PHYSICAL REVIEW E VOLUME 50, NUMBER 4 OCTOBER 1994

# Describing RNA sequential folding by dynamic coarse graining of the extended conformation space

Ariel Fernández
*Instituto de Matemática, Universidad Nacional del Sur–Consejo Nacional de Investigaciones Científicas y Técnicas, Avenida Alem 1253, Bahía Blanca 8000, Argentina*
*and Instituto de Investigaciones Bioquímicas, Universidad Nacional del Sur–Consejo Nacional de Investigaciones Científicas y Técnicas, Bahía Blanca 8000, Argentina*
(Received 31 May 1994)

We coarse-grain an extended conformation space for RNA chains that search for their structure concurrently with their sequential synthesis. Thus we describe the sequential folding dynamics *modulo* kinetic barriers of order $N^{\alpha}$ ($N$ is the number of monomers in the segment and $\frac{1}{4}\leqslant\alpha\leqslant\frac{1}{2}$). The random energy model is shown to hold for a range of $\alpha$'s which depends on the RNA primary sequence.

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

There is in principle no reason why a biopolymer should not start the search for its active structure while it is being assembled by progressive incorporation of monomers. Thus the sequential folding (SF) scenario, that is, the search in conformation space by a molecular that is concurrently being assembled, has been validated for real RNA's whose structure formation is subject to time constraints. In specific contexts where the synthetic machinery does not significantly interfere with biasing the upstream folding of the growing RNA chain, the SF scenario appears to account for the variable rate of chain elongation during progressive RNA synthesis [1–3].

The dearth of biological systems that have been addressed theoretically results from the fact that a proteic environment can radically alter the refolding of the growing chain and such complex protein-RNA interactions cannot be successfully modeled at present. Nevertheless, in the relatively simple context of RNA replication, the SF scenario has been successfully shown to account for pulse-chase kinetic experiments revealing the variable rate of chain elongation [1–3].

We intend to characterize the complex dynamics of sequential folding for a specific RNA molecule $N_0$ monomers long by introducing various coarse-grainings of the extended conformation space $\mathfrak{I}$. This space contains all plausible secondary structures (intrachain base-pair patterns subject to the Watson-Crick complementarity rules G—C, A—U, where G=guanosine, C=cytosine, A=adenine, and U=uracil) formed by segments of every length $N$, with $1\leqslant N\leqslant N_0$. We shall show that when elements in $\mathfrak{I}$ are regarded as *modulo* low kinetic barriers of interconversion, the resulting dynamics of transitions between clusters of conformations follow a type of relaxation characteristic of a random energy model (REM) [4]. This description becomes meaningful since a sufficiently large number—of order $2^N$—of energy levels has been considered. To understand the REM model as applied to RNA sequential folding we first group structures that interconvert on fast time scales of the order of $A^{-1}\exp(N^{1/4+\varepsilon})$, with $A=10^3\ \mathrm{s}^{-1}$ and $\varepsilon\geqslant 0$. In this context, the expected activation energy barriers for monitored transitions between equilibrated clusters depend on the progress of the relaxation process in such a way that at time $t$ the barrier to be surmounted is of logarithmic order in $t$ [5]. That is, we shall show that for a suitable level of coarse-graining the SF pathways are governed by a REM relaxation which entails surmounting larger and larger barriers which at time $t$ are of order $\ln t/q$, where $q$ is a characteristic time scale dependent on the size of the system and on the relaxation time range considered.

It should be pointed out that, since the number of plausible secondary structures for a chain of length $N$ is exponentially dependent on $N$, as simple combinatorial arguments show [5,6], we are not actually dealing with a thermodynamic limit but rather with a sufficiently large number of energy levels to apply REM statistics and, in general, to make probabilistic inferences. Within this context, an appropriate convention has been adopted [5]: Slowly diverging barriers of $O(N^{1/4+\varepsilon})$ are referred to as nonergodic, while fast diverging barriers of order $N^{1/2-\varepsilon}$ are referred to as ergodic, solely to indicate that the latter are not surmountable within the experimental time scales involved in the assembling of RNA molecules [3,5].

The REM description will be shown to break down as the size of the kinetic clusters is increased. As we approach ergodic cluster sizes, that is, as we identify conformations separated by barriers of order $N^{1/2-\varepsilon}$, a distinctively organized region of the energy spectrum is explored. Thus the activation energy barriers of significant transitions grow far more slowly than any multiple of $\ln[t/\Omega(N(t))]$, where $\Omega(N)=A^{-1}\exp(N^{1/2-\varepsilon})$ is the characteristic time scale for a chain that has reached length $N$ at time $t$.

Thus we shall conclude that only the upper portion of the extended energy spectrum for a real RNA chain that folds sequentially is random and uncorrelated (cf. [6]). SF delivers the molecule to organized states only after REM-like equilibration has taken place within clusters of rapidly interconverting states. Such observations are compatible with the tenets of statistical mechanics in the sense that organized behavior manifests itself in an averaged representation of the dynamics.

We shall represent each coarse-graining by a quotient space consisting of equivalence classes each of which is formed by conformations that have been grouped and thus are regarded as equivalent. A convenient conformation space $\mathfrak{I}$ contains all folded segments of various lengths regarded

1063-651X/94/50(4)/2435(4)/$06.00  

*modulo* their secondary structure. Thus, each equivalence class is labeled by a base-pairing pattern.

In order to represent the dynamics of sequential folding we shall now define a quotient space $\mathfrak{I}/\equiv_\alpha$ in which we regard secondary structures *modulo* the kinetic barriers associated to their interconversion. That is, the equivalence relation "$\equiv_\alpha$" is defined as follows: Let $s,s' \in \mathfrak{I}$, then $s \equiv_\alpha s'$ if and only if

$$\begin{aligned} -\ln[k(s\to s')/A] &= O(N_{\min}(s,s')^\alpha), \\ -\ln[k(s'\to s)/A] &= O(N_{\min}(s,s')^\alpha), \end{aligned} \tag{1}$$

where $\frac{1}{4}\leq\alpha\leq\frac{1}{2}$, and $k(s\to s')$ is the unimolecular rate constant [5,7] for the rate-limiting step in the interconversion between the member of minimal length in class $s$ and the member of minimal length in class $s'$. The integer $N_{\min}(s,s')$ is the minimum chain length in the reunion $s\cup s'$. Each equivalence relation $\equiv_\alpha$ defined on $\mathfrak{I}$ corresponds to a specific truncation of the activation energy landscape such that secondary structures are regarded *modulo* kinetic barriers of interconversion of order $N^\alpha$.

At this point we may describe the dynamics for different coarse-grainings of the activation energy landscape. In order to properly do this, we shall regard sequential folding pathways in $\mathfrak{I}$ as integral curves of a vector field defined over $\mathfrak{I}$ and describe different skeletal versions of this field corresponding to different coarse-grainings of $\mathfrak{I}$. The advantage of this geometric approach lies in the fact that all trajectories may be studied in a systematic fashion as will become apparent when the computational results are presented. In rigorous terms, if the map $\Lambda$: $\mathfrak{I}\to T\mathfrak{I}$ ($T$ is a tangent bundle, that is, the space of all plausible smooth vector fields defined over $\mathfrak{I}$) denotes the vector field whose trajectories are the SF pathways, we are interested in describing the vector field $\Lambda_\alpha$: $\mathfrak{I}/\equiv_\alpha \to T[\mathfrak{I}/\equiv_\alpha]$, a map that makes the following diagram commutative (by commutative we mean that, given a starting point in $\mathfrak{I}$, the same image in $T[\mathfrak{I}/\equiv_\alpha]$ is obtained regardless of the pathway we choose along the diagram):

$$\begin{array}{ccc} \mathfrak{I} & \overset{\Phi}{\to} & \mathfrak{I}/\equiv_\alpha \\ \Lambda\ \downarrow & & \downarrow\ \Lambda_\alpha \\ T\mathfrak{I} & \underset{T\Phi}{\to} & T[\mathfrak{I}/\equiv_\alpha] \end{array} \tag{2}$$

where $\Phi$ and $T\Phi$ denote the canonical projections which associate each element to its equivalence class. The commutativity of the diagram translates into the operator equation

$$\Lambda_\alpha\Phi = [T\Phi]\Lambda. \tag{3}$$

Thus for chain length $N$, the map $\Lambda_\alpha$ determines the possible events whose associated time scales are larger than $A^{-1}\exp(N^\alpha)$.

The $\Lambda$ dynamics has been simulated using kinetically controlled Monte Carlo methods [3,5]. Thus, a sequence of refolding and chain growth events becomes a realization of a Markov chain representing a trajectory in $\mathfrak{I}$. Such computations have been described elsewhere [3,5], thus only the basic tenets are sketched.

For each value of the contour variable $N$ we define a map $N\to J(N)=\{j:1\leq j\leq n(N)\}$, where $J(N)$ is a collection of elementary events which a segment of length $N$ might undergo, and $n(N)$ is the number of elementary events. Associated to each event, there is a unimolecular rate constant $k_j(N)$ which is equal to the rate constant for the $j$th event which may take place as the chain reaches length $N$. The only elementary events allowed are chain-elongation steps ($j=1$), or elementary refolding events ($j\geq 2$) that should satisfy $k_j(N)^{-1}\leq t_{\text{expt}}$; $t_{\text{expt}}$ is the experimental replication turnover time scale ($\approx 15$ s for an RNA sequence 220 nucleotides long) [1,3]. The mean time for an elementary refolding event is the reciprocal of its unimolecular rate constant. Since $\mathfrak{I}$ is made up of secondary structures for strands of various lengths, the mean time for an elementary refolding event is the sum of the mean time of a single helix-decay (or dismantling) event, which is zero in the particular case where no helix needs to be dismantled, plus the mean time of a helix-formation event.

The unimolecular rate constants for helix decay and helix formation have been obtained in analytical form [5,7] and used extensively in our computations. Their associated kinetic barriers depend respectively on the enthalpic loss associated to helix formation and the entropy loss associated to loop closure. Thus the compilation of thermodynamic parameters [8] begets the compilation of unimolecular rate constants upon which the Markov chain is constructed. The Markovian nature of the process is in accord with experimental evidence [1] and is defined as follows.

Let $r\in[0,\Sigma_{j=1}^{n(N)}k_j(N)]$ be a Poissonian random variable and let $r^*$ be a realization of $r$ such that if

$$\sum_{j=0}^{j^*-1} k_j(N) \leq r^* \leq \sum_{j=0}^{j^*} k_j(N) \tag{4}$$

[$k_0(N)=0$ for any $N$), then the event $j^*=j^*(N)$ is chosen as the growing RNA chain reaches length $N$. The sequence $\{j^*(1),j^*(2),j^*(3),\ldots,\}$ constitutes a realization of the Markov process.

A regular site $N=N(\text{reg})$ along the RNA chain corresponds to a segment for which chain elongation is the prevailing event, that is $j^*(N(\text{reg}))=1$. On the other hand, at a pause site $N=N(\text{pause})$, there exists at least one unimolecular rate constant for refolding which is comparable to $k_1(N(\text{reg}))=k_1(N(\text{pause})=50\ \text{s}^{-1}$ [5]. Thus, the $\Lambda$ dynamics are characterized by a relaxation process and the expected relaxation time, $\langle t(\text{relax})\rangle$, for each transition is computed as

$$\langle t(\text{relax})\rangle = [k_{j^*}(N(\text{pause}))]^{-1}. \tag{5}$$

A Markov chain $\{j^*(1),j^*(2),j^*(3),\ldots,\}$ determining a trajectory in $\mathfrak{I}$ induces another Markov chain $\{j_\alpha^*(N)\}$ in $\mathfrak{I}/\equiv_\alpha$: The event $j_\alpha^*(N)$ only exists and is equal to $j^*(N)$ if and only if $k_{j^*}(N)<A\exp(-N^\alpha)$. Thus, the $\Lambda_\alpha$ dynamics may be followed using the projection scheme defined by the diagram for specific RNA molecules where the SF scenario has been proved to hold [3,9]. This is shown in Figs. 1 and 2.

For convenienece, we monitor in real time the number $\ln(f\langle t(\text{relax})\rangle)$, where $f\approx 10^6\ \text{s}^{-1}$ is the rate constant for

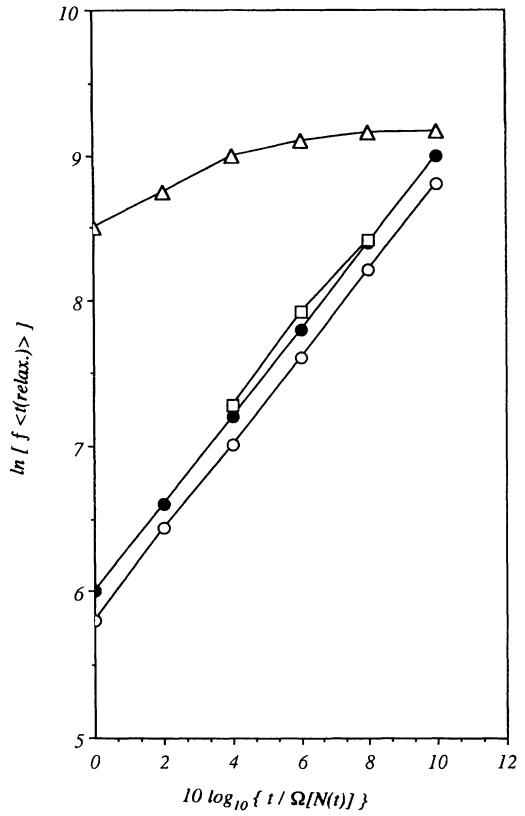


FIG. 1. Time dependence of the transition kinetic barrier when monitoring the dynamics in $\mathcal{I}/\equiv_\alpha$ for $Q\beta$ MDV1-RNA. The symbols $t$, $f$, $\langle t(\text{relax})\rangle$, and $\Omega[N(t)]$ denote real time, base-pair formation rate constant, expected relaxation time, and characteristic time scale, respectively. The REM behavior is revealed by the open-circle plot ($\alpha=0.25$), the filled-circle plot ($\alpha=0.28$), and the results of pulse-chase experiments indicated by open squares. The open-triangle plot ($\alpha=0.44$) reflects a high level of organization suggesting a correlated lower portion of the extended energy spectrum explored within large time scales.

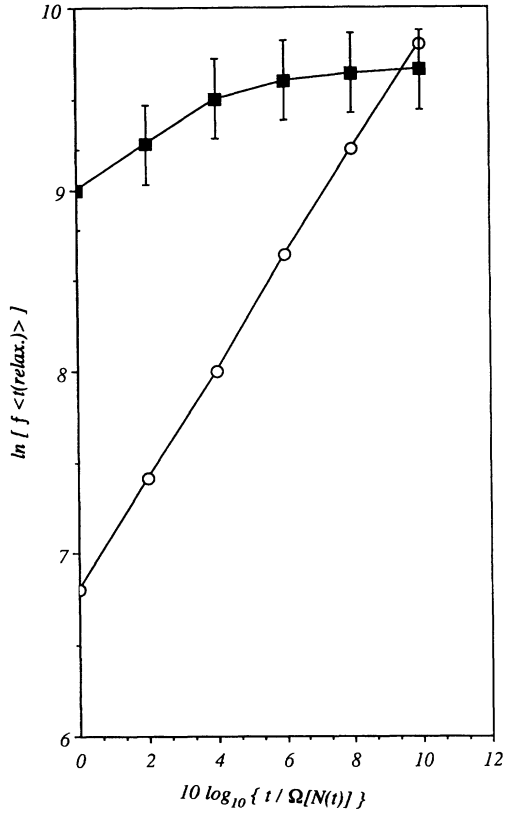


FIG. 2. Coarse-grained dynamics for the species *cob*I5. The same notation has been adopted. The open circles correspond to $\alpha=0.25$ and the filled squares correspond to $\alpha=0.35$.

single base-pair formation [3,5,7]. This quantity is proportional to the activation energy barrier $-\ln(k_{j*\alpha}(N)/A)$ of a transition in $\mathcal{I}/\equiv_\alpha$.

The results for the species $Q\beta$ MDV-1RNA ($N_0=220$) [1,3] are displayed in Fig. 1. The open and filled circle plots correspond to $\alpha=0.25$ and 0.28, respectively. The open squares are experimental results obtained by measuring the variable rate of chain elongation using pulse-chase techniques [1]. The chain elongation delay at specific sites along the RNA sequence [1] has been satisfactorily attributed to the occurrence of a refolding event, in accord with the simulations [3]. Thus the experimental results reported in [1] appear to correspond to a SF dynamics coarse-grained to the level $\alpha=0.28$. The logarithmic dependence of the activation barriers on real time is the signature of a REM-like relaxation which has been estimated to hold up to coarse-grainings of the order of $\alpha_{\text{crit}}\approx 0.31$ for this RNA species. Beyond this exponent, the kinetic barriers grow far more slowly than any multiple of the logarithm of real time, reflecting a considerable departure from REM behavior. This typical organized behavior is illustrated by the open-triangles plot corresponding to $\alpha=0.44$. This fact reveals the emergence of structural organization for larger time scales and a highly correlated lower portion of the extended energy spectrum which this species explores in longer times during its SF.

A similar behavior has been observed for the species *cob*I5, the fifth intron of yeast apocytochrome *b* gene [9], as shown in Fig. 2. Again, a REM behavior is detected for $\alpha=0.25$ (open circles), and a higher level of organization energies for more drastic coarse-graining at $\alpha=0.35$ (filled squares). The critical exponent has been estimated at $\alpha_{\text{crit}}\approx 0.27$ for this species.

The range of dynamic coarse-grainings of conformation space that yield REM dynamics is obviously dependent on the RNA primary sequence and its correlations, as the two examples above show. Thus for a purely random RNA sequence, we obviously have $\alpha_{\text{crit}}=0.5$. In the light of the results presented, we trust that dynamic coarse-graining will become an analytical tool in the general context of complex energy landscapes.

This work has been financially supported by the Camille and Henry Dreyfus Foundation (USA) and by Fundación Antorchas from Argentina.

[1] D. R. Mills, C. Dobkin, and F. R. Kramer, Cell **15**, 541 (1978).
[2] A. A. Mironov and A. Kister, J. Biomol. Struct. Dyn. **4**, 1 (1986).
[3] A. Fernández, Eur. J. Biochem. **182**, 161 (1989).
[4] B. Derrida, Phys. Rev. B **24**, 2613 (1981).
[5] A. Fernández, Phys. Rev. A **45**, R8348 (1992).
[6] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. USA **90**, 7195 (1993).
[7] V. V. Anshelevich, V. A. Vologodskii, A. V. Lukashin, and M. D. Frank-Kamenetskii, Biopolymers **23**, 39 (1984).
[8] D. H. Turner, N. Sugimoto, and S. M. Freier, Annu. Rev. Biophys. Biophys. Chem. **17**, 167 (1988).
[9] A. Fernández, A. Lewin, and H. Rabitz, J. Theor. Biol. **164**, 121 (1993).