*Proc. Nat. Acad. Sci. USA*
Vol. 68, No. 11, pp. 2682–2685, November 1971

# Prediction of RNA Secondary Structure

**(theoretical/free-energy estimates/oligonucleotides)**

CHARLES DELISI AND DONALD M. CROTHERS

Department of Chemistry, Yale University, New Haven, Connecticut 06520

*Communicated by Kenneth B. Wiberg, September 3, 1971*

**ABSTRACT Calculations of the free energy required to close single-strand loops by formation of a base pair in double-helical nucleic acids are reported. These results can be used to estimate the free energy of particular secondary structures for a given RNA molecule under conditions of high-salt concentration.**

A problem fundamental to physical biochemistry is the extent to which the spatial and orientational constraint imposed by catenation on two chemically reactive groups affects the equilibrium constant for the reaction (1–6). In particular, this problem presents itself in connection with RNA secondary structure. Consider the possible structures for the two sites of attachment to ribosomes of R17 RNA (7) shown in Fig. 1, which illustrates four conformations in which a sequence of nucleotides can be found: I. the intact helix, II. the internal loop, III. the hairpin loop, and IV. the bulge. Two bases that pair to close a hairpin loop, for example, are connected by the backbone chain, and this constraint must influence the free energy of the base-pairing reaction.

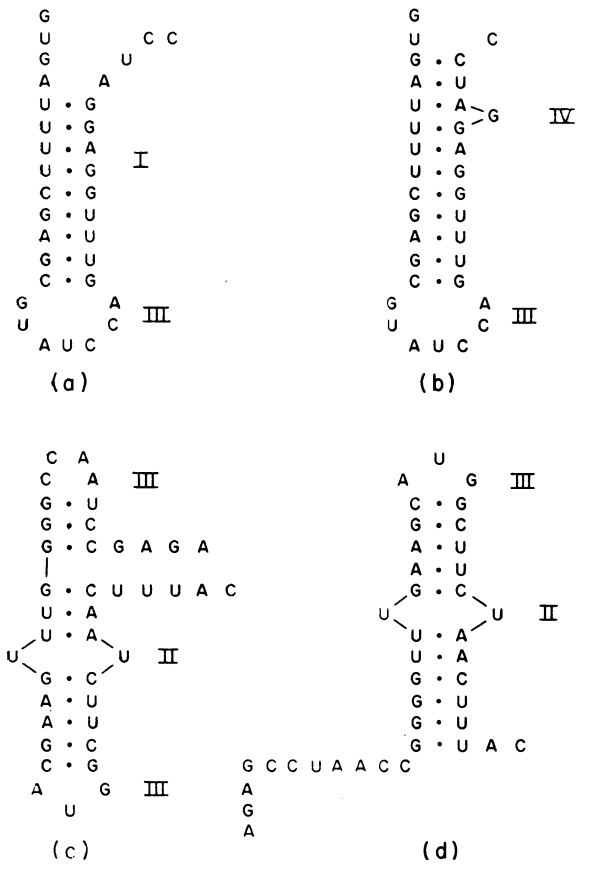


FIG. 1. (*a*) and (*b*) are two possible secondary structures of the initiation sequence for the A protein in R17 viral RNA (7), and and (*c*) and (*d*) are possible secondary structures of the initiation sequence for the coat protein from the same RNA (7). Regions labeled I–IV are, respectively: intact helix, internal loop, hairpin loop, and bulge loop.

Tinoco, Uhlenbeck, and Levine (8) have made substantial progress toward evaluating the free energy of RNA secondary structures in terms of the basic structural types I–IV shown in Fig. 1. Each such conformation is assumed to contribute an additive free-energy term to the molecular free-energy, a supposition that can be valid only at high-salt concentration, where the electrostatic interactions are of short range (4). A main weakness in their approach is the calculation of the free energy of single-strand loops formed in structures of type II–IV. In particular, they set for the free energy, $\Delta G$, of loop closure

$$\Delta G = -2.3\,RT\,[B - 1.5 \log (m + 1)] \qquad (1)$$

where $m$ is the number of unbonded bases and $B$ is a constant. Eq. (1) applies to a freely-jointed Gaussian chain (1); we earlier discussed its inadequacy for predicting loop free-energies (3).

We developed (3) a general formalism for computing ring-closure probabilities, using structural information about the backbone to calculate the distribution function for one base around the other. Our present purpose is to extend our earlier work by evaluating the angular part of the distribution and by taking account of the lack of spherical symmetry of the spatial part of the short-chain distribution function. On this basis, we calculate size-dependent closure free-energies for loops of types II–IV, and use these results to evaluate the free energies of typical RNA structures.

## THEORY

The free-energy increase on closure of the various types of loops could be evaluated if we knew the equilibrium constant, $K$, associated with each reaction. $K$ depends on the probability that the two reacting units are in the correct spatial arrangement relative to one another, as well as on the "intrinsic" equilibrium constant, $K_0$, for bond formation when the units are within the required spatial limits. Suppose that unit (*a*) (Fig. 2) is at the origin of a fixed coordinate system, and that unit (*b*), which has its own coordinate system rigidly attached, must be within a volume $\delta v$ at $\mathbf{R}^*$ in order to react. Reaction also requires that (*b*) is at an orientation specified by $\theta_1^*$,

$\theta_2^*$, and $\theta_3^*$ within the angular interval $\delta\omega\delta\alpha$. [$\delta\omega$ is a solid angle and $\delta\alpha$ is the rotational interval about the $z$ axis in the system of ($b$).] We can write that

$$K = S(\mathbf{R}^*)\ \delta v A(\theta_1^*, \theta_2^*, \theta_3^*)\delta\omega\delta\alpha K_0, \qquad (2)$$

where $S(\mathbf{R})$ and $A(\theta_1, \theta_2, \theta_3)$ are, respectively, the spatial and angular parts of the two-particle distribution function.

In general, $K_0$ cannot be calculated theoretically. Instead, we use an experimental value of $K$ for closing a loop of a given size, and calculate the variation in $K$ with loop size by computing the distribution functions $S$ and $A$ as functions of chain size.

## THE DISTRIBUTION FUNCTIONS

Finding adequate distribution functions is generally a formidable mathematical problem. For short chains, however, it can be accurately solved numerically by Monte Carlo (sampling) techniques (3). A model for the rotational freedom of the backbone bonds was chosen from crystallographic data (9). It has the feature that an exact calculation accurately predicts the experimentally determined radius of gyration of poly(U) (10). The model ignores changes in chain stiffness as the result of base stacking.

$S(R)$ was determined by calculation of the probability of particular positions $\mathbf{R}$ of the phosphorous atom in one nucleotide relative to the phosphorous in the other nucleotide of the base pair; $\mathbf{R}^*$ is the relative position in the double helix. For positions in which the phosphorous was within $\delta v$ around $\mathbf{R}^*$, the angular function $A$ was evaluated for the distribution of orientations of the 0 (5′) — P bond, relative to its orientation in the double helix. Note that rotation about the 0 (5′) — P bond is independent of chain length. Hence, we need not concern ourselves with the third angular variable, $\delta\alpha$. We also found that $A$ quickly approaches the value predicted for a uniform distribution of bond orientations ($^1/_4\ \pi$, when $\delta\alpha$ is neglected). The small deviations of $A$ from the uniform value were estimated for loops larger than seven nucleotides by interpolation from the linear plot of log $A$ versus the reciprocal of loop size, including the calculated values for smaller loops and the point at $A = {}^1/_4\ \pi$ for infinite loop size. (This plot was not linear, however, for small loops in the bulge defect.)

## LOOP-CLOSURE FREE ENERGIES

The free energy, $\Delta G$, for formation of the first base-pair that closes a loop can be expressed in terms of the experimental free-energy change, $\Delta G_{ref}$, for the reference reaction:

$$\Delta G = \Delta G_{ref} - RT \ln\left\{\frac{S^*A^*}{S^*_{ref}A^*_{ref}}\right\}, \qquad (3)$$

where $S^*$ and $A^*$ are the values of the distribution functions at the coordinates corresponding to base-pair formation.

As a reference reaction for hairpin- and internal-loop formation, we chose the formation of a hairpin loop of 4 unbonded bases by oligo(dT-dA) molecules (11). The equilibrium constant for formation of the first base-pair to close the loop has been estimated at $3 \times 10^{-3}$ (4, 11). Table 1 shows ring-closure free energies calculated on this basis. A notable feature of the results is the definite thermodynamic preference for internal loops over hairpin loops of the same size. The reason lies in chain stiffness: to nucleate a hairpin loop, a stiff chain must turn back on itself, whereas the two chains

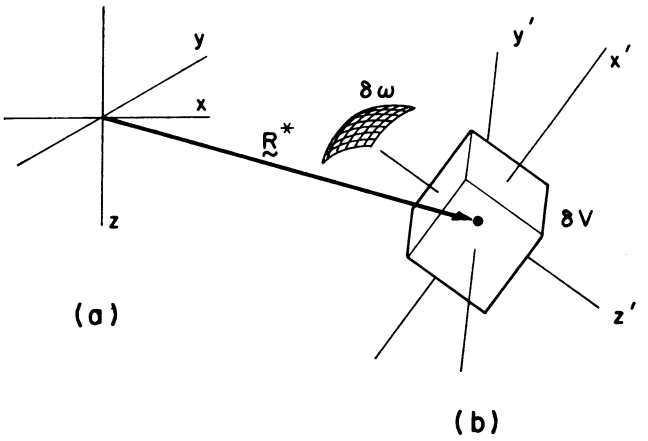


FIG. 2. The mutual orientation of two residues, (a) and (b), that are in position to be joined by a bond, including the permissible variation $\delta v$ and $\delta\omega$ in the spatial and angular positions of ($b$) in the coordinate system, fixed relative to ($a$). $\delta\alpha$ is the permissible rotation about the $z'$ axis.

in an internal loop can form a base pair even when both are in an extended conformation. Another interesting feature is the appearance of a most-favorable ring size (minimum free-energy) with five bases unbonded. Our earlier results (3) did not show this character because we neglected $A$ and assumed spherical symmetry of $S$.

The reference reaction for bulge formation must be different from that for the other two reactions, since one assumes in this case that the new base-pair stacks on the existing adjacent helix. According to a recent experimental estimate (Fink, T. R., and D. M. Crothers, in preparation) the free energy of formation of a base-pair adjacent to a bulge defect with one base looped out of the double helix is 2.8 kcal/mol, plus the free energy of formation of that base pair in the perfect helix. This latter term depends on whether the pair is A·U, G·C, or G·U; we assume values of −1.2 kcal, −2.4 kcal, and 0 kcal for each of these (8). Table 1 shows free energies for closure of various-sized bulge loops, calculated on this basis.

## EMPIRICAL TESTS OF THE THEORY

There are presently two experimental tests of our method of relating free-energy changes for different base-pairing reactions. One is to use the reference equilibrium constant for formation of the hairpin loop to calculate the bimolecular nucleation constant for formation of the first bond between two separate strands. We have described (3) the formalism for this comparison. Using our present values for the distribution functions $S^*$ and $A^*$, we calculate $\beta s = 1.7 \times 10^{-2}\mathrm{M}^{-1}$ ($\Delta G° = 2.5$ kcal at 25°C) for the bimolecular equilibrium constant, compared with a range of values from $4 \times 10^{-4}$ to $3 \times 10^{-3}\mathrm{M}^{-1}$ ($\Delta G° = 4.7$–$3.5$ kcal at 25°C) reported for oligomers that contain A·U base pairs (12). Considering the experimental inaccuracy in the equilibrium constant for both hairpin-helix nucleation and bimolecular nucleation, we consider the agreement between theory and experiment acceptable.

We can also use the theory to calculate the free energy of formation of a one-base bulge by using the base-pairing reaction in the perfect double helix as the reference reaction. On this basis, we calculate that the contribution of the bulge

TABLE 1. *Free energy of formation of the first base pair to close loops in RNA*

| | No. of bases unbonded | $S^*\delta v$ | $A^*\delta\omega$ | $S^*\delta v A^*\delta\omega$ | | $\Delta G$ (kcal/mol) | |
|---|---|---|---|---|---|---|---|
| 1. Hairpin loop ($\delta\omega/4\pi = 0.016$) | 4 | 0.037 | 0.040 | $1.48 \times 10^{-3}$ | | 3.43 | |
| | 5 | 0.05 | 0.033 | $1.65 \times 10^{-3}$ | | 3.37 | |
| | 6 | 0.04 | 0.030 | $1.2 \times 10^{-3}$ | | 3.6 | |
| | 7 | 0.03 | (0.026) | $0.78 \times 10^{-3}$ | | 3.8 | |
| | 8 | — | — | $0.62 \times 10^{-3}$ | | 3.9 | |
| | 9 | 0.02 | (0.023) | $0.46 \times 10^{-3}$ | | 4.1 | |
| 2. Internal loop ($\delta\omega/4\pi = 0.016$) (equal number of bases on each strand) | 2 | 0.471 | 0.027 | $1.9 \times 10^{-2}$ | | 1.9 | |
| | 4 | 0.33 | 0.024 | $1.2 \times 10^{-2}$ | | 2.2 | |
| | 6 | 0.20 | (0.022) | $6.6 \times 10^{-3}$ | | 2.5 | |
| | 8 | 0.10 | (0.021) | $3.2 \times 10^{-3}$ | | 3.0 | |
| | | | | | A·U | G·C | G·U |
| 3. Bulge loop ($\delta\omega/4\pi = 0.02$) | 1 | 0.13 | 0.18 | $2.3 \times 10^{-2}$ | 1.6 | 0.4 | 2.8 |
| | 2 | 0.04 | 0.06 | $2.4 \times 10^{-3}$ | 2.7 | 1.5 | 3.9 |
| | 4 | 0.021 | 0.018 | $3.2 \times 10^{-4}$ | 3.8 | 2.6 | 5.0 |
| | 6 | 0.013 | 0.019 | $2.5 \times 10^{-4}$ | 4.1 | 2.9 | 5.3 |
| | 8 | 0.008 | (0.019) | $1.5 \times 10^{-4}$ | 4.4 | 3.2 | 5.6 |

Free energies are calculated for $T = 25°C$ in high-salt concentration (about 1 M $Na^+$). $\delta v$ has a 0.6-nm radius. Values of $A^*\delta\omega$ in parentheses were determined from interpolation, as described in the text. Statistical error limits on $S^*\delta v$ and $A^*\delta\omega$ are 10% or less, corresponding to an error of less than 0.15 kcal in $\Delta G$.

loop is 1.7 kcal for one base and 3.0 kcal for two bases looped out, compared with the experimental estimate of 2.8 kcal for one base. Since we are pushing the theory severely by applying it to the perfect helix, we consider that this agreement is also reasonable.

## PREDICTION OF RNA SECONDARY STRUCTURE

One can in principle predict RNA secondary structure by showing that a particular base-pairing arrangement has minimum free energy (8). However, there remain serious limitations on our ability to assign accurate free energies to all the interactions. Tinoco *et al.* (8) discuss some of these difficulties, and since we presently see no strong reasons to differ from their compromise choice of the free energy contributed by helical base-pairs, we follow their procedure for the sake of consistency. This leads to the assignment of −2.4 kcal, −1.2 kcal, and 0 kcal for formation of G·C, A·U, and G·U pairs, respectively. We prefer, however, to account for the stabilization contributed by base pairs by counting stacking interactions—pairs at the end of a helix are involved in one, while those in the middle are involved in two. An A·U pair adjacent to a G·C pair contributes one G·C and one A·U stacking interaction, and the free-energy contribution is $-(2.4 + 1.2)/2$ kcal.

The following rules may be used to estimate the free energy of an RNA secondary structure at 25°C in about 1 M salt solution on the basis of our weighting functions for loops and bulges:

(*a*) For each continuous helix (which may include bulge defects) count the number of stacking interactions that involve each kind of base pair. Divide these numbers by two and multiply them by the free-energy contributions for each kind of base pair. Add together these contributions from double-helix formation.

(*b*) Count the number of bases in internal and hairpin loops, and add the corresponding free-energy term from Table 1.

(*c*) For bulge defects, add the number under the G·U column in Table 1, since the stacking interaction due to the actual adjacent base-pair has been included in rule (*a*).

For example, for conformation (*a*) in Fig. 1, there are 9 G·U, 4 A·U, and 3 G·C stacking interactions, and a hairpin loop of 7 bases. Hence, the free energy is $-(9 \times 0 + 4 \times 1.2 + 3 \times 2.4)/2 + 3.8$, or −2.2 kcal/mol, relative to the unbonded state. For conformation (*b*), there are 8 G·U, 8 A·U, and 4 G·C stacking interactions, a hairpin loop of 7 bases, and a bulge defect of one base. Hence, the free energy is $-(8 \times 0 + 9 \times 1.2 + 4 \times 2.4)/2 + 3.8 + 2.8$, or −3.0 kcal/mol. The two structures differ in free energy by only 0.8 kcal, and we therefore predict that both would be present in substantial amounts at conformational equilibrium. According to the rules given by Tinoco *et al.*, the two would differ by about 2.4 kcal/mol, again with (*b*) the favored form.

The results of this calculation should be sufficient to urge caution on those who would use such methods to predict RNA structure. In these molecules, one often finds several conformations with nearly equal predicted free-energy. Since the free-energy parameters in the calculation could be in error by as much as a kilocalorie, erroneous predictions may clearly be made unless one conformation is of decidely lower free-energy (5 kcal or more) than the others. The only way we can see to overcome this lack of precision in the calculation is to refine the parameters by means of extensive further experiments on the stability of oligonucleotides.

## COMPARISON WITH EARLIER RESULTS

According to our results, if other factors are equal the internal loop is favored over an equal-sized hairpin loop, which is in turn favored over the bulge loop. (For the bulge, the free-energy contribution from the loop itself is equal to the number under the G·U column in Table 1.) This progression has a reasonable physical basis, since the stiff nucleotide chain must return progressively closer to its point of origin to form a base pair when closing an internal-, hairpin-, or bulge-

loop, respectively. However, our ordering of stabilities is quite different from that assigned by Tinoco *et al.* (8), whose numbers make the bulge loop favored over internal- and hairpin-loops, the latter two being of equal stability. The free energy we assign to internal loops differs from their assignment by about 3 kcal; the difference is about 2 kcal for hairpin loops. The main source of this disagreement is that we have not given any weight to the results for melting high polymers in assigning free energies to small rings. In our view, the values for large rings (hundreds of nucleotides) cannot presently be correctly extrapolated to small rings. On the other hand, our free energies for the bulge loop are roughly the same as those of Tinoco *et al.*, since we made use of similar experimental information.

We should note parenthetically that our ordering of internal loops as preferable to bulge loops is consistent with the observation of the bulge loop in complexes of copolynucleotides that contain noncomplementary residues (13). For example, at stoichiometric equivalence between the *A* residues in poly(A,U) and the *U* residues in poly(U), more base pairs can be formed in structures containing bulge loops than in those containing internal loops. Application to such structures of the rules in the previous section leads to a clear prediction of preference for the bulge- over internal-loops, as long as the $T_m$ is not approached.

A second important difference between our results and those of Tinoco *et al.* is our observation of a generally stronger dependence of the free energy on loop size. (The exception is the free-energy minimum for the optimum hairpin-loop.) In general, closure free-energies for small rings depend more strongly on ring size than is predicted by Eq. (1) used by Tinoco *et al.* (8); that equation is evidently inappropriate to the present problem.

This work was supported by grants GB 8414 and GB 26535 from the National Science Foundation. C. DeL. was supported by a postdoctoral fellowship from the National Institutes of Health (GM 44137) and D. M. C. holds a Career Development Award (GM 19978) from the NIH.

1. Jacobson, H., and W. H. Stockmayer, *J. Chem. Phys.*, **18**, 1600 (1950).
2. Flory, P. J., and J. A. Semlyen, *J. Amer. Chem. Soc.*, **88**, 3209 (1966).
3. DeLisi, C., and D. M. Crothers, *Biopolymers*, in press.
4. DeLisi, C., and D. M. Crothers, *Biopolymers*, in press.
5. Crothers, D. M., and H. Metzger, *Immunochemistry*, in press.
6. Storm, D. R., and D. E. Koshland, Jr., *Proc. Nat. Acad. Sci. USA*, **66**, 445 (1970).
7. Steitz, J. A., *Nature (London)*, **224**, 957 (1969).
8. Tinoco, I., O. Uhlenbeck, and M. D. Levine, *Nature (London)*, **230**, 362 (1971).
9. Sundarlingam, M., *Biopolymers*, **7**, 821 (1969).
10. Inners, I. D., and G. Felsenfeld, *J. Mol. Biol.*, **50**, 373 (1970).
11. Scheffler, I., E. Elson, and R. L. Baldwin, *J. Mol. Biol.*, **36**, 291 (1968).
12. Craig, M. E., D. M. Crothers, and P. Doty, *J. Mol. Biol.*, in press.
13. Fresco, J. R., and B. M. Alberts, *Proc. Nat. Acad. Sci. USA* **46**, 311 (1960).