Random Energy Model for the Kinetics of RNA Folding

Ariel Fernández

Department of Chemistry, University of Miami, Coral Gables, Florida 33124
and Department of Biochemistry and Molecular Biology, Medical School, P.O. Box 016129, Miami, Florida 33101
(Received 23 May 1990)

I model the relaxation dynamics of metastable RNA folding by means of a master equation for the distribution of folded-state occupancies. The underlying primary sequence is considered to be random and uncorrelated. The model is tested *vis-à-vis* a Monte Carlo simulation of kinetically governed refolding events. The validity of a random energy model is established.

PACS numbers: 64.70.Pf, 36.20.Ey, 64.60.Cn

The kinetics of RNA folding remains poorly understood in spite of its paramount biological importance.[1-3] Crucial processes such as transcription and RNA replication yield RNA products folded into secondary structures which are biologically active. As a consequence of the short time scales of the enzymatic events involved, the formation of the active structure is kinetically governed and seldom thermodynamically controlled.[3,4] Thus, the folding of the product which emerges once transcription or replication has reached completion has been shown in some instances to be the most probable among the fast-formed secondary structures.[4] Moreover, this structure appears to be preserved for further polymerization events. In such instances, *the degree of folding* (as given by the number of Watson-Crick base pairs) *is not maximal,* the initiation sequences being usually unfolded for further recognition by the enzyme. Thus, the structure featuring maximal folding is biologically inert.[2] The essential point is that an understanding of the relaxation dynamics for intermediate folded structures is crucial to assess the molecular basis of regulation and control.

In order to treat the relaxation kinetics for metastable RNA folding with the tools of statistical mechanics of disordered condensed matter, I shall assume an underlying random uncorrelated primary sequence. Thus, the primary sequence can be viewed as a quenched disorder. The biological significance of such an assumption is apparent for biopolymers in general.[5-7] Recent results by Ptitsyn and Volkenstein[5] suggest that naturally occurring biopolymers might be more "disordered" than one might expect; that is, the dominance of specific folded forms which are biologically active appears to hold within vast domains in sequence space.

In this work I shall prove that a random energy model (REM) is suitable to study the relaxation of metastable RNA secondary structures. Such models have been implemented originally within the context of spin glasses and other condensed-matter systems with quenched disorder.[8] The tenets of my model are the following:

(a) Each folded structure or configuration has associated an energy level. The total number of energy levels is assumed to be 2^N, where N is the length of the chain.[8]

(b) The average number $\langle n(E)\rangle$ of energy levels with energy E is

$$\langle n(E)\rangle = 2^N[2\pi\Delta E^2]^{-1/2}\exp\{-[E-\langle E\rangle]^2/2\Delta E^2\}, \quad (1)$$

where

$$\Delta E^2 = \langle E^2\rangle - \langle E\rangle^2 \quad (2)$$

and $\langle\ \rangle$ denotes an average over the quenched disorder of the thermal or statistical average; that is, an average over sequence space of the average over all configurations for a given sequence.

(c) The relaxation of a given metastable RNA secondary structure (configuration) is studied considering the behavior of $p(E,t)$, the time-dependent distribution of the energy-level occupancies. Thus, we have

$$p(E,t)dE = dM(E,t)/M, \quad (3)$$

where $dM(E,t)$ is the number of molecules in a configuration with energy between E and $E+dE$, and M is the total number of molecules.

(d) The distribution $p(E,t)$ obeys the general master equation

$$dp(E,t)/dt = -p(E,t)\int K(E',E)dE' + \int K(E,E')p(E',t)dE', \quad (4)$$

where

$$K(E,E') = \langle n(E)\rangle A^{-1}\exp\{-(E_{\text{nonerg}}-E')/RT\} \quad (5)$$

and E_{nonerg} is the energy of a "nonergodic" transition state separating two substates.[3] The term nonergodic indicates the fact that we are not dealing with large conformational changes, with barriers of order $N^{1/2}$, but, rather, with conformational changes associated with kinetic barriers which scale as $N^{1/4}$. Such barriers diverge far more slowly than the ergodic barriers when the thermodynamic limit is approached. Their existence has been previously determined.[3] The nonergodic state is actually a collection of virtual random-coil intermediates whose existence is postulated to account for the fact that

a refolding event entails a partial unfolding of previously existing structures. It should be noticed, however, that for a specific chain it is not generally true that unfolding and, consequently, dismantling of previous structures is required for refolding: The new structure could emerge progressively, by slippage of base pairs, for instance. However, such progressive refolding obviously requires a highly ordered sequence, at least exhibiting periodicity in the distribution of base pairs. Such contexts are not considered in this work, where I focus strictly on random sequences, where the methods of disordered condensed-matter physics are relevant. The validity of the general master equation (4) supports a scenario in which refolding is not progressive in the sense of small consecutive unfolding and folding steps. Rather, existing structures in the form of hairpins are dismantled in a first stage and this process is followed by structure formation. Thus, quasi-random-coil structures mediate the transitions. Their common energy value determines the time scale in Eq. (4). The preexponential factor A is fixed at 1.12 s.[3] The activation energies $(E_{\text{nonerg}}-E)$ are analogous to the spin-glass kinetic barriers; that is, they scale with $N^{1/4}$:

$$E_{\text{nonerg}}-E=\mu RTN^{1/4},\quad \mu=\mu(E), \tag{6}$$

with $E_{\text{nonerg}}-E$ the kinetic barrier or activation energy associated with the relaxation of the folding pattern with energy E and μ the scaling factor. Throughout this work, I shall not consider large relaxation times, corresponding to vast changes in secondary structure, but only refolding events which are accessible within the Monte Carlo simulation; that is, those whose time span lies in a neighborhood of the already established nonergodic time scale.[3] Thus, the relaxation time t_{relax} for a configuration with energy E is

$$t_{\text{relax}}\sim\exp\{(E_{\text{nonerg}}-E)/RT\}. \tag{7}$$

The variance of relaxation times is directly accessible given its dependence on the variance in the energy of substates, as indicated by the functional dependence in Eq. (7). However, the variance in energies is regarded as an adjustable parameter in the model I am considering; that is, it is estimated indirectly, contrasting the numerical integration with a Monte Carlo simulation of kinetically governed refolding events.

(e) The final equilibrium energy E_{eq} for temperatures above the critical temperature for the frozen phase transition is a function of the dispersion of energies over the quenched disorder:

$$E_{\text{eq}}=\langle E\rangle-\Delta E^2/RT. \tag{8}$$

The corresponding overall relaxation time is

$$\tau=A\exp\{[\Delta E^2/(RT)^2-1]+E_{\text{nonerg}}/RT\}. \tag{9}$$

The REM determined by the tenets (a)–(e) predicts an increase in the average activation energy which is logarithmic in time. Such behavior is displayed in Fig. 1. The solid line represents the time dependence of the activation energy. The plot was obtained by numerical integration of Eq. (4) with the following choice of parameters: $N=512$, $T_{\text{nonerg}}=A\exp(E_{\text{nonerg}}/RT)$, $E_{\text{nonerg}}-E_{\text{eq}}=1.84RTN^{1/4}$, $\Delta E^2\approx 1.88(RT)^2$, and $T=25\,^{\circ}\text{C}$. It turns out that *this choice of parameters allows me to reproduce almost exactly the results of the Monte Carlo simulation of the kinetically governed refolding events.* The latter are revealed by the dashed-line plot in Fig. 1. The characteristic quantity τ, readily accessible from the simulation, is identical (within the uncertainty in the parametrization of free-energy contributions used in the simulation) to the result obtained by numerical integration of Eq. (4): $\tau=67.1\pm 1.6$ s.

For the sake of completeness, I shall describe the Monte Carlo simulation used to obtain the time-dependent probabilities for the transient secondary structures.[2,3] The case of interest is one in which the primary sequence is randomly generated, the limit of relatively long chains is explored, and the length of the chain is fixed throughout the simulation. The simulation mimics a Markov process in which, as new possibilities for folding arise, previously existing metastable structures are dismantled to allow for the formation of the emerging ones. The Markov process is comprised of two different

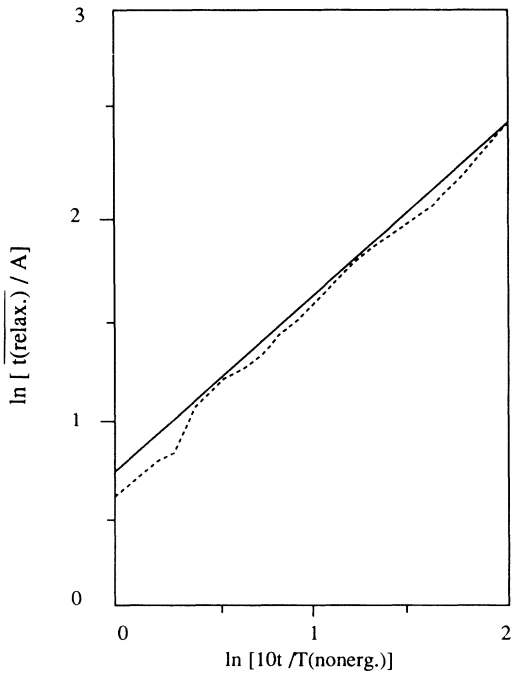

FIG. 1. Time dependence of the average relaxation time scale $\bar{t}_{\text{relax}}$. The choice of parameters is given in the text. The abscissas correspond to the range of real time (10%–80%) $\times T_{\text{nonerg}}$. The ordinates yield the activation barriers at each instant. The solid line is obtained by numerical integration of the master equation (4), representing the random energy model. The dashed line represents the result of the Monte Carlo simulation of the Markov chain of refolding events.

kinds of kinetically governed elementary events: (I) intrachain partial helix formation and (II) intrachain helix decay. In addition, I have incorporated certain features absent in previous work:[2] the possibility of G-T and A-C mispairs and the possibility of looped-out bases in the process of helix formation. The formation of new helices should always be topologically compatible with the pattern of existing ones in the sense that no knots may be allowed. This condition has been given proper combinatorial form and as such is incorporated in the algorithm in a standard manner. The present simulation differs from previous approaches[2,3] in that I have introduced a refinement in the kinetics of helix decay.

If intrachain helix decay is the chosen event, the inverse mean time can be obtained from an improved version of the expression for the kinetics, obtained by Anshelevich *et al.*[7] These authors give the equation

$$t^{-1}=fnS_{\mathrm{eq}}^{-n}, \tag{10}$$

where S_{eq} is the equilibrium constant for base-pair formation. However, their treatment does not properly distinguish between stacking and initiation of the base-pairing process.[9] Thus, I shall use instead the improved equation

$$t^{-1}=fn[KS^{n-1}]^{-1}, \tag{11}$$

where S is the geometrical mean of the base-stacking equilibrium constants (adequate for a random uncorrelated primary sequence) and K is the equilibrium constant for base-pairing initiation (nucleation equilibrium constant); $K(\mathrm{A\text{-}U})\approx 4\times 10^{-5}\ \mathrm{M}^{-1}$ and $K(\mathrm{G\text{-}C})\approx 2.5\times 10^{-4}\ \mathrm{M}^{-1}$.

The process made up of consecutive events of types (I) and (II) has been repeated 10^5 times for a random uncorrelated chain of length $N=512$. The actual time span for the simulation is 27 min of Cray-1S time.

The relaxation time t_{relax} can be obtained from the behavior of the time-dependent probability $U_n=U_n(t)$ for the most probable secondary structure $\underline{n}$ at time t. This function is readily accessible from our simulations. This is particularly crucial since the kinetic barriers for interconversion between metastable secondary structures also become accessible: The activation energy for the transition between two structures is given by

$$\{E_{\mathrm{nonerg}}-E(\underline{n})\}/RT=E_a(n\rightarrow n+1)\propto|U_n'(t^*)-U_{n+1}'(t^*)|^{-0.33}, \tag{12}$$

where the prime denotes time derivative and t^* is the actual instant when one structure is superseded by another occurring with a subsequently higher probability. This fit was found empirically.[2,3]

The relaxation of metastable secondary structures is completely characterized once the activation-energy landscape for transitions is described. Thus, the kinetic barrier encountered at any given instant should be calculated. The real-time parameter t must be properly scaled in order to determine the range of transitions accessible from the simulation. An adequate scaling factor, T_{nonerg}, the therefore introduced. This factor is conveniently defined as follows:

$$T_{\mathrm{nonerg}}=A\exp\{[E_{\mathrm{nonerg}}-E(t=0)]/RT\}. \tag{13}$$

Only transitions which occur within a certain vicinity of the nonergodic time scale $[(10\%\text{-}80\%)T_{\mathrm{nonerg}}]$ are accessible computationally. Such transitions correspond to relatively small changes in structure; that is, they entail substates exclusively. The regime corresponds to the range of abscissas given in Fig. 1, where the results of the simulation are revealed by the dashed-line plot. The time dependence of the activation energies for refolding events is thus displayed in Fig. 1. The activation-energy landscape in the range of time scales considered can be most adequately described in terms of the REM. The signature of this assumption is the logarithmic growth of the activation energy for transitions, as indicated by the solid line. Thus, the validity of the REM model has been established. The fact that no slippage in the base pairs or progressive refolding is contemplated in the master equation is entirely compatible with the all-or-none stepwise refolding described by the Markov chain in the simulation. The entire character of the problem would be distorted if short-ranged periodicities in the primary sequence occurred, since that could make slippage possible. Nevertheless, this situation is precluded given the uncorrelated and random nature of the chain.

A far more arduous task would be to characterize the relaxation in regimes which entail strongly divergent kinetic barriers in the thermodynamic limit (the "ergodic" regime). Carefully biased means of exploring conformation space might be required.

This work has been financially supported by a distinguished New Faculty Award from the Camille and Henry Dreyfus Foundation.

[1]S. E. LaFlamme, F. R. Kramer, and D. R. Mills, Nucleic Acids Res. **13**, 8425 (1985).

[2]A. Fernández, Eur. J. Biochem. **182**, 161 (1989).

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

[4]A. Fernández, Arch. Biochem. Biophys. **280**, 421 (1990).

[5]O. B. Ptitsyn and M. V. Volkenstein, J. Biomol. Struct. Dyn. **4**, 137 (1986).

[6]J. B. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).

[7]V. V. Anshelevich, A. V. Vologodskii, A. V. Lukashin, and M. D. Frank-Kamenetskii, Biopolymers **23**, 39 (1984).

[8]B. Derrida, Phys. Rev. B **24**, 2613 (1981).

[9]S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner, Proc. Natl. Acad. Sci. U.S.A. **83**, 9373 (1986).