Ascribing weights to folding histories: explaining the expediency of biopolymer folding



1994 J. Phys. A: Math. Gen. 27 6039

(http://iopscience.iop.org/0305-4470/27/18/014)

J. Phys. A: Math. Gen. 27 (1994) 6039–6052. Printed in the UK

# Ascribing weights to folding histories: explaining the expediency of biopolymer folding

Ariel Fernández

Instituto de Matemática, UNS-CONICET, Bahia Blanca 8000, Argentina and The Frick Laboratory, Princeton University, Princeton, NJ 08544, USA

Received 16 February 1994

**Abstract.** We develop a novel scheme of statistical inference whereby statistical weights are assigned to folding pathways. Evidence is presented that supports the fact that this scheme accounts for the robustness and expediency of biopolymer folding processes. The essential properties of folding are captured by showing that the weight is concentrated over a very limited domain of closely related folding pathways. To make probabilistic inferences, we constructively define a measure $\eta$ over the space of folding pathways. Such a scheme stands in contrast to traditional methods built upon a Boltzmann measure over conformation space. In order to implement and validate this new approach we combine analytical theory and computations that successfully reproduce pulse-chase kinetic experiments. We first present a rigorous analytical result by proving that an appropriate measure exists over the space of folding pathways. This existence theorem is shown to hold in two general scenarios: (i) the unbiased folding (UF) scenario, in which the complete chain starts its search in conformation space in an unbiased manner; (ii) the sequential folding (SF) scenario, in which the chain starts searching in conformation space concurrently with its own sequential assembling by progressive incorporation of monomers. A systematic coarse-graining simplification of the space of folding pathways is implemented to make the computations feasible and to validate our theory as a means of accounting for the expedient way of searching for the functionally competent conformation.

## 1. Weighting folding pathways

The dearth of theoretical approaches to explain the robustness and expediency with which a biopolymer chain finds its active folding is apparent, as burgeoning research indicates [1–5]. Thus, statistical-mechanical methods based upon the construction of a Boltzmann measure over conformation space [6] cannot account for the fact that the active structure is formed expeditiously under severe time constraints. The inadequacies of a thermodynamic approach rooted in stability considerations become obvious since time constraints force the chain to circumvent the so-called Levinthal scenario of random search in conformation space [7].

To address this problem, we have focused on recent evidence [1, 2, 8–10] that prompts us to introduce a measure $\eta$ on the space of folding pathways itself. This evidence indicates that, out of the manifold possibilities, the search in conformation space begets only a discrete and small number of competing folding pathways. For instance, in the context of RNA catalysis, recent experimental evidence [8] and computer simulations [9, 10] show that RNA cyclization at an internal position and RNA self-splicing are basically the only two processes pervasive in ribozyme (catalytic RNA) function governed by just two significant competing folding pathways. A meaningful measure should therefore be concentrated *exclusively* over

0305-4470/94/186039+14$19.50  

these pathways. Thus, the theoretical approach rooted in the construction of a measure $\eta$ should deal with the evaluation of integrals of the form

$$\Pr(A) = \int_A d\eta(\vartheta) \tag{1}$$

where a generic notation has been adopted in which $\vartheta$ denotes any folding pathway and $\Pr(A)$ indicates the probability of an event $A$ which is realized by an $\eta$-measurable bunch [11] (an open set in a suitable topology) $A$ of folding pathways. In the context of ribozyme function, the 'event A' might either be internal cyclization or RNA self-splicing.

In view of these considerations, the purview of this work is to establish the existence of a measure $\eta$ over the space of folding pathways [11] and to prove that the concentration of this measure is limited to a restricted domain. These facts account for the expediency and robustness of the search in conformation space. Moreover, such a measure will be defined *constructively*, based on the stochastic process whose realizations are the folding pathways themselves.

We shall consider only two scenarios in which the expediency of biopolymer folding is manifested. Although tentative, each addresses a different context and has been supported by experimental evidence using either denaturation–renaturation [6] or pulse-chase kinetic probes [12]. The two generic situations are:

(a) *The unbiased folding* (UF) scenario. Plausible folding events concurrent with the synthesis of the polymer chain do not bias the search for the destination structure since they are ultimately suppressed by the synthetic machinery which denaturizes structure and/or by the environment [3, 5, 6].

(b) *The sequential folding* (SF) scenario. The folding events that take place during the sequential assembling of the polymer chain bias the way in which conformation space is explored [4, 9, 10].

Both situations will be addressed by means of a theoretical approach in order to rigorously establish the existence of measure $\eta$. Obviously, the space of folding pathways is constructed differently depending on whether the context warrants the UF or the SF scenario. However, the rigorous proof of the main result is very similar in both cases. Thus, we shall present the proof of existence only for the UF scenario and show that it holds valid, *mutatis mutandis*, for the SF scenario.

This work is organized as follows: sections 2 and 3 deal with the analytical result that establishes the existence of a measure over the space of folding pathways under general conditions. Sections 4–7 deal with an actual computation of the measure distribution for a specific example: the SF of the species MDV-1 RNA, a small template that instructs the RNA synthesis performed by the enzyme $Q\beta$ replicase [4, 12]. Since hard facts revealed by pulse-chase experiments confirm the SF scenario for this system [12], we shall use this illustrative case to validate our novel approach to statistical inference by showing that the pathway that concentrates the measure is indeed the one detected in pulse-chase experiments. The complexity of the actual computation of the measure distribution is due to the fact that actual MD simulations may be run for at most 100 ps using state-of-the-art technology. This problem calls for a systematic coarse graining of conformation space, as implemented in sections 5 and 6.

The computation is essential to show that the measure $\eta$ is effectively concentrated on a very restricted and reduced manifold of experimentally probed folding pathways. This fact in itself suggests that a measure on the space of folding pathways indeed furnishes the

proper scheme of inference, as it accounts for the expediency and robustness of the folding process.

Obviously, establishing the validity of the theory in as broad a context as would be demanded for any statistical mechanical treatment will require laborious computations with case-dependent simplifications of the space of folding pathways, such as the coarse-graining procedure described in this work. Thus, although suggestive, the results of this work should be regarded as preliminary.

## 2. Describing the space of folding pathways

Throughout this section and the next one, we shall concentrate on the UF scenario in order to formally establish the existence of the measure $\eta$. The results are valid, *mutatis mutandis*, for the SF scenario, which will be studied computationally in sections 4–7. Although the space of folding pathways is different in the latter case, the properties of *separability* and *compactness*, necessary to establish the existence of the measure, still hold valid for the space of SF pathways, as shown in section 3. Thus, the existence theorem holds true in the SF context.

We consider a polymer chain made up of $N$ monomeric units whose conformation is defined by $M(N)$ degrees of freedom. Each of these internal variables corresponds to a dihedral angle representing rotation around a specific bond. Such bonds might be part of the backbone chain, like those forming the sugar–phosphate backbone of RNA, or might be inherent only to the internal conformation of each residue, as the glycosidic base–sugar bond of an RNA nucleotide [6]. Since vibrational degrees of freedom equilibrate on far shorter time-scales than rotational ones, it has been rightly assumed that rotational internal variables suffice to specify a polymer conformation [6].

Thus, we may consider in principle a conformation space $X$, which, given the angular nature of the degrees of freedom that specify a conformation, constitutes a torus of dimension $M(N)$:

$$X = M(N) - \text{torus}. \tag{2}$$

A folding pathway becomes a trajectory on $X$ defined by a map $\vartheta : I \to X$, where $I$ denotes a time interval. In the physically unrealistic case of an infinitely slow pathway made up of successively equilibrated states, the trajectory is determined entirely by thermodynamic or stability control. This means that the trajectory is tangent at point $x$ to the vector field $\Phi(x) = -\text{grad}_x U(x)$, where $U(x)$ is the potential energy functional. This potential, in turn, determines the Boltzmann measure on $X$, the object upon which classical methods of statistical inference are based [6].

In a more realistic context, the search in conformation space obeys a stochastic process $\xi : X \times I \to X$ which must be particularly robust since only a small assortment of destination structures occur reproducibly regardless of the initial state and perturbations of the folding pathways [4, 5, 9].

In accord with the introductory discussion, we shall focus on devising a proper scheme that will allow us to assign weights to folding pathways themselves. Thus, we need to introduce a proper space $\Theta$ containing all trajectories in $X$, define its topology $\Im(\Theta)$, and finally, endow it with a measure $\eta$ *induced* by the stochastic process $\xi : X \times I \to X$ which generates the trajectories.

Let $\Im(X)$ be the topology on $X$ induced by the metric topology $\Im(\Re^{M(N)})$ of $\Re^{M(N)}$ ($\Re$ = real numbers), the space in which $X$ is embedded. That is,

$$\Im(X) = \{A \cap X;\, A \in \Im(\Re^{M(N)})\}. \tag{3}$$

Let us now define a product topological space of copies or replicas of $X$ which contains in principle all continuous and discontinuous folding pathways with associated time span $|I|$:

$$Y = \prod_{t \in I} X_t \qquad X \equiv X_t. \tag{4}$$

Thus, $Y \supset \Theta$, where $\Theta = C(I \to X)$ is the space of continuous maps of the interval $I$ on $X$. This space $\Theta$ is endowed with the topology $\Im(\Theta)$ *inherited* from the product topology $\prod_{t \in I} \Im(X_t)$ of $Y$. Moreover, $\Theta$ is *naturally* endowed with a measure $\mu$ induced by the product Boltzmann measure $\mathrm{Pr}_B = \prod_{t \in I} \mu_{B,t}$ defined on $\wp(\prod_{t \in I} \Im(X_t))$, the minimal sigma algebra of sets generated by the product topology.

For every $x \in X$, let $\xi_x \in \Theta$ be a specific realization of the stochastic process $\xi : X \times I \to X$. This realization represents a specific folding pathway with associated time span $|I|$ starting with conformation $x$ at $t = 0$. The collection of such realizations constitutes a subset $\xi(X)$ of $\Theta$ which is comprised of all the folding pathways that are determined by the generating rules that define the stochastic process $\xi$ [4, 9].

It is not the purview of this section to actually specialize the map to any specific folding process [4, 5, 9]. It suffices to indicate that in the specific case where folding operates under time constraints and kinetic control governs the folding pathways, a realization $\xi_x$ may be defined and simulated computationally by means of the following general Markov process.

For each time $t \in I$, we define a map $t \to J(x, t) = \{j : 1 \leqslant j \leqslant n(x, t)\}$, where $J(x, t)$ = collection of elementary events representing conformational changes which are feasible at time $t$ given that the initial conformation $x$ has been chosen at time $t = 0$, and $n(x, t)$ = number of possible elementary events at time $t$. Associated with each event there is a unimolecular rate constant $K_j(x, t)$ = rate constant for the $j$th event [4] which may take place at time $t$ for a process that starts with conformation $x$. The mean time for an elementary refolding event is the reciprocal of its unimolecular rate constant. Thus, the only elementary events allowed are elementary refolding events that satisfy $k_j(x, t)^{-1} \leqslant |I|$.

At this point we may define the Markov process by introducing a Poissonian random variable $r \in [0, \sum_{j=1}^{n(x,t)} k_j(x, t)]$. let $r^*$ be a realization of $r$ such that if

$$\sum_{j=0}^{j^*-1} k_j(x, t) < r^* \leqslant \sum_{j=0}^{j^*} k_j(x, t) \qquad (k_0(x, t) = 0 \text{ for any } x, t) \tag{5}$$

then the event $j^* = j^*(x, t)$ is chosen at time $t$ for the folding process that starts with conformation $x$. Thus the map $t \to j^*(x, t)$ for fixed initial condition $x$ constitutes a realization of the Markov process which unambiguously determines the trajectory $\xi_x$.

## 3. The existence of a measure on the space of folding pathways

At this point we shall formulate and prove the following theorem.

The stochastic process $\xi$ indexed by a starting conformation $x \in X$ induces a measure $\eta$ on $\Theta$ which satisfies the relation

$$\eta A = \int_A \chi_{\xi(X)}(\vartheta)\, \mathrm{d}\mu(\vartheta) \tag{6}$$

where $\chi_{\xi(X)}(\vartheta) = 1$ if there exists $x \in X$ such that $\vartheta = \xi_x$, and $\chi_{\xi(X)}(\vartheta) = 0$ otherwise.

In precise terms, the $\mu$-measurable function $\chi_{\xi(X)}$ is the Radon–Nikodym derivative of $\eta$ with respect to $\mu$.

*Proof.* The space $X$ is compact when endowed with topology $\Im(X)$, thus, by Tikhonov's theorem, $Y$ is compact with the product topology, and $\Theta$ is also compact when endowed with the topology inherited from the product topology. Since $\Theta$ is also Hausdorff, we shall apply the Riesz–Markov representation theorem [11]. Consider the space of continuous functionals $C(\Theta)$, then, given a functional $F$ in the dual space $C(\Theta)^*$, there exists a measure $\eta$ on $\Theta$ such that

$$F(h) = \int_{\Theta} h(\vartheta)\,\mathrm{d}\eta(\vartheta) \qquad \text{for any } h \text{ in } C(\Theta). \tag{7}$$

Since there are no restrictions on $F$, we take

$$F(h) = \int_X \langle h(\xi_x)\rangle_x\,\mathrm{d}\mu_{\mathrm{B}}(x). \tag{8}$$

In equation (8), the symbol '$\langle\ldots\rangle_x$' denotes the average over the ensemble of realizations $\xi_x$ for fixed initial condition $x$. Thus, we have shown that $\eta$ is induced by the stochastic process $\xi$.

The measure $\eta$ may be constructed as follows.

Let $A \in \Im(\Theta)$, then we define its measure as

$$\eta A = \mathrm{Sup}\{F(h), 0 \leqslant h \leqslant 1, h \in C(\Theta), A \supset \mathrm{support}(h)\}. \tag{9}$$

This real functional defined on open sets may be canonically extended to a *regular* measure over $\wp(\prod_{t\in I} \Im(X_t) \cap \Theta)$ [11].

Consider now the set $D(A)$ of functionals $f(\vartheta)$ of the form

$$f(\vartheta) = \left\{ \int_I \chi_{\pi_t(A)}(\pi_t\vartheta) f(t) \exp(-\beta U(\pi_t\vartheta))\,\mathrm{d}t \right\} \Big/ |I| \int_X \exp(-\beta U(x))\delta x \tag{10}$$

where $\pi_t : Y \to X_t$ is the canonical projection, $\beta = 1/k_{\mathrm{B}}T$ ($T$ = temperature, $k_{\mathrm{B}}$ = Boltzmann constant), $0 \leqslant f(t) \leqslant 1$ is *any* continuous real function, $\chi_{\pi_t(A)}$ is the characteristic function of the projection of $A$ on the replica $X_t$, and $\delta x$ is the differential volume in conformation space $X$.

The set $D(A)$ is *dense* in $G(A) = \{0 \leqslant h \leqslant 1, h \in C(\Theta), A \supset \mathrm{support}(h)\}$ with respect to the norm determined by the measure $\mu$. Therefore, we have

$$\eta A = \mathrm{Sup}\{F(h), h \in D(A)\}. \tag{11}$$

This equation enables us to compute the measure of $A$, thus verifying equation (6):

$$\begin{aligned}\eta A &= \int_X \int_I \chi_{\pi_t(A)}(\pi_t\xi_x) \exp(-\beta U(\pi_t\xi_x))\,\mathrm{d}t\delta x \Big/ |I| \int_X \exp(-\beta U(x))\delta x \\ &= \int_A \chi_{\xi(X)}(\vartheta)\,\mathrm{d}\mu(\vartheta).\end{aligned} \tag{12}$$

This completes the proof of the theorem. □

The validity of this theorem in the case of the SF scenario for a chain of total length $N_0$ is apparent once we realise that the space $Y = \prod_{1\leqslant N\leqslant N_0}(M(N) - \text{torus})$ is in this case also Hausdorff and compact with the product topology.

## 4. A measure within the SF scenario

Kinetic experiments determining the uneven rate of RNA replication furnish compelling evidence that RNA structure is often formed under stringent time constraints [12–14]. This observation is corroborated by the metastable nature of the biologically competent emerging structure [14]. These facts have prompted the author to define an SF scenario in which the search for the structure starts during the very synthesis of the molecule [4, 14]. This scenario implies an exploration of conformation space concurrent with the sequential assembling of the RNA chain. In other words, the growing chain folds as it is being assembled and such early events inevitably bias the search for the active structure of the fully formed chain.

The SF scenario has been confirmed by reinterpreting the experiments that reveal a variable rate of chain elongation [12, 13]: it has been demonstrated that the experimentally probed pause sites during progressive template-instructed RNA replication are due to upstream refolding events concurrent with progressive elongation of the RNA chain. An SF pathway is a sequence of such events and has been confirmed to lead to the biologically active structure [4, 12–14]. Thus, computer simulations of SF have been readily turned into algorithms for structure prediction [15] of paramount interest whenever the structure is searched for under time constraints.

The previous analysis suggests that the Boltzmann weights assigned to foldings of the fully formed chain might not agree with the statistical weights resulting from SF. This is indeed the case whenever SF has been confirmed experimentally [4, 12–15]. The biologically active structure is the result of a kinetically controlled pathway which reflects the opportunistic search in a conformation space of increasingly higher dimension. Thus, the destination structure, being kinetically determined, is often metastable, unless the relevant experimental timescale is long enough to allow for full relaxation.

In this context an appropriate measure $\eta$ will be one that is essentially concentrated on the SF pathway probed by pulse-chase experiments whose destination structure coincides with the biologically competent folding.

A feasible construction of $\eta$ will demand a systematic simplification of the conformation space of increasingly higher dimensions. Thus, a suitable coarse graining will be implemented in section 2 so that rapidly interconverting secondary structures will be clustered together by means of an identification formalized as an equivalence relation. Within this representation, SF will be simulated computationally as a sequence of transitions between rapidly equilibrated clusters. Thus, each sequence of transitions will be considered to be a specific realization of a stochastic process. This stochastic process is actually the projection onto the coarse-grained space of a Monte Carlo-simulated stochastic process representing refolding events concurrent with chain elongation events [4, 14, 15].

## 5. Coarse-graining conformation space

We intend to characterize the complex dynamics of sequential folding for a specific RNA molecule $N_0$ monomers long by introducing a coarse graining of the extended conformation space $X$. The space $X$ contains all plausible secondary structures (intrachain base pair patterns subject to the Watson–Crick complementarity rules G–C, A–U) formed by segments of every length $N$, with $1 \leqslant N \leqslant N_0$. As might be obvious to the reader, we have altered the representational framework with respect to the one presented in sections 2 and 3. This, however, does not affect the validity of the existence theorems, since $X$ remains Hausdorff and compact.

The space $X$ itself may be viewed as a preliminary coarse graining of the formal reunion of coordinate spaces for all chain lengths. This construction is formalized dividing the reunion of coordinate spaces by an equivalence relation:

$$X = \bigcup_{1 \leqslant N \leqslant N_0} \Re^{3M(N)} / \approx \tag{13}$$

where $\Re$ is the set of real numbers and $\Re^{3M(N)}$ the conformation space for an RNA chain of length $N$ made up of $M(N)$ atoms. The equivalence relation '$\approx$' is defined as follows: two conformations $\sigma$ and $\rho$ adopted by RNA chains of lengths $N(\sigma)$ and $N(\rho)$, respectively, are regarded as equivalent ($\sigma \approx \rho$) if their secondary structure is identical, that is, if $\sigma$ and $\rho$ have the same base-pairing pattern.

We shall now simplify the dynamical description by first noting that as the elements in $X$ are regarded modulo low kinetic barriers of interconversion, the resulting dynamics of transitions between clusters of conformations follow a random energy model (REM) [16]. That is, if we group structures that interconvert on fast timescales of the order of $A^{-1} \exp(N^{1/4+\varepsilon})$, with $A = 10^3\ s^{-1}$ and $\varepsilon \geqslant 0$, the expected activation energy barriers for monitored transitions between equilibrated clusters grow logarithmically in real time [4].

The REM description will be shown to break down as the size of the 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 timescale 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 [17]). SF delivers the molecule to organized states only after REM-like equilibration has taken place within clusters of rapidly interconverting states.

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 $X$ contains all folded segments of various lengths regarded modulo their secondary structure. Thus, each equivalence class is labelled by a base-pairing pattern.

In order to represent the dynamics of sequential folding we shall now define a quotient space $X/\equiv_\alpha$ 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 $s, s' \in \mathfrak{S}$, then $s \equiv_\alpha s'$ if and only if

$$\begin{aligned} -\ln(k(s \to s')/A) &= O(N_{\min}(s, s')^\alpha) \\ -\ln(k(s' \to s)/A) &= O(N_{\min}(s, s')^\alpha) \end{aligned} \tag{14}$$

where $\frac{1}{4} \leqslant \alpha \leqslant \frac{1}{2}$, and $k(s \to s')$ is the unimolecular rate constant [8, 10] 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 $X$ 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 rigorous terms, if the map $\Lambda : X \to TX$ ($T =$

tangent bundle) denotes the vector field whose trajectories are the SF pathways, we are interested in describing the field $\Lambda_\alpha : X/\equiv_\alpha \rightarrow T[X/\equiv_\alpha]$, a map that makes the following diagram commutative:

$$\begin{array}{ccc} X & \xrightarrow{\Phi} & X/\equiv_\alpha \\ \Lambda\downarrow & & \downarrow\Lambda_\alpha \\ TX & \xrightarrow[T\Phi]{} & T[X/\equiv_\alpha] \end{array} \tag{15}$$

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:

$$\Lambda_\alpha \Phi = [T\Phi]\Lambda. \tag{16}$$

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

The $\Lambda$-dynamics has been simulated using kinetically controlled Monte Carlo methods [4, 14]. 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 [4, 14], and thus only the basic tenets are outlined.

For each value of the contour variable $N$ we define a map

$$N \rightarrow J(N) = \{j : 1 \leqslant j \leqslant n(N)\}$$

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 with each event there is a unimolecular rate constant $k_j(N)$ = 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 ($j \geqslant 2$) that should satisfy $k_j(N)^{-1} \leqslant t_{\text{exp}}$, where $t_{\text{exp}}$ is the experimental replication turnover timescale ($\approx$ 15 s for an RNA sequence 220 nucleotides long) [12, 14]. The mean time for an elementary refolding event is the reciprocal of its unimolecular rate constant. Since $X$ 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 [4, 18] and used extensively in our computations. Their associated kinetic barriers depend, respectively, on the enthalpic loss associated with helix formation and the entropy loss associated with loop closure. Thus, the compilation of thermodynamic parameters [19] 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 $r\varepsilon[0, \sum_{j=1}^{n(N)} k_j(N)]$ be a Poissonian random variable and let $r^*$ be a realization of $r$ such that if

$$\sum_{j=0}^{j^*-1} k_j(N) \leqslant r^* \leqslant \sum_{j=0}^{j^*} k_j(N) \qquad (k_0(N) = 0 \text{ for any } N) \tag{17}$$

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 $k_1(N(\text{reg})) = k_1(N(\text{pause})) = 50\ \text{s}^{-1}$ [4].

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

$$\langle t(\text{relax})\rangle = [k_{j^*}(N(\text{pause}))]^{-1}. \tag{18}$$

A Markov chain $\{j^*(1), j^*(2), j^*(3), \ldots\}$ determining a trajectory in $X$ induces another Markov chain $\{j^*_\alpha(N)\}$ in $X/\equiv_\alpha$: 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 proven to hold [10, 14]. This is shown in figures 1 and 2.

For convenience, 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 single base pair formation [4, 14, 18]. This quantity is proportional to the activation energy barrier $-\ln(k_{j^*\alpha}(N)/A)$ of a transition in $X/\equiv_\alpha$.

The results for the species Q$\beta$ MDV 1-RNA ($N_0 = 220$) [12, 14] are displayed in figure 1. The open and solid circles 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 [12]. The chain elongation delay at specific sites along the RNA sequence [12] has been satisfactorily attributed to the occurrence of a refolding event, in accord with the simulations [14]. Thus, the experimental results reported in [12] appear to correspond to an 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 behaviour. This typical organized behaviour is illustrated by the open triangles, corresponding to $\alpha = 0.44$. This fact reveals the emergence of structural organization for larger timescales and a highly correlated lower portion of the extended energy spectrum which this species explores in longer times during its SF.

A similar behaviour has been observed for the species *cob*I5, the fifth intron of yeast apocytochrome *b* gene [10], as shown in figure 2. Again, an REM behaviour is detected for $\alpha = 0.25$ (open circles), and a higher level of organization emerges for more drastic coarse graining at $\alpha = 0.35$ (solid 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 $\alpha_{\text{crit}} = 0.5$.

## 6. The measure over coarse-grained SF pathways

Given the $\Lambda_\alpha$ dynamics we now consider the space of $\Lambda_\alpha$ pathways, which we denote $\Theta_\alpha$:

$$\Theta_\alpha = \left( \prod_{N(\text{in}) \leqslant N \leqslant N_0} \mathfrak{N}^{3M(N)}/\approx \right)/\equiv_\alpha \tag{19}$$

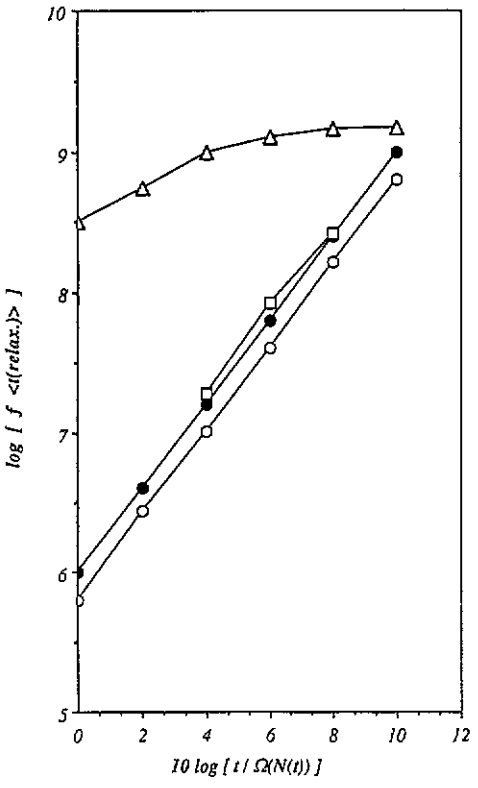


**Figure 1.** Time dependence of the transition kinetic barrier when monitoring the dynamics in $X/\equiv_\alpha$ for Q$\beta$ MDV1-RNA. The symbols $t$, $f$, $\langle t(\text{relax})\rangle$ and $\Omega(N(t))$ denote real time, base pair formation rate constant, expected relaxation time and characteristic timescale, respectively. The REM behaviour is revealed by the open circles ($\alpha = 0.25$) and the solid circles ($\alpha = 0.28$), and the results of pulse-chase experiments are indicated by open squares. The open triangles ($\alpha = 0.44$) reflect a high level of organization, suggesting a correlated lower portion of the extended energy spectrum explored within large timescales.

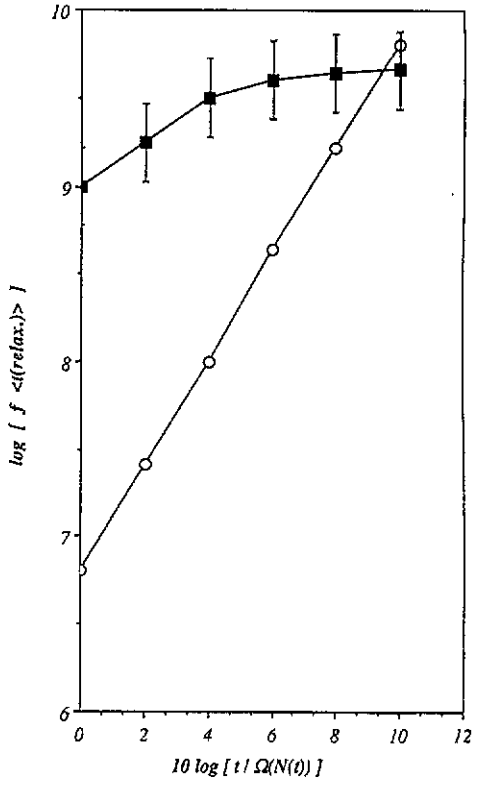


**Figure 2.** Coarse-grained dynamics for the species *cobI5*. The same notation as in figure 1 has been adopted. The open circles correspond to $\alpha = 0.25$ and the solid squares to $\alpha = 0.35$.

where $N(\text{in})$ is the minimal length for the cluster of structures that constitutes the set of initial conditions. Thus, an element $\theta_\alpha = (s_{N(\text{in})}, s_{N(\text{in})+1}, s_{N(\text{in})+2}, \ldots)$ of $\Theta_\alpha$ is a sequence of $N^\alpha$ clusters which determines unambiguously an SF pathway.

In order to construct a measure $\eta$ over $\Theta_\alpha$ we shall start by considering the Boltzmann measure $\mu_B$ defined over $X(\text{in}) = \Re^{3M(N(\text{in}))}/_\approx$:

$$\mathrm{d}\mu_B\,(s_{N(\text{in})}) = \exp(-\beta\varepsilon(s_{N(\text{in})}))\,\mathrm{d}v(s_{N(\text{in})})/Z(N(\text{in})). \tag{20}$$

In equation (20), $Z(N(\text{in}))$ is the partition function resolved up to secondary structure for the RNA chain segment of length $N(\text{in})$, $\varepsilon(s_{N(\text{in})})$ is the energy of the secondary structure $s_{N(\text{in})} \in X(\text{in})$ and $\mathrm{d}v(s_{N(\text{in})})$ is the differential volume in $X(\text{in})$ around $s_{N(\text{in})}$.

The measure defined by equation (20) induces a measure $\mu_\alpha$ over $X(\text{in})/\equiv_\alpha$ as follows:

$$d\mu_\alpha(s) = \sum_{s' \equiv_\alpha s} d\mu_B(s') \tag{21}$$

where $s$ in equation (21) is a representative structure of the cluster or class $s$.

At this stage we may define the coarse-grained measure $\eta$. Let $h \in C(\Theta_\alpha)$ be *any* continuous functional defined over $\Theta_\alpha$, then we may define a functional $F$ in the dual of $C(\Theta_\alpha)$, $F \in C(\Theta_\alpha)^*$, as follows:

$$F(h) = \int_{X(\text{in})/\equiv_\alpha} h(\theta_\alpha^*(s))\, d\mu_\alpha(s) \tag{22}$$

where

$$\theta_\alpha^*(s) = (s^{(0)} = s, s^{(1)}, s^{(2)}, \ldots). \tag{23}$$

In the sequence $\Theta_\alpha^*(s)$, $s^{(i)} \in s^{(i)}$ is the resulting structure starting from $s^{(i-1)} \in s^{(i-1)}$ after the event $j_\alpha^*(N(s) + i - 1)$ has taken place. If the event $j_\alpha^*(N(s) + i - 1)$ happens to be projected out of the $\Lambda_\alpha$ dynamics, then $s^{(i)}$ is taken to be identical to $s^{(i-1)}$, since they both belong to the same $N^\alpha$ cluster.

In accordance with the results of sections 2 and 3, and making use of Riesz's theorem [11], given the functional $F$, there exists a *regular* measure $\eta$ over the space $\Theta_\alpha$ satisfying

$$F(h) = \int_{\Theta_\alpha} h(\theta_\alpha)\, d\eta(\theta_\alpha) \qquad \text{for } \textit{any } h \text{ in } C(\Theta_\alpha). \tag{24}$$

Thus the measure $\eta$ defined by equation (24) is induced by the Boltzmann measure $\mu_\alpha$ on the space of initial conditions and the kinetically controlled stochastic process whose realizations are the SF pathways obtained for chosen initial conditions.

## 7. The SF pathways that concentrate the measure $\eta$

We shall illustrate and assess the power of our approach by determining the SF pathway for the species $Q\beta$ MDV-1 RNA [12–14] that concentrates the measure $\eta$, and show that the destination structure for this pathway coincides with the biologically active structure shown in figure 3.

The $\Lambda_\alpha$ dynamics are described here by an inverted tree where each vertex represents an $N^\alpha$ cluster and each edge represents the fastest transition between clusters. Each path along the tree corresponds to a realization of the stochastic process determined by a specific choice in $X(\text{in})/\equiv_\alpha$. The particular tree for $Q\beta$ MDV-1 RNA is displayed in figure 4 for $\alpha = 0.28$, the precise value for which the $\Lambda_\alpha$ dynamics reproduces the experimental data presented in [12] and displayed in figure 1.

The base of the tree represents the space of initial conditions $X(\text{in})/\equiv_\alpha$ for $N(\text{in}) = 25$. In this space we distinguish 11 clusters, each of which represents a different folding of the primer ($N = 25$) region. The primer region exhibits a very high level of Watson–Crick G–C self-complementarity, as can be seen in figure 3.

The fully formed hairpin that takes maximum advantage of this complementarity is displayed on the LHS of figure 3 and is denoted 'folded primer' in figure 4. The pathway indicated by the bold line resulting from successive refoldings and chain elongations built upon this structure leads to the active destination structure denoted $A$ and displayed in full

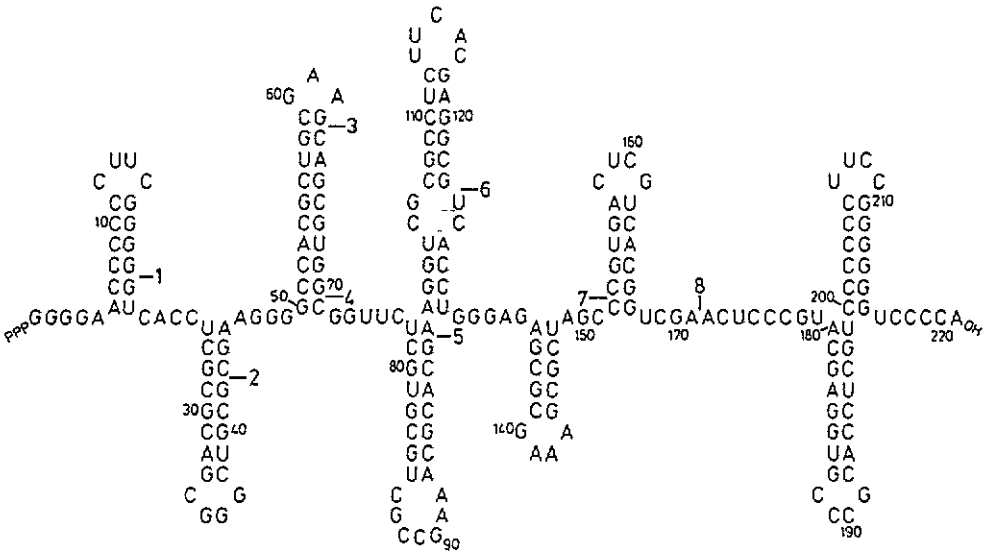

**Figure 3.** Active secondary structure for Qβ MDV-1 RNA, a template directing RNA synthesis by Qβ replicase. The sites indicated by numbers denote experimentally determined pause sites where the occurrence of an upstream refolding event causes the enzyme to idle for a relatively long period of time until the next nucleotide is incorporated. These sites are in perfect agreement with the computationally determined $N$(pause)s, where $j_\alpha^*(N(\text{pause})) \geqslant 2$, $\alpha = 0.28$. The structure displayed is identical to the destination structure which results from the pathway that concentrates 81% of measure $\eta$.

in figure 3. This structure is reached well within the experimental timescale coresponding to a replication turnover $t_{\text{exp}} \approx 15$ s [12]. Structure $A$ eventually relaxes to the global free energy minimum $I$, albeit in a longer and unrealistic timescale, as indicated in figure 4. This inactive structure is obtained by binding to each other the two 25 nucleotides long extremes of the molecule. The reader may notice that both extremes are highly complementary and thus their binding to each other is energetically (but not entropically) advantageous with respect to adopting the active conformation.

Of paramount importance is the fact that the pathway developed by starting with the folded primer concentrates 81% of the measure $\eta$ and coincides with the experimentally determined pathway for template-instructed Qβ MDV-1 RNA replication [12–14]. Within the uncertainties of the computation, the remaining 19% of measure $\eta$ is evenly distributed between the remaining pathways. The initial conditions leading to the latter pathways are represented by clusters of structures containing lower degrees of base pairing corresponding to slippage of the folded primer. Such slippage produces more unstable initial structures and the resulting pathways have each an ascribed $\eta$-weight of approximately 2%.

The extreme situation is encoutnered when the primer is totally unfolded. The resulting pathway is represented by a bold line on the RHS of figure 4. This pathway leads expeditiously to the inert global minimum since the primer is readily available to bind to the opposite extremity as soon as the molecule is totally assembled. This pathway concentrates 48% of the measure $\mu$, while the biologically relevant pathway concentrates 44%.

The inert nature of the destination structure for the most heavily $\mu$-weighted pathway is apparent: as the two extremities of the molecule are bound to each other, no initial signal

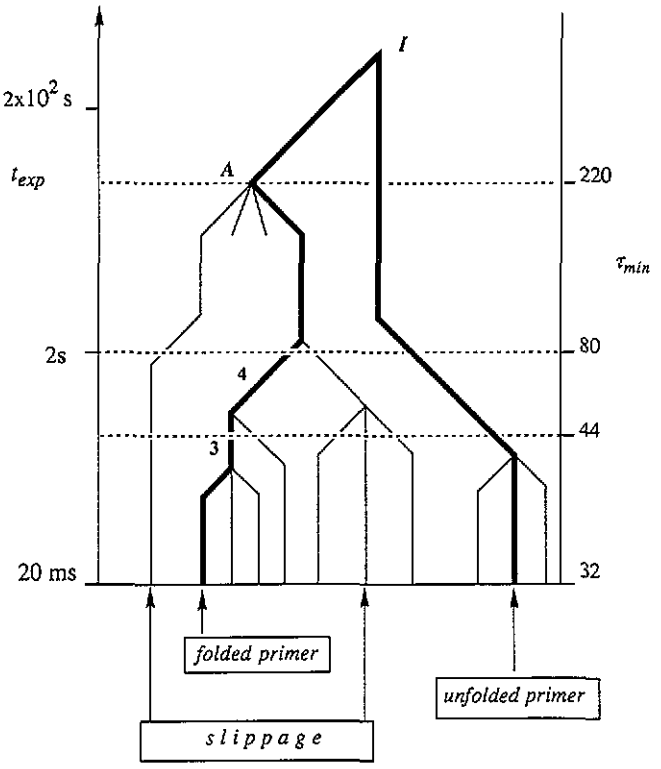


**Figure 4.** Graphic representation of the $\Lambda_\alpha$ dynamics with $\alpha = 0.28$ for the species Q$\beta$ MDV-1 RNA. The inverted tree represents the coarse-grained SF pathway space $\Theta_\alpha$. The vertical axis denotes real time, the horizontal lowest layer of vertices is the space $X(\text{in})/\equiv_\alpha$ of initial conditions and each edge in the graph represents the fastest transition between two $N^\alpha$ cluster.

for replication is possible [14, 15].

To summarize, this analysis reveals that the measure is $\eta$, and not the thermodynamically determined $\mu$, which truly captures the essential physical reality of the folding of sequentially assembled RNA structure. The measure $\eta$ should be regarded as the starting point to develop a new scheme of probabilistic inference suitable for biopolymer folding subject to realistic time constraints.

The fact that this measure is actually concentrated only upon experimentally confirmed pathways suggests that the approach presented in this work might indeed be promising as a means to account for the expediency and robustness of the folding process.

## Acknowledgments

Financial support for this work has been provided by the Camille and Henry Dreyfus Foundation through a Teacher-Scholar Award made to the author, and by Fundación Antorchas from Argentina. AF is principal investigator of CONICET, the National Research Council of Argentina.

## References


[1] Jaenicke R 1984 *Angew. Chem. Intl. Ed. Engl.* **23** 295
[2] Creighton T E 1988 *Bioessays* **8** 57
[3] Creighton T E 1988 *Proc. Natl Acad. Sci. USA* **85** 5082
[4] Fernández A 1990 *Phys. Rev. Lett.* **64** 2328; 1992 *Phys. Rev.* A **45** R8348
[5] Shakhnovich E I, Farztdinov G, Gutin A M and Karplus M 1991 *Phys. Rev. Lett.* **67** 1665
[6] Cantor C R and Schimmel P R 1980 *Biophysical Chemistry* (New York: Freeman)
[7] Levinthal C 1968 *J. Chim. Phys. (Paris)* **65** 44
[8] Partono S and Lewin A 1988 *Mol. Cell. Biol.* **8** 2562
[9] Fernández A 1992 *J. Theor. Biol.* **157** 487
[10] Fernández A, Lewin A and Rabitz H 1993 *J. Theor. Biol.* **164** 121
[11] Nelson E 1959 *Ann. Math.* **69** 630
[12] Mills D R, Dobkin C and Kramer F R 1978 *Cell* **15** 541
[13] Mironov A A and Kister A 1986 *J. Biomol. Struct. Dyn.* **4** 1
[14] Fernández A 1989 *Eur. J. Biochem.* **182** 161
[15] Fernández A 1993 *Phys. Rev.* E **48** 3107
[16] Derrida B 1981 *Phys. Rev.* B **24** 2613
[17] Shakhnovich E I and Gutin A M 1993 *Proc. Natl Acad. Sci. USA* **90** 7195
[18] Anshelevich V V, Vologodskii V A, Lukashin A V and Frank-Kamenetskii M D 1984 *Biopolymers* **23** 39
[19] Turner D H, Sugimoto N and Freier S M 1988 *Ann. Rev. Biophys. Biophys. Chem.* **17** 167