Towards an action principle governing biopolymer folding *in vitro*

Ariel Fernández

Instituto de Investigaciones Bioquímicas, INIBIBB,
Universidad Nacional del Sur, CONICET, CC857, and Instituto de Matemática, INMABB,
Universidad Nacional del Sur - CONICET, Avenida Alem 1253, Bahía Blanca 8000,
Argentina and The Frick Laboratory, Princeton University, Princeton, NJ 08544, USA

Received 17 February 1995

The exploration of conformation space performed by biopolymers is biased towards a confined region. This property is paramount in providing theoretical underpinnings of the time-constrained nature of folding. By introducing an *action principle* in the space of folding pathways, we show how the folding process is guided expeditiously within realistic time frames.

1. The need for an action principle governing exploration of conformation space

The search in conformation space performed by biological polymers that fold intramolecularly *in vitro* is expeditious once renaturation conditions are established in the environment [1,2]. The folding process leads effectively to an active structure within timescales far shorter than those that would be actually compatible with thermodynamic control. The context mentioned above suggests the existence of an *action principle* that governs or biases the search in conformation space, a space upon which a complex multi-minima energy landscape is constructed. Such rugged landscapes have been considered previously in the field of polymer folding [3]. *If such a variational principle actually exists, one must ultimately prove that each experimentally-probed folding pathway, constitutes an extreme of an action integral.*

In this work we provide a theoretical strategy that enables us to define a suitable action by means of a Lagrangian defined on the space of folding pathways. This Lagrangian is shown to be induced by a probability measure previously defined over the space of folding pathways. The measure weights systematically entire pathways and is actually the probability measure associated to a stochastic process [4,5]. The latter is shown to yield different realizations each of which corresponds to a different kinetically-controlled pathway. Here kinetic control refers to the fact that, given a specific state of the system, the weight of any *a priori* plausible transition depends on the height of the kinetic barrier to be surmounted in order to realize the transition. This stochastic process has been shown to reproduce experimentally-determined folding pathways in such a way that the pathway that carries the highest statistical weight is identical to the one that contains the experimentally-identified folding intermediates [4,6].

Once the Lagrangian structure of the stochastic process has been determined, the results are specialized to illustrate the convergence of folding pathways to a specific pathway whose destination secondary structure is known to be biologically competent [4,6]. The experimental counterpart of such results is available for selected RNA molecules enabling us to test the theoretical predictions. The illustrative example serves to show how the action principle underlies the search in conformation space, thus providing the theoretical underpinnings of expeditious folding under the severe time constraints which are relevant in the biological context.

2. A measure on the space of folding pathways

We introduce a general scheme to assign statistical weights to all *a priori* pathways in conformation space for a polymer that folds intramolecularly. Thus, we endow the space of kinetically-controlled pathways with a regular measure induced by a stochastic process built upon a rugged potential energy landscape. This process simulates the progressive and opportunistic exploration of basins of attraction of critical points. *The measure determines a scheme of statistical inference whereby the ensemble of physically-relevant kinetically-arrested states [6] becomes a cross-section of the ensemble of pathways at a fixed instant.*

To fix notation, the space of pathways or histories is the space Θ of continuous maps $\vartheta : I \rightarrow X$, where $I = [0, t']$ is the parametric time interval associated to the experimentally-relevant timescale and X denotes a compact conformation space. The type of inferences that one can make based on the ensemble of kinetic pathways depends on the evaluation of integrals of the form

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

where ϑ denotes a generic pathway and $\Pr(A)$ indicates the probability of an event which is realized by an η-measurable bunch A of pathways.

The measure η has been shown [5] to be determined via a representation theorem [7] by the Boltzmann measure μ_B defined on X and a stochastic process governing the exploration of basins by surmounting barriers in the molecular potential energy landscape $U : X \rightarrow \mathfrak{R}$ ($\mathfrak{R}$ = set of real numbers). The conformation space is best described as follows: Consider a polymer chain made up of N monomeric units. Each conformation of the chain may be specified by $M(N)$ degrees of freedom. If we view conformation space modulo high frequency motions (vibrational modes and planar angular distortion modes) which are averaged out on the timescales of folding events, the relevant internal variables correspond to dihedral angles representing rotations around specific bonds preserving planar angles and distances between adjacent atoms. Thus, conformation space constitutes a torus of dimension $M(N)$:

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

In several realistic contexts [4,6], the search in conformation space is severely time-constrained and has been shown to obey a stochastic process $\xi: I \to \{g: X \to X\}$ of interval I on the space of automorphisms (g's) of X. This process may be alternatively viewed as a multi-vector field such that each possible transition that starts at a point in X is assigned a vector whose length is related to the kinetic barrier associated to the transition. Since many transitions are possible from a given conformation onto adjacent basins, the field must be multi-vectorial. Whenever the exploration of conformation space takes place under time constraints and kinetic control governs the pathways, a realization ξ_x, that is, one of the integral curves of ξ that starts in x, may be defined by means of a general Markov process which is defined below [4,6]. This is done in such a way that the stochastic process ξ becomes the continuous extension of a diffusional process defined over the network of critical points of the potential U:

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 or transitions between minima of U 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 to each event, there is a unimolecular rate constant $k_j(x, t)$ = rate constant for the jth event [4–6] which may take place at time t for a process that starts with conformation x. The mean time for an elementary event (transition between two free energy minima) is the reciprocal of its unimolecular rate constant. Thus, the only elementary events with significant probability are elementary events that satisfy: $k_j(x, t)^{-1} \leqslant |I|$.

To compute the unimolecular rate constant $k_j(x, t)$ we assume that the jth event corresponds to the transition $x_j \to x_j'$, where x_j and x_j' represent respectively the initial and final stable conformations for the jth event. Thus, an Arrhenius-type derivation [4–6,8] of the thermally-induced mean passage rate yields in the parabolic approximation:

$$k_j(x, t) = A \exp\{-\beta[U(x_j'') - U(x_j)]\}. \tag{3}$$

Where A is the preexponential Arrhenius factor, x_j'' is the critical point of U corresponding to the top of the barrier that must be surmounted to realize the transition $x_j \to x_j'$; $\beta = 1/k_B T$ and k_B is the Boltzmann constant.

At this point we may define the Markov process by first discretizing time and conformation space. The construction is made in such a way that the collection of events which are plausible at time t depends *exclusively* on the state or conformation of the system at time $t-1$. The state of the system at time $t-1$ is obviously determined by the sequence of choices made starting with conformation x at time 0. However, *the mechanism for choosing the event at time t* must not depend on the pathway that led to the conformation at time $t-1$. To devise the mechanism for choosing the event at generic time t, we introduce a random variable $r \in [0, \sum_{j=1}^{n(x,t)} k_j(x,t)]$ uniformly distributed over the interval. Let r^* be a realization of r at time t. Then there exists an event j^* such that r^* satisfies the inequalities

$$\sum_{j=0}^{j^*-1} k_j(x,t) < r^* \leqslant \sum_{j=0}^{j^*} k_j(x,t)\,, \tag{4}$$

$$(k_0(x,t) = 0 \quad \text{for any } x, t)\,.$$

In this case the event $j^* = j^*(x,t)$ is chosen at time t for the process that starts with conformation x. *This mechanism fulfills the Markovian tenets.* Thus the map $t \rightarrow j^*(x,t)$ for fixed initial condition x constitutes a realization of the Markov process which defines a series of transitions between minima. Again, the choice at time t was made based on the rates of all plausible events which the conformation at time $t-1$ might undergo. This conformation, in turn, was the result of a sequence of choices following the same mechanism defined by eq. (4) starting with conformation x at time $t=0$. The entire sequence of transitional events is continuously extendible to ξ_x [5].

To do statistical mechanics on folding pathways we need to construct an ensemble. This program requires endowing the space described above with a measure. In this regard, it has been shown [5] that the stochastic process ξ, jointly with the Boltzmann measure μ_B over the space of initial states $X_0 = X$, induces a measure η on Θ. Succintly, this measure is constructed as follows: We first consider the space of continuous functionals or actions $C(\Theta)$ (a functional over a space S associates a real number to each object of this space), then, given a linear functional F over $C(\Theta)$, there exists a measure η on Θ [7] such that

$$F(h) = \int_\Theta h(\vartheta)\, d\eta(\vartheta), \text{for } \textit{any } h \text{ in } C(\Theta)\,. \tag{5}$$

Since there are no restrictions on F, we take

$$F(h) = \int_X \langle h(\xi_x)\rangle_x \, d\mu_\mathrm{B}(x) = \int_X \Big[\sum_{\xi_x} h(\xi_x) p_x(\xi_x)\Big]\, d\mu_\mathrm{B}(x)\,. \tag{6}$$

In eq. (6), the symbol "$\langle \ldots \rangle_x$" denotes the average over the ensemble of realizations ξ_x for fixed initial condition x. This average is determined by the probabilities of the type $p_x(\xi_x)$, the probability that the pathway ξ_x will be realized if we start with conformation x. For fixed x, each realization is weighted according to the probabilities of the events chosen for every t. Given that the probability that event j occurs at time t is $k_j(x, t)/\sum_{j'\in J(x,t)} k_{j'}(x, t)$, the actual probability $p_x(\xi_x)$ is given by

$$p_x(\xi_x) = \prod_{j^*=j^*(t)} \left[k_{j*}(x,t) / \sum_{j'\in J(x,t)} k_{j'}(x,t) \right], \tag{7}$$

where the set $\{j^* = j^*(t)\}$ is the set of chosen events that defines ξ_x.

Thus, we have shown that η is induced by the stochastic process ξ. Explicitly, given a bunch A of trajectories, its weight or measure is

$$\eta A = \{\text{Supreme of the } F(h) \text{ with } 0 \leqslant h \leqslant 1, h \in C(\Theta) \text{ and } A \supset \text{ support}(h)\}, \tag{8}$$

where support $(h) =$ set of the limits of sequences of trajectories w's for which $h(w) \neq 0$.

3. Constructing the action over the space of folding pathways

At this point we shall construct a Lagrangian based on the measure η over the space of folding pathways. We proceed as follows: Let $\mathfrak{D}$ denote a disc of dimension $M = M(N)$: $\mathfrak{R}^M \supset \mathfrak{D}$; consider monoparametric families of smooth maps Φ_t: $\mathfrak{D} \rightarrow X$, known as families of embeddings; then the space of all such embeddings and their tangent vectors constitutes the so-called principal fiber bundle TP: TP $= \{\Phi, \Phi'\}$ ($\Phi' =$ tangent vector to Φ). At this point we define the lagrangian $\mathcal{L}$: TP $\rightarrow \mathfrak{R}$ over the principal fiber bundle induced by the measure η: Let us denote by $A_\Phi(t)$ the tube $A_\Phi(t) = \prod_{0\leqslant t'\leqslant t} \Phi_{t'}\mathfrak{D}$, and $\eta_t =$ restriction of η to $\prod_{0\leqslant t'\leqslant t} X_{t'}$ (the $X_{t'}$'s are identical copies of X indexed by the parameter t'), then

$$\mathcal{L}(\Phi_t, \Phi_t') = \int_{\mathfrak{D}} L(\Phi_t(y), \Phi_t'(y))\, d^M y = \lim_{\Delta\to 0} -\Delta^{-1}[\eta_{t+\Delta} A_\Phi(t+\Delta) - \eta_t(A_\Phi(t))], \tag{9}$$

where L is the Lagrangian defined on the space of folding pathways which induces $\mathcal{L}$. If we impose the condition

$$\min_{\{\xi_x\}} \int_I L(\xi_x(t), \xi_x'(t))\, dt = \int_I L(\xi_x^*(t), \xi_x^{*\prime}(t))\, dt, \tag{10}$$

where ξ_x^* is the most probable realization of the stochastic process starting with x, we obtain

$$L(x, x') = 1/2(\text{sign}\, u' + 1) u/c\, d/dt\, [\exp(u/c)], \tag{11}$$

where $U(x(t)) = u(t)$; $u' = U_x x'$ and $c = N^{1/2} k_B T$. The subsidiary condition is

$$\int_I S(x(t), x'(t))\, dt = \text{constant}, \tag{12}$$

where $S(x, x') = 1/2(\text{sign}\, u' + 1)u'/c$.

The actual computation of an action requires that we introduce the following notation:

∂I^+ = reunion of boundaries of the subintervals of I in which $u'(t) \geqslant 0$;

$B_i = u(t_{i+1}) - u(t_i) = i$th barrier to be surmounted along the pathway $x(t)$.

Thus, the action along a generic pathway $x(t)$ is given by

$$\begin{aligned}
&\int_I L(x(t), x'(t))\, dt \\
&= \sum_{t_i \in \partial I^+} \sum_{p \geqslant 2} [u(t_{i+1})^p - u(t_i)^p]/(p-1)!c^p \\
&= \sum_{t_i \in \partial I^+} \sum_{p \geqslant 2} [u(t_{i+1}) - u(t_i)] \left[\sum_{k=1,2,\ldots,p} u(t_{i+1})^{p-k} u(t_i)^{k-1} \right] / (p-1)!c^p \\
&= \sum_{t_i \in \partial I^+} \sum_{p \geqslant 2} B_i \left[\sum_{k=1,2,\ldots,p} u(t_{i+1})^{p-k} u(t_i)^{k-1} \right] / (p-1)!c^p .
\end{aligned} \tag{13}$$

This action defined by the Lagrangian L favors pathways with the lowest barriers within a family of pathways {smooth map: $I \to X$} satisfying the isoperimetric condition:

$$\text{Sum of kinetic barriers along pathway } x(t) = \int_I S(x(t), x'(t))\, dt = \text{constant}.$$

To prove this crucial property it suffices to consider two generic pathways (all energies are given in c-units):

(I) $X(t)$ involves a single barrier of height $n\Delta$ starting at an energy level with energy e and ending at an energy level with energy e.

(II) $x(t)$ involves n identical barriers of height Δ separating wells with zero point energy e starting and ending at the same states as pathway $X(t)$.

In this generic case we obtain

$$\begin{aligned}
&\int_I L(x(t), x'(t))dt = 2en\Delta + n\Delta^2 + O(\Delta^3) \\
&\qquad < \int_I L(X(t), X(t)')dt = 2en\Delta + n^2\Delta^2 + O(\Delta^3) .
\end{aligned} \tag{14}$$

Thus, within a family of pathways for which the sum of all barriers is a constant, the Lagrangian favors the pathway involving the lowest barriers regardless of their number.

4. Verifying the results for a specific RNA

A quantitative analysis of the results requires graphic representation. For this purpose we introduce the base-pair probability matrix (BPPM) $P_{ab}(t) = \int_{\Theta} M_{ab}(\vartheta(t))\, d\rho(\vartheta)$, where $\rho(A_{\Phi}(t)) = \int_0^t \mathcal{L}(\Phi_{t'}, \Phi'_{t'})\, dt'$ and $M_{ab}(x) = 1$ if a pairs with b in structure x and $= 0$ otherwise. That is, we choose a cross section of the ensemble obtained by fixing the indexing time parameter at a particular value t. Each $a - b$ entry in the matrix $P_{ab}(t)$ represents the probability for monomer a to pair with monomer b (a, $b = 1, 2, \ldots, N$), within secondary structures weighted according to the action defined. Since the BPPM is symmetric, the ensemble of structures weighted according to L will be conventionally represented in the upper right triangle of a square N×N matrix and the active structure, in the lower left triangle.

To compute the BPPM for an RNA species that folds intramolecularly *in vitro*, we make use of a compilation of thermodynamic parameters [9] and use it to generate the kinetic barriers associated to the formation and dismantling of hairpins [4,12]. The activation energy barrier for the rate-determining step in the formation of a hairpin [4,10] is known to be $-T\Delta S$(loop), where ΔS(loop) indicates the loss

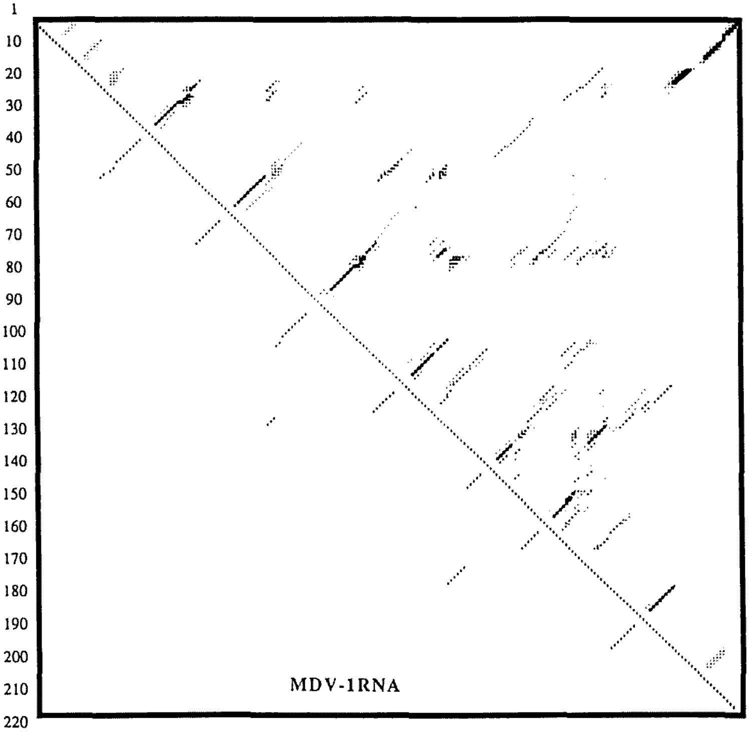

Fig. 1. The 221×221 BPPM $P_{ab}(t)$ for the species QβMDV-1RNA at real times $t = 0$ s. The upper right triangle of the matrices represent the ensembles generated by the lagrangian L at different times. The active structure is represented in the lower left triangle for comparison. Notice the convergence of the nonstationary pathways to a single destination structure identical to the active structure.

of conformational entropy associated to closing a loop. On the other hand, the activation energy barrier associated with the melting of a hairpin is $-\Delta H$(stem), the amount of heat released when forming all intramolecular contacts in the stem. The unimolecular rate constants for helix decay and helix formation have been obtained in analytical form [4,8,10] 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 [4,8,10].

The compilation of rate constants is built upon a given primary sequence. This requires prior elucidation of all *a priori* plausible no-knotted secondary structures associated to the sequence, a relatively canonical combinatorial problem. The sequence indicates the position along the chain of the residues of four types denoted A, U, G, C, where A = adenine, U = uracil, G = guanosine and C = cytosine. Each secondary structure is determined by identifying complementary regions following the Watson–Crick binding scheme: A–U, G–C.

The time evolution of the BPPM has been monitored for the species QβMDV1-RNA, a template for the technologically-crucial enzyme Qβ-replicase [4,11]. The BPPM has been computed at $t = 0$ s, $t = 10$ s and $t = 15$ s in real time, a realistic time frame for the folding of QβMDV-1RNA. The results obtained adopting a thermodynamic ensemble at the starting point (each structure is initially weighted

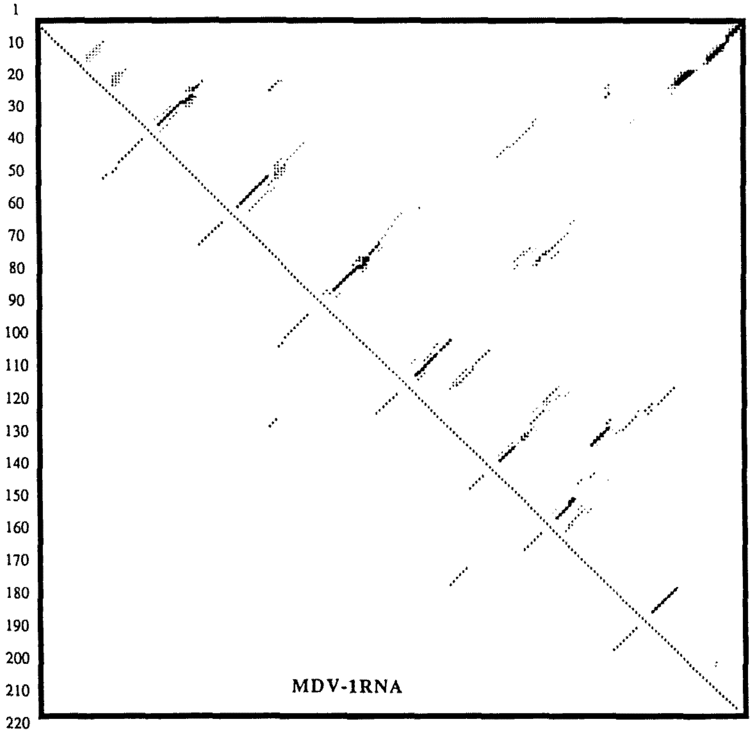

Fig. 2. Same as fig. 1 but for real time $t = 10$ s.

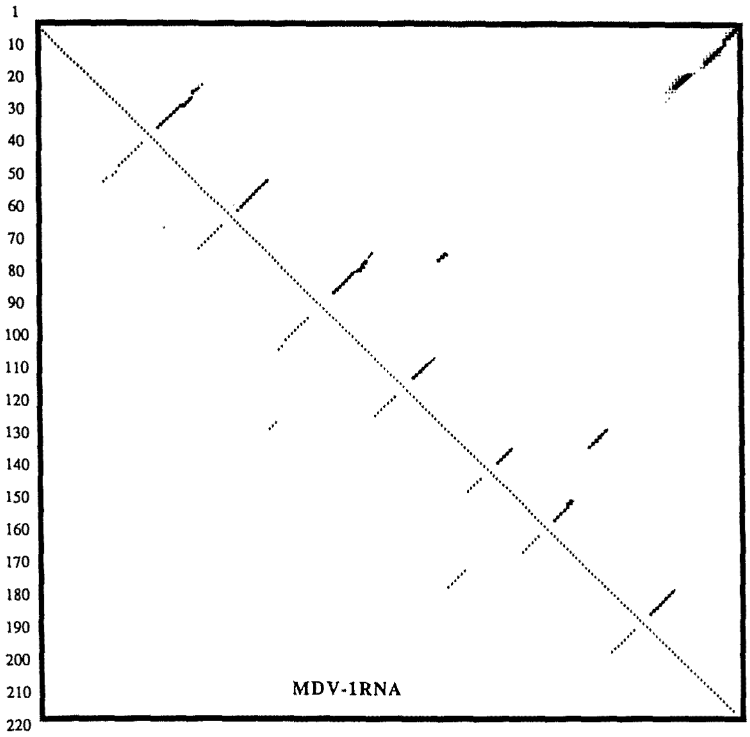

Fig. 3. Same as fig. 1 but for real time $t = 15$ s.

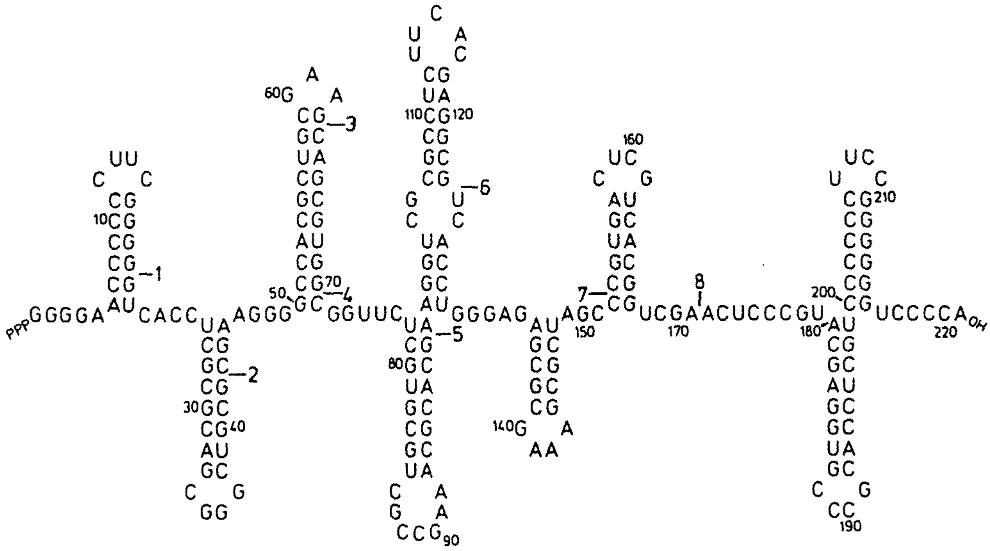

Fig. 4. The active secondary structure for Qβ MDV-1RNA [4,6,11].

according to the Boltzmann measure) are displayed in figs. 1–3, respectively. We must distinguish two types of pathways. One is stationary and starts at the global free energy minimum, the conformation for which the two highly complementary extremities (see fig. 4) are bound to each other [4]. The other pathways start at metastable structures and are therefore, nonstationary. All nonstationary pathways converge to the experimentally-determined active secondary structure shown in fig. 4 [4,11], as direct inspection of figs. 1–3 reveals. This result supports the existence of an action principle guiding the exploration in conformation space. Moreover, it supports the conjecture that the pathway whose destination structure is biologically relevant is actually the extreme of the action integral.

Acknowledgements

The author gratefully acknowledges financial support from Fundación Antorchas, Argentina, and from the J.S. Guggenheim Memorial Foundation. AF is the principal investigator of CONICET, the National Research Council of Argentina. During his sporadic stays at Princeton, funds from the Camille and Henry Dreyfus Foundation have been used.

References

[1] R. Jaenicke, Angew. Chem. Intl. Ed. Engl. 23 (1984) 295.
[2] T.E. Creighton, Bioessays 8 (1988) 57; Proc. Natl. Acad. Sci. USA 85 (1988) 5082; E.O. Purisima and H.A. Scheraga J. Mol. Biol. 186 (1987) 697.
[3] P.G. de Gennes, J. Stat. Phys. 12 (1975) 463.
[4] A. Fernández, Eur. J. Biochem. 182 (1989) 161; Phys. Rev. A: Rapid Comm. 45 (1992) R8348; Phys. Rev. Lett. 64 (1990) 2328.
[5] A. Fernández, J. Stat. Phys. 77 (1994) 1079.
[6] A. Fernández, Physica A201 (1993) 557.
[7] E. Nelson, Ann. Math. 69 (1959) 630.
[8] N.G. van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1984).
[9] D.H. Turner, N. Sugimoto and S.M. Freier, Ann. Rev. Biophys. Biophys. Chem. 17 (1988) 167.
[10] V.V. Anshelevich, V.A. Vologodskii, A.V. Lukashin and M.D. Frank-Kamenetskii, Biopolymers 23 (1984) 39.
[11] D.R. Mills, C. Dobkin and F.R. Kramer, Cell 15 (1978) 541.