PHYSICAL REVIEW E VOLUME 52, NUMBER 2 AUGUST 1995

Folding RNA with the minimal loss of entropy

Ariel Fernández,[1,2] Hugo Arias,[2] and Diego Guerín[3]
[1]The Frick Laboratory, Princeton University, Princeton, New Jersey 08544
[2]Instituto de Investigaciones Bioquímicas—INIBIBB, Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional del Sur, Bahía Blanca 8000, Argentina
[3]Departamento de Física, Universidad Nacional del Sur, Bahía Blanca 8000, Argentina

(Received 14 April 1995)

The principle of sequential minimization of entropy loss (SMEL) is introduced and justified within the context of biopolymer folding *in vitro*. This principle implies that at each stage in the dominant folding pathway, the conformational entropy loss associated with loop closure, ΔS_{loop}, is minimized while the number of effective contacts is maximized. The applicability of the SMEL principle is contingent upon a rigorous and reliable derivation of the contribution ΔS_{loop}. This derivation is carried out in this work for RNA by taking into account the orientational restrictions associated with the self-energy of charged phosphate moieties within a loop. The predictive potential of the principle is revealed by showing that the theory reproduces the biologically competent secondary structures of specific catalytically competent RNA's.

PACS number(s): 87.10.+e, 87.15.He, 87.15.Da

I. INTRODUCTION

The aim of this paper is to postulate and justify a plausible expedient by means of which a biopolymer chain can fold itself by selecting at each stage of the process the step that entails the minimal loss of conformational freedom. In this regard, the principle of sequential minimization of entropy loss (SMEL) is introduced. This is essentially a least-action approach in the spirit of Helmholtz's minimum principles in thermodynamics [1,2]. The SMEL principle relates to the context of biopolymer folding *in vitro* and, in plain terms asserts that folding proceeds at each stage by minimizing the loss of conformational freedom while forming as many new favorable intrachain contacts as possible. This tenet does not necessarily conform to existing algorithms for structure prediction rooted in free energy minimization. To cast the principle in proper terms, one should emphasize that the enthalpy loss associated with contact formation and the entropic contribution are not placed on equal footing: SMEL control implies that each event along the folding pathway is chosen so that the conformational entropy loss, denoted ΔS_{loop}, associated with loop closure is minimized and the number n of effective contacts maximized so that the quantity $Q=n^{-1}\exp(-\Delta S_{\text{loop}}/R)$ is minimized at each stage in the folding process. This requirement is justifiable: Equilibrium thermodynamics cannot dictate a sequence of events in a time-constrained situation unless the contributions to the thermodynamic potential represent kinetic parameters themselves [3].

The SMEL principle has been anticipated in the context of RNA secondary structure formation which often takes place under stringent time constraints [3–5]. This fact can be readily shown as follows. The unimolecular rate constant k of formation of an intramolecular stem [5,6] is

$$k=fn\ \exp(-B/RT)=fn\ \exp(\Delta S_{\text{loop}}/R), \tag{1}$$

where $f\approx 10^5\ \text{s}^{-1}$ is the rate constant for base formation within a disrupted helix [6,7], n is the number of contacts or base pairs stacked in the helical stem, T is the absolute temperature, and B is the activation energy barrier for closure of the loop associated with stem formation. Since the closure of the loop is the rate-determining step in the formation of an intramolecular stem [4–6], we have $B\approx\Delta G_{\text{loop}}\approx -T\Delta S_{\text{loop}}$, and thus, Eq. (1) follows. Thus the choice of an intramolecular folding event becomes dictated by the minimization of the quantity Q whenever the time span of each consecutive elementary unimolecular step is minimized. In other words, the SMEL principle holds whenever RNA folding is subject to kinetic control.

A direct and useful application of the SMEL principle for RNA folding prediction has been lacking since it is contingent upon a proper derivation of the entropic contribution for *any* size loop. The unreliability of the compilation of thermodynamic parameters [8] for conformational entropy loss, especially outside the excluded-volume regime [9], makes the SMEL principle particularly unsuitable given the exponential dependence of Q on the entropic term. Accordingly, we shall obtain the entropic contribution that corrects or supersedes, depending on the loop size, the excluded-volume effects. We shall show (Sec. II) that *a paramount effect is introduced by the orientational constraints associated with the self-energy of the charged phosphate moieties within an unstrained loop.* To conclude, the operational version of the SMEL principle will be applied to predict catalytically competent folding of specific ribozymes.

II. ENTROPY AND THE SELF-ENERGY OF POLAR MOIETIES

Loop closure is paramount among folding events that a flexible polymer chain undergoes. From a thermodynamic perspective, the entropic contribution $-T\Delta S_{\text{loop}}$ is a positive term in the free energy change for loop closure. Thus the loop closure becomes feasible only if sufficient contacts are made so that the enthalpic loss is smaller than minus the entropic contribution: $\Delta H<T\Delta S_{\text{loop}}$, where the enthalpy

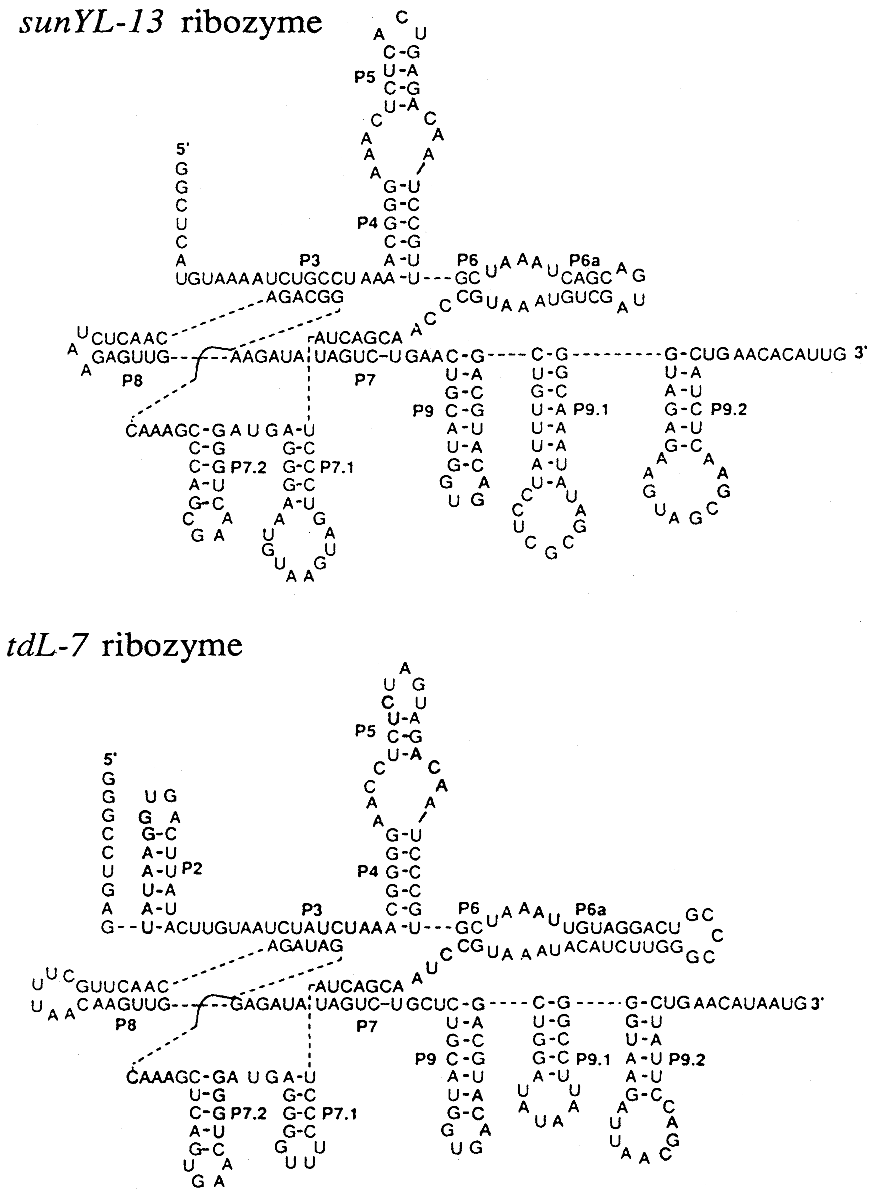
sunYL-13 ribozyme
5'
P5
P4
P3
P6
P6a
P8
P7
P9
P9.1
P9.2
P7.2
P7.1
UGUAAAAUCUGCCU AAA -U---GC
GAGUUG---AAGAUA UAGUC-U GAAC-G----C-G---------G-CUG AACACAUUG 3'
tdL-7 ribozyme
5'
P5
P4
P2
P3
P6
P6a
P8
P7
P9
P9.1
P9.2
P7.2
P7.1
G--U-ACUUGUAAUCUAUCUAAA -U---GC
CAAGUUG---GAGAUA UAGUC-U GCUC-G----C-G-----G-CUG AACAUAAUG 3'

FIG. 1. Secondary structure for the *sun YL-13* and the *tdL-7* group I ribozyme as obtained from the SMEL-based sequential algorithm.

change ΔH accounts for the heat released upon formation of the contacts.

Despite this simple scenario, the thermodynamics of a folding event involving loop closure could only be calculated rather crudely and for moderately large loops, where the effects of discrete solvent organization inside the loop may be neglected [8,9]. Thus, in the Jacobson-Stockmayer approximation, one assumes an unrestricted Gaussian coil of length N= size of the loop. This gives the well-known result [9]:

$$\Delta S_{\text{loop}} = -(3/2)R \ \ln N + R \ \ln[(3/2\pi l^2)^{3/2} v], \qquad (2)$$

where l is the effective length of a monomer and v is the effective contact volume within which two monomers are assumed to have made contact [9]. In a good solvent, where excluded-volume effects are to be taken into account, the logarithmic dependence on N must be corrected to $-\mu R \ \ln N$, with $\mu \approx 1.75$ [9].

At this point one can pose the question: Why can we not extrapolate these results for small loops? The answer is obviously no, since discrete solvent structure effects will become apparent, changing the situation in a qualitative way. In this section we offer a quantitative approach to encompass such effects. The discussion is first cast in general terms and subsequently specialized for RNA.

Let us assume that the solvent is water and that the polymer chain is able to selectively orient its charge or polar moieties whenever the solvent surrounding the molecule represents more than one distinctive dielectric environment. These requirements are fulfilled by RNA or proteins under *in vitro* renaturation conditions [9]. Provided the number of solvent molecules confined by the polymer rod inside the loop is sufficiently small, two distinctive domains are formed: the inner domain, of clusterlike dimensions, and the outer bulklike domain. The outer domain is a distinctively better dielectric and thus charges or polar moieties would, if at all feasible, tend to orient themselves towards the outer domain, where they can be better solvated.

There is a critical size loop beyond which both domains are indistinguishable. Thus, above the critical size loop, the confined water must be indistinguishable from the bulk. Given specific dimensions of the polymer rod and admitting a maximum of four solvation layers per polar moiety, it becomes straightforward to calculate the critical size $N=N_0$ beyond which the domain differences break down: it suffices to take a planar loop such that the inner domain of confined solvent molecules is at least seven molecules across for any pair of opposite points on the chain. Thus, in the particular case of a circular loop, its critical size would be the one for which the inner domain is seven solvent molecules in diameter. If the solvent is water, this is the cross section of a 28-molecule cluster which should be considered bulk for most thermodynamic implications [10].

The separation of environments produced as a loop of size $N<N_0$ is formed must have conspicuous entropic effects for polymer chains that are able to selectively orient their polar groups. Thus, if we assume two distinctive orientations for each residue, we obtain the following reduction of conformational entropy:

$$\Delta S_{\text{loop}}=R\ \ln(\Omega'/\Omega)=R\ \ln(2^{M-N_p}/2^M)=-RN_p\ln 2, \tag{3}$$

where Ω' is the number of conformations with the ends of the loop constrained to be in contact with one another, Ω is the total number of available conformations, M is the total length of the chain, and N_p is the number of residues in the loop that contain a polar or charged group which is able to orient itself facing the most favoring solvent domain. While nonpolar residues may have two distinctive orientations (pointing inwards or outwards), polar residues will invariably point outwards, and this fact drastically reduces the number of available conformations once a loop is formed. Since in general the number of polar side chains is proportional to the total number of residues, the orientational effect as estimated with Eq. (3) must be far more dominant (linear in N versus logarithmic in N) for small loops than the well-studied excluded-volume effect. The linear dependence of orientational effects on the size of the loop appears to be a consideration of paramount importance in theories that attempt to infer the folding events taking place in the initial stages of the search for the native conformation.

That proteins might exhibit this drastic entropy reduction upon loop closure is not unexpected because it is well established since Kauzmann's seminal observations [9] that polar groups always point to the bulk solvent in the native conformation. The situation is similar in RNA, since the charges are in this case located on the phosphate moieties of the backbone itself and thus are susceptible in unstrained loops ($N\geqslant 4$) of orienting themselves towards the bulk, minimizing the self-energy. Two pieces of evidence support this picture: (a) For small unstrained loops $N(=5-12)$, the indirectly measured [8] free energy associated with loop formation is proportional to N, and not to $\ln N$, as would be the case for a chain with excluded volume [9]. (b) There is strong kinetic evidence suggesting a substantial orientational contribution to ΔS_{loop}. Taking into account orientational effects, a straightforward computation of the unimolecular rate constant for the formation of a hairpin with loop size $N<N_0$ gives (cf [3–5].)

$$\begin{aligned} k&=fn\ \exp[-B/RT]=fn\ \exp[T\Delta S_{\text{loop}}/RT]\\ &=10^5\ \text{s}^{-1}n2^{-N}. \end{aligned} \tag{4}$$

In this computation, it has been rightly assumed that the rate-determining step in the formation of the hairpin is loop closure, which should be regarded as the nucleation event in intrachain helix formation (cf. Sec. I). Thus, for a loop of size 5–12, Eq. (4) brings the mean time of formation of the hairpin precisely in to the millisecond range, in good agreement with kinetic experiments as well as computations [3–5].

The estimation of the critical size N_0 for RNA follows readily from a simple computation incorporating molecular dimensions and regarding the molecule as a rod of sectional radius equal to the mean phosphate-base distance [9]. Thus, assuming a phosphate-base distance of 10 Å, a phosphate-phosphate distance of 5.9Å, and a diameter corresponding to seven water molecules across the critical loop we obtain a critical size $N_0\approx 17$. This computation leaves us with the following working equations:

$$\Delta S_{\text{loop}}(N)=-\mu R\ \ln[N/L]+\Delta S_{\text{loop}}(L)-RN\ \ln 2, \tag{5a}$$

valid for $5\leqslant N<N_0$ with $L\gg N$, $L\gg N_0$ and $\Delta S_{\text{loop}}(L)$, the entropy loss for closure of a loop of size L, known with good degree of confidence [9]. Above the critical size we get

$$\Delta S_{\text{loop}}(N)=-\mu R\ \ln[N/L]+\Delta S_{\text{loop}}(L) \tag{5b}$$

valid for $N>N_0$ and $L\gg N$.

Reliable measurements for triloops and tetraloops [8,9], together with Eqs. (5a) and (5b), respectively taking into account orientational effects as well as excluded-volume effects for sizes below criticality, or incorporating only excluded volume for sizes above criticality, allow us to implement the SMEL principle within the full range of folding events in RNA.

III. RESULTS AND DISCUSSION

A SMEL-based predictive algorithm requires the minimization of the quantity Q at each stage of the folding process using the working equations (5a) and (5b). This requires solving the combinatorial problem of searching for all Watson-Crick (WC) complementary regions with antiparallel orientation and following the pathway that allows for the maximization in the number of WC contacts with a minimal loss of conformational freedom for each step. Given these considerations, the dominant folding pathway is the one initiated by the folding event that minimizes Q when starting from a random coil. Thus the initial stages of folding favor short-range ($N\ll N_0$) and moderately long-range interactions ($N>N_0$). Since the latter are dominated by the slowly grow-

TABLE I. Sequence of events dictated by the SMEL principle for the most economic folding of the *sunYL-13* and *tdL-7* ribozymes.

sun YL-13 ribozyme:
$[P4][P6a,\ P7.2,\ P9] \rightarrow [P3] \rightarrow [P7][P7.1,\ P8,\ P9.1,\ P9.2]$
$[P5]$
tdL-7 ribozyme:
$[P4][P2,\ P7.1,\ P7.2,\ P9][P6a] \rightarrow [P3] \rightarrow [P7]$
$[P8,\ P9.1,\ P9.2][P5]$

ing term proportional to $\ln N$, loops involving solely excluded-volume effects have a high probability of occurring even if N is larger than N_0 by one order of magnitude.

By contrast, medium range interactions in the proximity of the critical size ($N \approx N_0$, $N < N_0$) and very long range interactions ($N \gg N_0$) become excruciatingly difficult at any stage, as direct inspection of Eqs. (5a), and (5b) reveals. However, the contacts that result form such interactions may form at later stages, induced by short- and moderately long-range interactions that may form first, shortening the loop that needs to be formed. The following two examples describing the SMEL folding of two catalytically competent RNA molecules or ribozymes illustrate this point. Standard notation has been adopted: the four possible RNA residues (nucleotides) are denoted G,C,A,U, where G=guanosine, C=cytosine, A=adenine, and U=uracil. The WC pairing is based upon the complementarity G-C, A-U. Thus, finding all plausible *a priori* foldings of an RNA chain is tantamount to solving the combinatorial problem of finding all WC complementary antiparallel regions of the chain.

Figure 1 displays the secondary structures for the *sun YL-13* and *tdL-7* ribozyme [11], as obtained making use of the SMEL-based algorithm. Certain WC complementary regions denoted by P of such molecules are conserved within a generic family of RNA catalysts, the group I ribozymes. Both structures predicted by means of the SMEL algorithm contain all conserved paired regions which are known to be required for RNA catalysis within group I introns.

The sequence of events determined by SMEL control and followed when a random coil is adopted as the starting point is displayed in Table I. Interactions which occur within the same stage of SMEL folding are grouped in square brackets. The arrows indicate that an interaction occurring at a certain stage has induced an interaction which takes place at the next stage. Thus the arrow indicates that an interaction of initially unfavorable range becomes more feasible once a loop within the purported loop of the former interaction is closed first.

We shall discuss the SMEL folding for *sun YL-13* and simply present the results for the *tdL-7* as the discussion is entirely analogous: In the first stage $P4$ forms ($Q \approx 40$) since this event requires closure of an $N = 18$-loop, for which the inner and outer domains are indistinguishable and thus the associated entropy loss does not contain the costly orientational contribution. In the next stage $P6a$, $P7.2$ and $P9$ form since they involve closure of the smallest unstrained loops for which inner and outer solvent domains are differentiated (tetraloops). The occurrence of $P4$, $P6a$, and $P7.2$ brings $P3$ well into the favorably moderately long-range interactions with $Q \approx 8.2$. Starting from a random coil, $P3$ would have a long-range interaction with $Q \approx 18.8$ and thus, highly improbable. In turn, the formation of $P3$ induces the formation of $P7$ ($Q \approx 8.8$), which now involves closure of two internal loops belonging to the moderately-long range domain with $Q = 4.4$ and $Q = 4.2$, respectively. Obviously, if an n-interaction entails closure of two loops, A and B, the quantity Q becomes $Q = n^{-1}\exp\{-[\Delta S_{\text{loop}}(A) + \Delta S_{\text{loop}}(B)]/R\}$. The next stage of SMEL folding entails closure of moderately short-range loops $P7.1$, $P8$, $P9.1$, and $P9.2$, where inner and outer solvent domains are clearly different and Q lies in the range $Q \approx 18.8$–19.2. In the final stage, the most unfavorable $P5$ forms. This interaction involves closure of a strained triloop ($Q \approx 18.0$) and also of a loop which requires the orientation of ten phosphate groups towards the outer solvent domain. Applying Eq. (5a), we obtain $Q \approx 19.8$ for this event.

These examples illustrate a physically intuitive principle at work. The principle demands from each folding event the maximum effectiveness, that is, the maximization of the number of contacts at the expense of a minimal loss of conformational freedom.

ACKNOWLEDGMENTS

The authors have received financial support from the National Research Council of Argentina (CONICET). A.F. acknowledges support from the U.S. Air Force Office of Scientific Research and from the J. S. Guggenheim Memorial Foundation.

[1] A. Helmholtz, J. Crelle **100**, 137 (1886).

[2] B. Levich, *Theoretical Physics, Vol. 4: Quantum Statistics and Chemical Kinetics* (Wiley, New York, 1973).

[3] A. Fernández, Phys. Rev. Lett. **64**, 2328 (1990).

[4] A. Fernández, J. Phys. A.: Math. Gen. **27**, 6039 (1994).

[5] A. Fernández, Phys. Rev. E **48**, 3107 (1993).

[6] V. V. Anshelevich, V. A. Vologodskii, A. V. Lukashin, and M. D. Frank-Kamenetskii, Biopolymers **23**, 39 (1994).

[7] D. Pörschke, Ph.D. thesis, University of Braunschweig, Germany, 1968 (unpublished).

[8] M. Zuker and P. Stiegler, Nucleic Acids Res. **9**, 133 (1981); D. H. Turner, N. Sugimoto, and S. M. Freier, Annu. Rev. Biophys. Biophys. Chem. **17**, 167 (1988).

[9] C. Cantor and P. R. Schimmel, *Biophysical Chemistry* (Freeman, San Francisco, 1980), Vol. III.

[10] O. Sinanoglu, Chem. Phys. Lett. **81**, 188 (1981).

[11] M. Belfort, J. Tomaschewski, J. Pedersen-Lane, D. A. Schub, J. M. Gott, M. Q. Xu, B. F. Lang, and F. Michel, Proc. Natl. Acad. Sci. USA **85**, 1151 (1988).