RAPID COMMUNICATIONS

PHYSICAL REVIEW A VOLUME 46, NUMBER 8 15 OCTOBER 1992

Computation of the fraction of RNA sequences that fold sequentially into a unique free-energy minimum

Ariel Fernández

Department of Chemistry, University of Miami, Coral Gables, Florida 33124,
*Department of Biochemistry and Molecular Biology, The Medical School, P.O. Box 016129, Miami, Florida 33101-6129,**
and Max-Planck-Institut für Biophysikalische Chemie, Am Fassberg 11, W-3400 Göttingen, Germany

(Received 6 January 1992)

We compute the fraction of RNA sequences of fixed length $N=220$ that fold expeditiously and reproducibly into a unique structure within realistic RNA replication time spans. The search in conformation space is assumed to be concurrent with the assembling of the molecule which occurs by sequential incorporation of nucleotides. Monte Carlo simulations of the folding events concomitant with polymerization events are used to determine the evolutionary consequences of the principle of expeditious and reproducible folding hereby introduced.

PACS number(s): 87.15.Da, 36.20.Ey

There is, in principle, no reason why the search for RNA conformations should start only once the molecule has been fully synthesized. There are a number of instances in RNA synthesis [1-3] where sequential folding, partially retaining intramolecular upstream structures, yields a biologically relevant structure at the end of a replication turnover. The intramolecular folding intermediates formed during the sequential assembling of RNA molecules seem to play a significant regulatory role in fundamental processes such as replication and transcription and appear to be responsible for variable rates of chain elongation [1-3]. The statistical weight of a particular folding intermediate depends on the feasibility of folding pathways determined normally during the early stages of folding. For instance, in template-instructed replication or transcription, they are determined by initiation signals. These events, which bias decisively the final outcome of sequential folding, may be compared to the fast compactization in protein folding which affects the statistical weight of secondary structures prior to the sequence of transitions of distorted nativelike conformations [4].

In this work we shall introduce a set of conditions that warrant, within realistic replication time spans, the expeditious and computationally reproducible sequential folding into a unique structure. The evolutionary constraints determined by this set of conditions will be discussed thereafter. In this way, this work attempts to establish the physical basis for effective folding leading to the free-energy minimum. Thus, the criterion put forth in this work is crucial because of its evolutionary implications. Specifically, it allows us to identify the set of sequences that might be targeted in a selection process aimed in principle at reproducing a specific structural context.

We start the analysis by coarse-graining conformation space to the level of secondary structure and describing the activation energy barriers that separate different conformations for a given RNA sequence. Accordingly, a range of transition time scales should be defined. Thus, we introduce a lower cutoff of the order of $1\ \mu s \approx T_{\text{nonerg}}$, a typical nonergodic time scale [3] corresponding physically to the unimolecular time scale of base pair formation [5]. The upper limit in the time scale window depends on the chain lengths considered. For computationally accessible lengths $N \approx 200$–250, the upper limit could be taken to be the overall replication time span, which is of the order of a few tens of seconds, depending on conditions [6].

Within this time scale window, our preliminary aim is to predict whether a specific folding intermediate, generically denoted S, will form expeditiously and reproducibly in sequential RNA folding and will become significantly populated. The problem could be laboriously solved in principle by determining the activation energy landscape [3] at each intermediate length $N' \leq N$. Instead, we have devised an alternative criterion rooted in Monte Carlo simulations which accounts for the fact that *the time constraints characteristic of RNA synthesis force the concomitant exploration of conformation space to be kinetically driven* [2,3]. The criterion rests on the general observation that in realistic situations, initial refolding events determine decisively the folding pathway. More explicitly, the structure-induced initiation signals for template-directed RNA synthesis correspond to a highly structured initial portion of the growing chain [1,2]. Consequently, we shall impose a restriction on the RNA sequences considered solely to address this situation.

Some elements need to be defined before the restriction can be formulated: We shall denote by $N_0 \leq N$ the minimal length of a chain able to sustain a *substate,* that is, a conformation with minimal lifetime flanked by kinetic barriers that scale as $N_0^{1/4}$ [3]. These substates are relatively flat minima in conformation space and the activation energy barriers that must be surmounted to escape from such kinetically trapped foldings are the lowest ones. Thus, the lifetimes for such conformations may be written as $\tau = A \exp\{N_0^{1/4}\}$, where $A \approx 10^{-6}$ s is the lifetime of a single base pair [2,3,5]. We may adopt as minimal lifetime for a substate the time span of the weakest nontrivial stem: $\tau = 10^{-5}$ s [5]. Given the nonergodic time scale adopted, $N_0 \approx [\ln 10]^4 \approx 28$.

This initial fragment should be dominated by a single

conformation Ω in the following sense: Let $E(\Omega)$ be the energy of Ω and let $M=M(E(\Omega), E(\Omega)+kTN_0^{1/4})$ be the set of substrates [3] with energies in the range $(E(\Omega), E(\Omega)+kTN_0^{1/4})$ which are kinetically connected to Ω. Then, the dominance of Ω is reflected in the inequality

$$\int_{I(M)} P_{N_0}(Q)dQ \geq 1-\epsilon(N_0), \quad \epsilon(N_0) \ll 1 . \tag{1}$$

Here the argument Q denotes overlap between conformations, $P_{N_0}(Q)$ is a Parisi-type distribution [7] and the subset $I(M)$ of the interval [0,1] is defined as:

$$I(M)=\{Q_{\beta\Omega};\ \beta=\text{any conformation in } M\} . \tag{2}$$

The distribution is defined by (cf. [7])

$$P_{N_0}(Q)=\sum_{\beta,\beta'} \delta(Q_{\beta\beta'}-Q)p_\beta p_{\beta'} , \tag{3}$$

where $Q_{\beta\beta'}=$(number of base pairs common to conformations β and β')/(number of base pairs in Ω) and $p_\beta, p_{\beta'}$ are Boltzmann weights for conformations β and β', respectively.

In order to establish a criterion to determine which folding intermediates induced by Ω will be significantly populated, we have made use of Monte Carlo simulations of refolding events concurrent with polymerization events [2,3] for RNA sequences with $N=220$. Such simulations become feasible since the chain of kinetically controlled events constitutes a Markov process, as confirmed in pulse-chase experiments [1]. The specifics on these simulations have been described extensively in [2,3] and need not be reproduced here. The simulations mimick progressive refolding which takes place since the RNA chain explores different folding alternative as they become available concomitantly with the sequential incorporation of nucleotides.

Restricting the discussion to RNA sequences that satisfy relations (1)–(3), our main result may be formulated as follows: Let S be the conformation of the complete RNA chain with lowest energy $E(S)$ which results from a sequential folding pathway retaining Ω. Then, for a specific sequence, all the RNA molecules samples (each one corresponding to one of the n runs of the simulation) will fold into S if and only if

$$n(S|\Omega)/n \leq \delta^{-1}(\epsilon(N_0))\tilde{n}(S)/n , \tag{4a}$$

$$\tilde{n}(S)/n \leq n(S|\Omega)/n , \tag{4b}$$

where $\tilde{n}(S)/n=\exp[-\beta E(S)]/Z(N)$ is the Boltzmann probability of an RNA molecule to fold into S with $Z(N)$ being the partition function for the complete chain resolved up to secondary structure, and $n(S|\Omega)/n$ is the time-dependent probability that Ω induces the formation of S along the kinetically dominant folding pathway. The function $\delta(\epsilon(N_0))$ (Table I) has been obtained by computing the dominant folding pathway for 80 sequences which share the initially generated subsequence of length $N_0=28$ with the sequence given in Fig. 1. Thus, all the sequences generated obey relations (1)–(3), since the sequence indicated in Fig. 1 does so.

Relation (4a) may be interpreted in the sense that S

TABLE I. The function $\delta(\epsilon(N_0))$, which gives the minimal fraction of thermodynamic dominance required from structure S and depends on the degree of dominance of the structure which represents the initiation signal, as indicated by Eq. (1). Obviously, a refinement of the values could be achieved by performing the simulation on a sample in sequence space larger than the set of 80 RNA sequences considered.

$\epsilon(N_0)$ (%)	$\delta(\epsilon(N_0))$
10	0.301
20	0.440
30	0.688
40	1.000

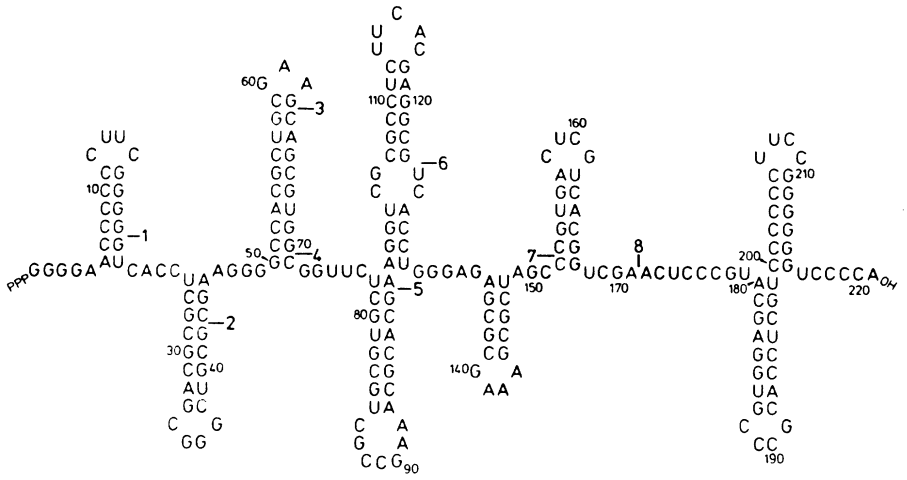

FIG. 1. Secondary structure of a natural RNA sequence emerging 20 s after initiation of sequential assembling of the molecule. The sequence optimizes the regulatory role of transient structures formed concurrently with the replication of the molecule in template-instructed synthesis. The nonadjacent base pairs brought to proximity are engaged in Watson-Crick interactions G-C and A-U (G=guanosine, C=cytosine, A=adenosine, U=uracil). The digits denote sites determined by the simulation where further polymerization is precluded until a refolding event has taken place.

will form expeditiously if it is not too fragile a structure, that is, if its Boltzmann probability is at least a fraction $\delta(\epsilon(N_0))$ of the probability of being formed by sequential folding. On the other hand, the latter probability should surpass the probability that S is formed solely under thermodynamic control, as indicated by condition (4b). This should be demanded if Ω is to determine the kinetically controlled pathway that leads to S. As we have emphasized, thermodynamic control should be superseded by kinetic control when the time constraints are far more stringent than the time scale allowed for random exploration of conformation space. Figure 2 displays runs averaged over $n=20$ for three different sequences subject to restrictions (1)-(3). The time span of the simulations is invariably 20 s, a typical time span for RNA replication [6]. Plot *a* corresponds to the sequence given in Fig. 1, a natural template which optimizes the regulatory function in template-directed replication [1-3]. Relations (4a) and (4b) are indeed satisfied for this sequence which folds after 20 s into the structure depicted in Fig. 1. Plots *b* and *c* correspond to sequences which violate conditions (4a) and (4b), respectively. Therefore, as expected, the expeditious folding into S does not take place appreciably within the time span indicated.

It should be noted that the significantly populated structure S given in Fig. 1 is indeed a metastable structure, a folding intermediate. The free-energy minimum has been identified [2] and it is obtained by bringing together the extensively complementary initiation and termination subsequences, thus requiring the dismantling of the seed structure Ω.

Of the 80 sequences considered, only four of them (5%) meet the restrictions given by relations (4a) and (4b). Those which did not meet the constraints invariably gave nonreproducible results for the 20 runs consisting each of them in 10^7 Monte Carlo steps. More importantly, when the constraints were not met, not a single run yielded the structure S associated with the particular sequence, not even when the number of steps was extended to 10^8. Thus, the evolutionary constraints implied by the set of conditions (1)-(4) are apparent.

A vast sample in sequence space would be required to verify whether all sequences which fold expeditiously and reproducibly actually share further homology with the sequence presented in Fig. 1, beyond the initial 28 nucleotides.

This work introduces a dynamical criterion to determine whether a specific RNA sequence will search for its conformation following a folding pathway which is reproducible and expeditious. The criterion is rooted in an algorithm which consists in the simulation of kinetically controlled sequential folding. Thus, our conclusions are only valid for sequences which search for their conformation as they are being progressively assembled by successive incorporation of nucleotides. Specifically, the criterion is applicable for RNA transcripts and template-directed replication products but it does not hold true for RNA species whose folding pathway is not sequential or subject to time constraints. In this sense, our folding algorithm is complementary to existing algorithms rooted in free-energy minimization.

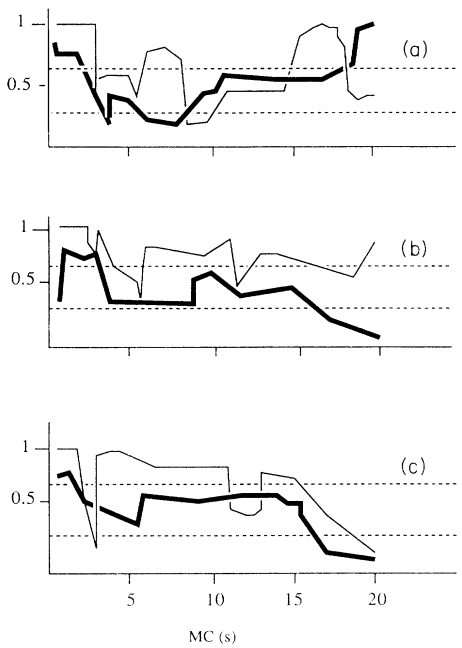

FIG. 2. Time dependence of the probability of the most probable secondary structure S' formed concomitantly with the progressive polymerization leading to chain growth. The thick line plots represent the fraction $n(S')/n$ ($S'=\Omega$ at length N_0, $S'=S$ at length N) while the thin line plots represent $n(S'|\Omega)/n$. The lower dashed line corresponds to the Boltzmann probability $\tilde{n}(S)/n$ and the higher dashed line to $\delta^{-1}(\epsilon(N_0))\tilde{n}(S)/n$. Plot *a* corresponds to the sequence indicated in Fig. 1, while plots *b* and *c* correspond, respectively, to sequences for which conditions (4a) and (4b) are violated. MC (s) stands for Monte Carlo steps equivalent to real time measured in seconds.

The specialization of the results to the specific example of the wild type midivariant-1 RNA, a natural template in RNA replication [1], should indeed instill confidence in the validity of the criterion. The active structure for this species is represented in Fig. 1 and has been confirmed by site-directed mutagenesis. Moreover, the folding pathway generated by our algorithm has also been confirmed by kinetic pulse-chase experiments aimed at trapping structure-induced replication intermediates [1,2]. In accord with these findings, our sequential folding algorithm reveals that the species midivariant-1 RNA, an obvious target of natural selection, satisfies our dynamic criterion for expeditious folding.

The author received support from the Camille and Henry Dreyfus Foundation.

*Present address.

[1] D. R. Mills, C. Dobkin, and F. R. Kramer, Cell **15**, 541 (1978).
[2] A. Fernández, Physica A **176**, 499 (1991).
[3] A. Fernández and E. I. Shakhnovich, Phys. Rev. A **42**, 3657 (1990).
[4] D. M. Brems and H. A. Havel, Proteins: Struct. Funct. Genet. **5**, 93 (1989).
[5] V. V. Anshelevich, V. A. Vologodskii, A. V. Lukashin, and M. D. Frank-Kamenetskii, Biopolymers **23**, 39 (1984).
[6] J. Fernández, J. Theor. Biol. **134**, 419 (1988).
[7] E. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).