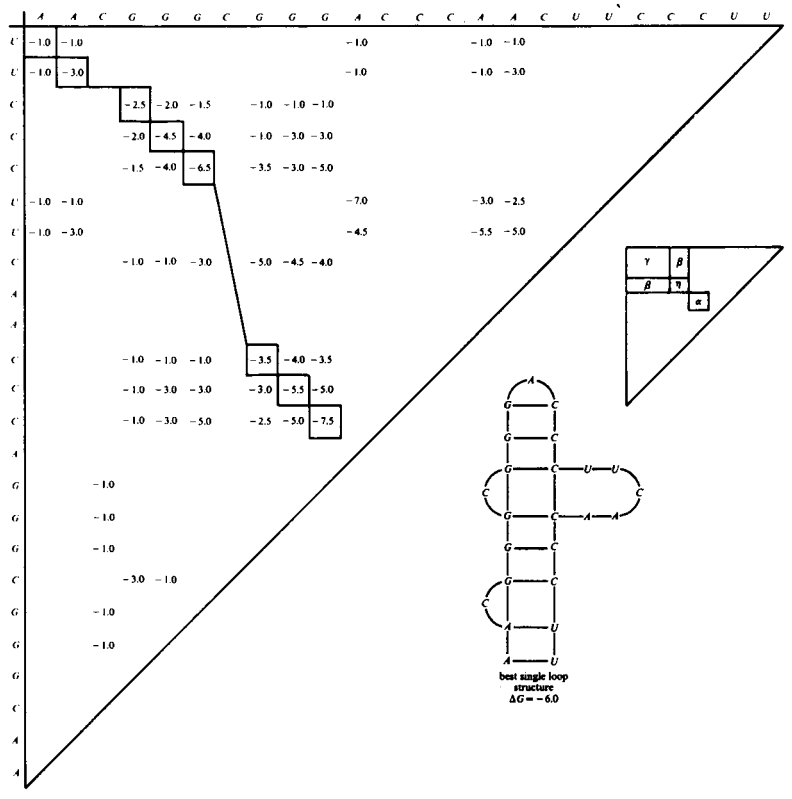


FIG. 1. The f_{ij} matrix calculated for the given test sequence using the component free energies in Table 1, column A. The first order structure obtained is shown, along with the traceback path through the f_{ij} matrix. The small triangular inset illustrates the relationship between the positions in the matrix and the contributions of the various component free energies to the calculation of each term with a nonzero α in Eq. (2). For this test sequence there are, for example, 1344 eight bound pair, 89 nine pair, and only 1 ten pair structures as enumerated by Eq. (1).

each step. There is no guarantee that the minimum free energy structure is unique, but a traceback procedure can locate all such structures.

As noted in the introduction the application of even a mathematically correct algorithm requires a knowledge of component free energies. While these have been extracted from experimental data by many workers [3, 2, 10–13], the distinction between secondary and tertiary contributions is somewhat arbitrary. Yet considerable success has been had in understand-

ing the quantitative nature of Watson-Crick pairing and the weaker pairings [16] which probably are formed only later [19]. It therefore seems essential that the mathematics of the secondary structure algorithm be independent of the continuing investigation of the various free energy contributions.

APPLICATIONS

The general algorithm has been applied through second order on the R17 viral RNA fifty-five base subsequence originally studied by Tinoco et al. [2]. The predicted structure is the single loop structure identical to that proposed by Tinoco with a free energy only slightly lower. This is due of course to the fact that the free energies in column B of Table 1 are not identical to those used by Tinoco et al. [2]. As a check on both the validity of the free energies used and the physical-chemical nature of the algorithm, the Phe tRNA from yeast [18] with known secondary structure was investigated. The minimum free energy single loop structure predicted by Eq. (3) is shown in Fig. 2(a). This first order structure contains the anticodon loop and stem as well as the acceptor stem. The midregion predicted pairing is of low thermostability (short interspersed helices) This structure is highly suggestive of an intermediate melting sequence structure proposed by Crothers [21]. In Fig. 2(b), the predicted minimum free energy second order structure shows the familiar four stem, three loop form. It should be noted that for the substructure ΔG's in column B of Table 1, used for this calculation, the D loop is favored over an open bulge by less than 2 kcal. This also supports the melting model of Crothers in which the stability of the D loop is a function of the folded three dimensional conformation. Thus, by comparing the best single loop structure with the best higher order structure, one may be able to investigate the dynamics of the tRNA melting stages proposed in [21], [22].

The proposed method of investigating minimum free energy RNA structures not only rests on a rigorous mathematical foundation, but with the use of experimental component ΔG's yields predictions compatible with the best known structural configurations. This analysis will allow a major investigative effort of the numerous RNA sequences now available.

Finally, the computation time limitations as a function of sequence length may in fact be related to nature's own problem of searching for a global free energy minimum for large RNA molecules. At physiological ionic conditions and temperatures one expects many local minima of less than -20 kcal. This suggests that for large RNA the stable structures may be in part kinetically determined. In conclusion, we note that a computer program for this algorithm is available from the first author [24].

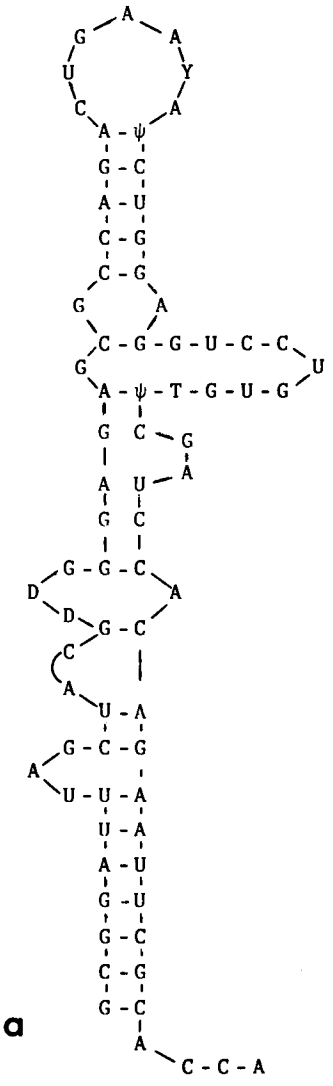
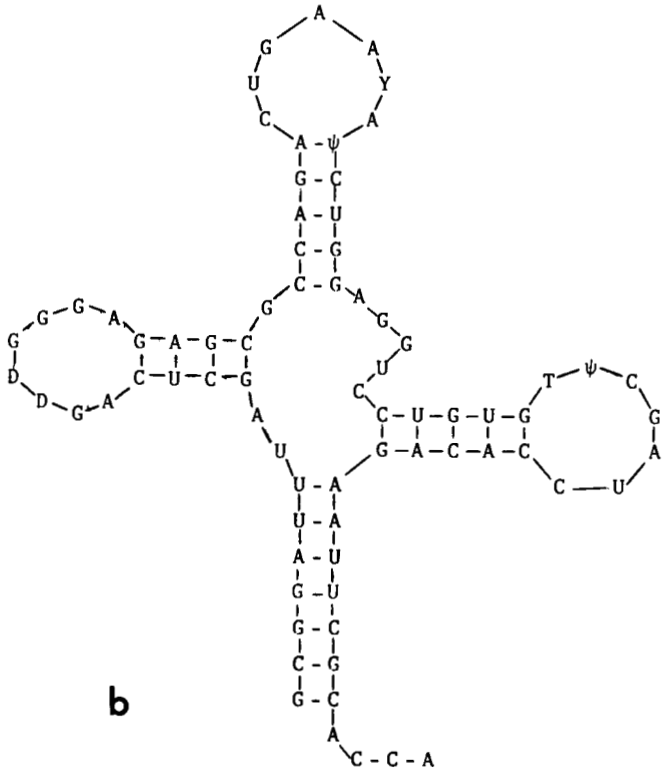


FIG. 2. Minimum ΔG structures obtained for yeast Phe tRNA. (a) The best first order single loop structure predicted from Eq. 3, having a ΔG of -16.95 kcal, with 42 paired bases. (b) The best second order structure found, having a ΔG of -22.15 kcal, with 42 paired bases.



b

REFERENCES

1 J. R. Fresco, B. M. Alberts, and P. Doty, *Nature* 188:98 (1960).
2 I. Tinoco, Jr., O. C. Ahlenbeck, and M. D. Levine, *Nature* 230:362 (1971).
3 C. DeLisi and D. M. Crothers, *Proc. Nat. Acad. Sci. U.S.A.* 68:2682 (1971).
4 R. Lapidus, R. Bernard, and R. Hepperle, *J. Theoret. Biol.* 64:587 (1977).
5 G. E. Fox and C. R. Woese, *Nature* 256:507 (1975).
6 E. G. Richards, *Eur. J. Biochem.* 10:36 (1969).
7 M. S. Waterman, *Advances in Math.*, Supplementary Studies Vol. I, Studies in Foundations and Combinatorics, 1978, p. 167.
8 Mitiko Gō, *J. Phys. Soc. Japan* 32:597 (1967).
9 D. Poland and H. A. Scheraga, *Theory of Helix-Coil Transitions in Biopolymers*, Academic, 1970.

10 F. H. Martin, O. C. Uhlenbeck, and P. Doty, *J. Mol. Biol.* 57:201 (1971).
11 P. N. Borer, B. Dengler, and I. Tinoco, Jr., *J. Mol. Biol.* 86:843 (1974).
12 J. Gralla and D. Crothers, *J. Mol. Biol.* 78:301 (1973).
13 E. Wickstrom and I. Tinoco, Jr., *Biopolymers* 13:2367 (1974).
14 P. H. Sellers, *SIAM J. Appl. Math.* 26:787 (1974).
15 S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.* 48:443 (1970).
16 R. G. Shulman, C. W. Hilbeus, D. L. Miller, S. K. Yang, and D. Soll, Assignment of the hydrogen-bonded proton resonances in $tRNA^{Glu}$ by sequential melting, in *Structural and Conformation of Nucleic Acids and Protein–Nucleic Acid Interactions* (M. Sundaralingam and S. T. Rao, Eds.), Univ. Park Press, Baltimore, 1975, p. 149.
17 D. Klambt and O. Richter, *J. Theor. Biol.* 58:319 (1976).
18 R. Lapidus, B. Rosen, and R. Hepperle, *J. Theor. Biol.* 64:587 (1977).
19 M. S. Waterman, T. F. Smith, and W. A. Beyer, *Advances in Math.* 20:367 (1976).
20 G. J. Quigley and A. Rich, *Science* 194:796 (1976).
21 D. M. Crothers, RNA structure and structural changes, in *Structure and Conformation of Nucleic Acids and Protein–Nucleic Acid Interactions* (M. Sundaralingam and S. T. Rao, Eds.), Univ. Park Press, Baltimore, 1975, p. 215.
22 C. DeLisi, *Biopolymers* 12:1713 (1973).
23 J. M. Pipas and J. E. McMahon, *Proc. Nat. Acad. Sci. U.S.A.* 72:2017 (1975).
24 M. S. Waterman and T. F. Smith, Los Alamos Scientific Laboratory Publication LA-7153-MS, 1978.