Journal of Mathematical Chemistry 20 (1996) 95–116

An action principle for biopolymer folding *in vitro*: A new perspective on the design of expeditiously-folded RNA molecules

Ariel Fernández [a,b] and Gustavo Appignanesi [a]

[a] *Instituto de Matemática de Bahía Blanca (INMABB), Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional del Sur, Av. Alem 1253, Bahía Blanca 8000, Argentina*

[b] *The Frick Laboratory, Princeton University, Princeton, NJ 08544, USA*

Received 2 August 1995; revised 17 April 1996

The exploration of conformation space performed by a biopolymer becomes rapidly biased towards a confined region and takes place under a stringent schedule incompatible with the thermodynamic limit. The theoretical underpinnings of such properties have been missing to a considerable extent. By introducing an *action principle in the space of folding pathways*, we show how folding is guided expeditiously within realistic time frames. The variational principle is constructed in three stages: (a) An appropriate space of folding histories is defined. (b) The space is endowed with a measure and, in this way, an ensemble is defined. (c) This measure induces a Lagrangian which, in turn, defines the underlying action principle. The theory is specialized to account for the expeditious folding of an RNA species resolved to the level of secondary structure. Thus, using the Lagrangian, a time-dependent Base-Pair Probability Matrix (BPPM) is generated. This representational tool is introduced to display all RNA structures contributing to the cross section of the ensemble of pathways at each instant in time. The BPPM is contrasted *vis-a-vis* experimental information on biologically-competent RNA conformations. The results reveal that the statistical weight is concentrated on a very limited domain of folding pathways which yield the biologically-relevant destination structure within realistic timescales. To conclude, we assess in a preliminary fashion the potential of the action principle as a tool to aid the design of RNA species capable of folding within experimental timeframes.

1. The need for an action principle governing the 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. This stringent schedule of folding events often leads to metastable conformations [1–6], which prompts us to think that predictive algo-

rithms of the active structure should incorporate other principles besides stability criteria.

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 potential energy landscape is constructed. Such rugged landscapes have been considered previously by de Gennes in the field of polymer folding [3]. If such a variational principle holds and any random search scenario is to be relinquished, one must ultimately be able to prove that an experimentally-probed folding pathway constitutes an extreme of an action integral. Evidence along these lines is supplied in this work, although a vast task lies ahead before the action principle can be implemented with complete versatility.

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 to be defined over the space of folding pathways. This measure weights systematically entire pathways and is actually the probability measure associated with 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 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 relevant [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.

As we have discussed, constructing the Lagrangian requires that we previously endow the space of folding pathways with a measure [7]. This problem will be dealt with in sections 2–4. The Lagrangian will be introduced in section 5 and the theory will be critically evaluated *vis-a-vis* experimental facts in section 6. Section 7 will be devoted to assess, at least in a preliminary fashion, the potential of the Lagrangian approach as a means of aiding the design of expeditiously-folded RNA species. Thus, different recombinant RNA species will be studied to assess their ability to fold effectively and the results will be contrasted with experimental evidence on the species selected in an *in vitro* evolutionary reactor.

2. Assigning weights to folding histories

Statistical-mechanical methods based upon the construction of a Boltzmann measure over conformation space cannot generally account for the fact that the active structure is a suboptimal folding formed expeditiously under severe time constraints in most significant biological contexts [1,2,4–6,8–11].

To address this problem, we focus on recent evidence and observations on the functional relevance of folding pathways [1,2,8–11]. This viewpoint stands in contrast with the pervasive structure-function relationship and prompts us to introduce a measure η on the space of folding pathways itself. Thus, we shall construct an ensemble of folding histories upon which a statistical scheme will be defined.

Existing studies suggest that, out of the burgeoning possibilities, the search in conformation space begets in reality only a discrete and small number of often competing folding pathways [1,2,4,8–10]. Thus, a good theory must be primarily concerned with proving that the measure is concentrated on a very limited domain of the space of folding histories.

For instance, in the context of RNA catalysis, recent experimental evidence [8,9] and computer simulations [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 each by a single significant folding pathway. Thus in this context, a meaningful theory should warrant that the measure over the space of folding pathways be concentrated exclusively over the catalytically-relevant pathways.

In general, the type of inferences that one can make based on the ensemble of folding pathways is contingent upon the evaluation of integrals of the form

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

Here a generic notation has been adopted in which ϑ denotes any folding pathway and $\Pr(A)$ indicates the probability of an event A which is realized by an η-measurable bunch [7] (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.

The aims stated above are too vast to be dealt with in general. Eventually we shall specialize our results to the context of RNA folding, where satisfactory dynamic modeling of folding events has proven possible [4,9].

The purview of this work is to establish the existence of a measure η over the space of folding pathways [7] and to prove that the concentration of this measure is limited to a restricted domain of biological significance. These properties by themselves can 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 used to model time-dependent folding resolved up to secondary

structure [4,9–11], a stochastic process whose realizations are the folding pathways themselves.

3. Describing the space of folding pathways

We consider a polymer chain made up of N monomeric units whose conformation is defined by $M(N)$ degrees of freedom. Since the inherent timescales for vibrational degrees of freedom and planar angular distortions are far shorter than those associated to torsional degrees of freedom, it can be rightly assumed that torsional dihedral variables suffice to specify a polymer conformation. Thus, each of the internal variables represents a rotation around a specific bond regarding the remaining molecular frame as a rigid body. The bonds considered might be part of the backbone chain, like those forming the sugar-phosphate backbone of RNA, or might be inherent to the internal conformation of each residence, as the glycosidic base-sugar bond of an RNA nucleotide.

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.

In a more realistic context, the search in conformation space obeys a stochastic process $\xi: I \to \{\text{Automorphism on } X\}$, (we denote $\xi(t) \in \text{Aut}(X)$), 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 [10,11].

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 Θ containing all trajectories in X, define its topology $\mathcal{T}(\Theta)$, and finally, endow it with a measure η *induced* by the stochastic process ξ which generates the trajectories.

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

$$\mathcal{T}(X) = \{A \cap X; A \in \mathcal{T}(\mathcal{R}^{M(N)})\}. \tag{3}$$

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

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

Thus, $Y \supset \Theta$, where $\Theta = C(I \to X)$ is the space of continuous maps of the interval I on X. This space Θ is endowed with the topology $\mathcal{T}(\Theta)$ *inherited* from the product topology $\prod_{t \in I} \mathcal{T}(X_t)$ of Y. Moreover, Θ is *naturally* endowed with a measure μ induced by the product Boltzmann measure $\Pr_B = \prod_{t \in I} \mu_{B,t}$ defined on $\wp(\prod_{t \in I} \mathcal{T}(X_t))$, 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 ξ: $X \times I \to X$. This realization represents a specific folding pathway with associated timespan $|I|$, starting with conformation x at $t = 0$. The collection of such realizations constitutes a subset $\xi(X)$ of Θ which is comprised of all the folding pathways that are determined by the generating rules that define the stochastic process ξ [4].

It is not the purview of this section to specialize the map ξ to any specific folding process [4], that aspect will be dealt with in the illustrative sections 6 and 7. Here it suffices to indicate that in the specific case where folding is subject to time constraints and kinetic control is exerted, a realization ξ_X may be computed 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 to each event, there is an unimolecular rate constant $k_j(x, t)$ = rate constant for the jth event [4] which may take place at time t for a process that starts with conformation x. The mean time for an elementary refolding events 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 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 such that if

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

then the event $j^* = j^*(x, t)$ is chosen at time t for the folding process that starts at 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 ξ_X.

4. Proving the existence of a measure on the space of folding pathways

To do statistical mechanics on folding pathways we need to construct an appropriate ensemble. This program requires endowing the space described above with a measure. In this regard we shall formulate and prove the following theorem:

THEOREM

The stochastic process ξ induces a measure η on Θ which satisfies the relation:

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

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 μ-measurable function $\chi_{\xi(X)}$ is the Radon–Nikodym derivative of η with respect to μ.

Proof

The space X is compact when endowed with topology $\mathcal{T}(X)$, thus, by Tikhonov's theorem, Y is compact with the product topology, and Θ is also compact when endowed with the topology inherited from the product topology. Since Θ is also Hausdorff, we shall apply the Riesz–Markov representation theorem [7]. First we consider the space of smooth real-valued functions over Θ. This space is denoted $C(\Theta)$ and it consists physically of all possible smooth actions. This space strictly contains the set of all smooth path integrals. The representation theorem asserts that given a linear functional F over $C(\Theta)$, that is, a smooth correspondence between actions and real scalars, there exists a measure η on Θ such that

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

In other words, the scalar $F(h)$ could be regarded as an expectation value of h with respect to η, and this identification is valid for any action h.

Since there are no restrictions on F, we take

$$F(h) = \int_X \langle h(\xi_x)\rangle_x \, d\mu_B(x) = \int_X \left[\sum_{\xi_x} h(\xi_x) p_x(\xi_x)\right] d\mu_B(x)\,. \tag{8}$$

In eq. (8), the symbol "$\langle\cdots\rangle_x$" 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

$$p_x(\xi_x) = \prod_{j^*=j^*(t)} \left[k_{j^*}(x,t) \Big/ \sum_{j' \in J(x,t)} k_{j'}(x,t) \right]. \tag{9}$$

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.

The measure η may be constructed as follows: Let $A \in \mathcal{T}(\Theta)$, then we define its measure as

$$\eta A = \mathrm{Sup}\{F(h), 0 \leqslant h \leqslant 1, h \in C(\Theta), A \supset \mathrm{support}(h)\}. \tag{10}$$

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

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

$$f(\vartheta) = \left\{ \int_I \chi_{\pi_t(A)}(\pi_t \vartheta) f(t) \exp[-\beta U(\pi_t \vartheta)] \, dt \right\} \Big/ |I| \int_X \exp[-\beta U(x)] \, \delta x, \tag{11}$$

where π_t: $\Theta \to X_t$ is the canonical projection; $\beta = 1/k_B T$ (T = temperature, $k_B \doteq$ 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 replica X_t and δx 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 μ. Therefore we have

$$\eta A = \mathrm{Sup}\{F(h), h \in D(A)\}. \tag{12}$$

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

$$\begin{aligned} \eta A &= \int_X \int_I \chi_{\pi_t(A)}(\pi_t \xi_x) \exp[-\beta U(\pi_t \xi_x)] \, dt \, \delta x \Big/ |I| \int_X \exp[-\beta U(x)] \, \delta x \\ &= \int_A \chi_{\xi(X)}(\vartheta) \, d\mu(\vartheta). \end{aligned} \tag{13}$$

This completes the proof of the theorem. □

5. 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 $\mathcal{D}$ denote a disc of dimension $M = M(N)$: $\mathcal{R}^M \supset \mathcal{D}$; consider monoparametric families of smooth maps Φ_t: $\mathcal{D} \to X$, known as families of embeddings; then the space of all such embeddings and their tangent vectors constitutes the so-called principal fiber bundle TP: $\mathrm{TP} = \{\Phi, \Phi'\}$ (Φ' = tangent vector to Φ).

At this point we define the Lagrangian $\mathcal{L}$: TP $\to \mathcal{R}$ over the principal fiber bundle induced by the measure η: Let us denote by $A_\Phi(t)$ the tube $A_\Phi(t) = \prod_{0 \leqslant t' \leqslant t} \Phi_{t'}\mathcal{D}$, and η_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

$$\mathcal{L}(\Phi_t, \Phi_t') = \int_{\mathcal{D}} L(\Phi_t(y), \Phi_t'(y)) d^M y = \lim_{\Delta \to 0} -\Delta^{-1}[\eta_{t+\Delta} A_\Phi(t+\Delta) - \eta_t(A_\Phi(t))], \tag{14}$$

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

$$\operatorname*{Min}_{\{\xi_x\}} \int_I L(\xi_x(t), \xi_x'(t))\, dt = \int_I (\xi_x^*(t), \xi_x^{*\prime}(t))\, dt, \tag{15}$$

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

$$L(x, x') = 1/2(\text{sign } u' + 1)u/c\; d/dt[\exp(u/c)], \tag{16}$$

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

$$\int_I S(x(t), x'(t))\, dt = \text{constant}, \tag{17}$$

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

$$\begin{aligned}
\int_I L(x(t), x'(t))dt &= \sum_{t_i \in \partial I^+} \sum_{p \geqslant 2} [u(t_{i+1})^p - u(t_i)^p]/(p-1)!c^p \\
&= \sum_{t_i \in \partial I^+} \sum_{p \geqslant 2} [u(t_{i+1}) - u(t_i)] \left[\sum_{k=1,2,\ldots,p} u(t_{i+1})^{p-k} u(t_i)^{k-1} \right] \Big/ (p-1)!c^p \\
&= \sum_{t_i \in \partial I^+} \sum_{p \geqslant 2} B_i \left[\sum_{k=1,2,\ldots,p} u(t_{i+1})^{p-k} u(t_i)^{k-1} \right] \Big/ (p-1)!c^p .
\end{aligned} \tag{18}$$

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

$$\int_I L(x(t), x'(t))dt = 2en\Delta + n\Delta^2 + O(\Delta^3) < \int_I L(X(t), X(t)')dt$$

$$= 2en\Delta + n^2\Delta^2 + O(\Delta^3). \tag{19}$$

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.

6. Verifying the theory

A quantitative analysis of the results requires graphic representation. For this purpose we introduce the Base-Pair Probability Matrix (BPPM) $P_{ab}(t) = \int_\Theta M_{ab}(\vartheta(t))\, d\gamma(\vartheta)$, where $\gamma(A_\Phi(t)) = \int_0^t \mathcal{L}(\Phi_{t'}, \Phi'_{t'})\, dt'$ and $M_{ab}(x)$ is equal to 1 if unit a pairs with unit b in structure x and is 0 otherwise. That is, we choose a cross section resolved up to secondary structures of the ensemble of pathways. This cross section is obtained by fixing the indexing time parameter t at a particular value. 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 $N \times N$ 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 [12,13] and use it to generate the set of kinetic barriers associated to the formation and dismantling of hairpins [4,12], the elementary events in this context. Thus, the activation energy barrier for the rate-determining step in the formation of a hairpin [4,12] is known to be $-T\Delta S(\text{loop})$, where $\Delta S(\text{loop})$ indicates the loss of conformational entropy associated to closing a loop. On the other hand, the activation energy barrier associated with the melting of a hairpin is $-\Delta H(\text{stem})$, the amount of heat released when forming all contacts in the stem. The unimolecular rate constants for helix decay and helix formation have been obtained in analytical form [4,12] 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 [4,13].

For completion we shall display the analytic expressions for the unimolecular rate constants. If the jth step or event happens to be a helix decay process, we obtain

$$k_j = fn \exp[G_h/RT], \tag{20}$$

where f is the kinetic constant for base pair formation (estimated at 10^6 s^{-1} [4,12], n is the number of base pairs in the helix formed in the jth step and G_h is the (negative) free energy contribution resulting from stacking of the base pairs in the helix. Thus, the essentially enthalpic term $-G_h = -\Delta H(\text{stem})$ should be regarded as the activation energy for helix disruption. If an admissible hairpin formation happens to be the event designated by the jth step, the inverse of the mean time for the transition will be given by

$$k_j = fn \exp(-\Delta G_{\text{loop}}/RT), \tag{21}$$

where $\Delta G_{\text{loop}} \approx -T\Delta S_{\text{loop}}$ is the change in free energy due to the closure of the loop concurrent with helix formation. Loop closure, being the nucleating step [4,12], is the rate-limiting event, therefore $-T\Delta S_{\text{loop}}$ is essentially the activation energy of helix formation.

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. Thus, the compilation of thermodynamic parameters shown succinetly in Tables 1–3 and used in the computations begets the compilation of unimolecular rate constants upon which the Markov process is constructed for a given sequence.

The time evolution of the BPPM has been monitored for the species QβMDV1-RNA, a template for the technologically-crucial enzyme Qβ-replicase [4]. 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 according to the Boltzmann measure) are displayed in Figs. 1–3, respectively. The random coil configuration is present in the ensemble although it is obviously not displayed in the BPPM. 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 twenty-one nucleotides long (see Fig. 4) are bound to each other [4]. Other pathways start at metastable structures and are nonstationary. All nonstationary pathways converge to the experimentally-deter-

Table 1
Compilation of thermodynamic parameters based on [13]. These parameters are used to generate the rate constants, in turn used to determine realizations of the Markov process. Thermodynamic basis set for stacking of nearest neighbors pairs (1M NaCl, pH 7, 37°C).

Nearest-neighbor pair	ΔG^0 (kcal/mol)
5'-A-A→ 3'←U-U-	−0.9
5'-A-U→ 3'←U-A-	−0.9
5'-U-A→ 3'←A-U-	−1.1
5'-C-A→ 3'←G-U-	−1.8
5'-C-U→ 3'←G-A-	−1.7
5'-G-A→ 3'←C-U-	−2.3
5'-G-U→ 3'←C-A-	−2.1
5'-C-G→ 3'←G-C-	−2.0
5'-G-C→ 3'←C-G-	−3.4
5'-G-G→ 3'←C-C-	−2.9

Table 2
Thermodynamic basis set for the three types of loops (1M NaCl, pH 7, 37°C).

Loop size	$-T\Delta S^0$ in kcal/mol		
	Internal	Bulge	Hairpin
1	–	+3.3	–
2	+0.8	+5.2	–
3	+1.3	+6.0	+7.4
4	+1.7	+6.7	+5.9
5	+2.1	+7.4	+4.4
6	+2.5	+8.2	+4.3
7	+2.6	+9.1	+4.1
8	+2.8	+10.0	+4.1
9	+3.1	+10.5	+4.2
10	+3.6	+11.0	+4.3
11	+4.4	+11.8	+4.9

Table 3
Thermodynamic basis set for nearest neighbors pairs involving GU base pairs (1M NaCl, pH 7, 37°C).

GU mismatch in first position

First position	Second position			
	5′-U→ / 3′← A-	5′-A→ / 3′← U-	5′-G→ / 3′← C-	5′-C→ / 3′← G-
5′-U→ / 3′← G-	−0.5	−0.7	−1.5	−1.3
5′-G→ / 3′← U-	−0.7	−0.5	−1.5	−1.9

GU mismatch in second position

First position	Second position	
	5′-U→ / 3′← G-	5′-U→ / 3′← G-
5′-U→ / 3′← A-	−0.5	−0.7
5′-A→ / 3′← U-	−0.7	−0.5
5′-G→ / 3′← C-	−1.9	−1.3
5′-C→ / 3′← G-	−1.5	−1.5

GU mismatch in first and second position

First position	Second position	
	5′-U→ / 3′← G-	5′-G→ / 3′← U-
5′-U→ / 3′← G-	−0.5	−0.6
5′-G→ / 3′← U-	−0.5	−0.5

mined active secondary structure shown in Fig. 4 [4], as direct inspection of Figs. 1–3 reveals. Such pathways are actually attracted to one minimum of the action given by eq. (18), which concentrates 48% of the measure η. This extreme starts with a random coil configuration and ends up in the ative structure shown in Fig. 4. The other extreme is the trivial stationary pathway that starts and ends in the most stable conformation. These results support the existence of an action principle guiding the exloration of conformation space.

7. Action principle and preliminaries on molecular design

Another clearcut example of an RNA species whose biologically-competent secondary structure is *metastable* is the SV-11RNA, a recombinant RNA which evolves (technically, it is the product of a re-design) from the better-known species MNV-11RNA, a natural template for the enzyme Qβ-replicase [16]. The meta-

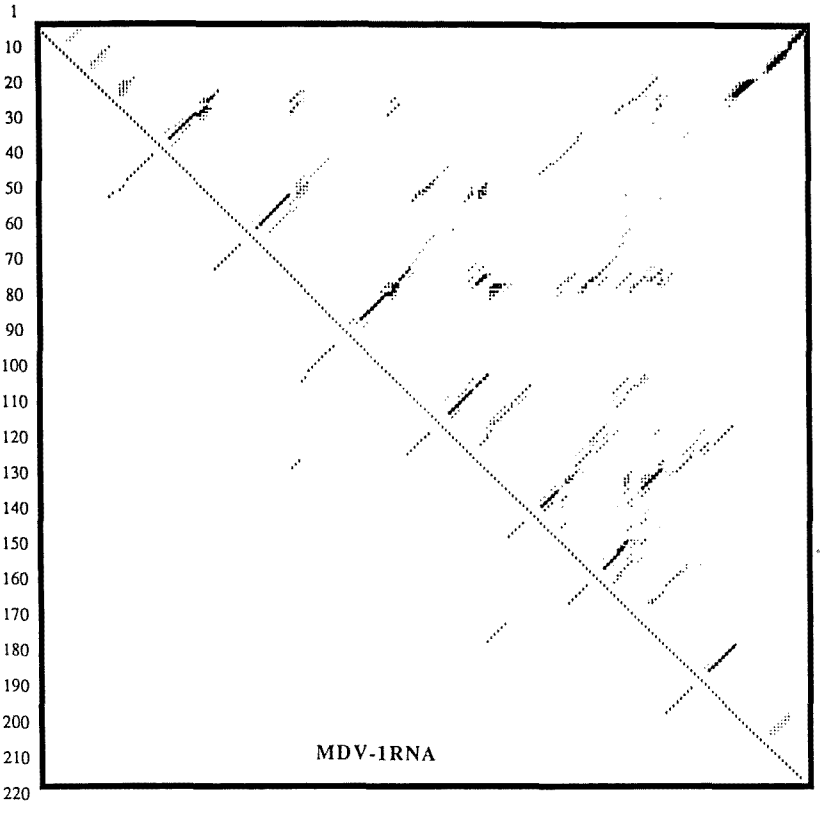

Fig. 1. The 221 × 221 BPPM $P_{ab}(t)$ for the species QβMDV-1RNA at real time $t = 0$ s. The upper right triangle of the matrices represent the ensembles generated by the Lagrangian L at different times. The active structure is represented in the lower left triangle for comparison. Notice the convergence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

stable and most stable structures of SV-11RNA are displayed in Fig. 5. In this section we deal with this species with two purposes:

(a) We first show that the active structure for SV-11 is actually the destination structure of an extreme of the action integral defined over the space of sequential folding pathways. By sequential folding pathway, we mean a pathway that results when conformation space is explored *concurrently* with the synthesis of the molecule which occurs by progressive incorporation of nucleotides.

(b) By first examining how SV-11 has evolved from MNV-11 (see [16]), we derive a criterion for the molecular design of expeditiously-folding species which evolve from a natural template in a template-directed RNA replication system.

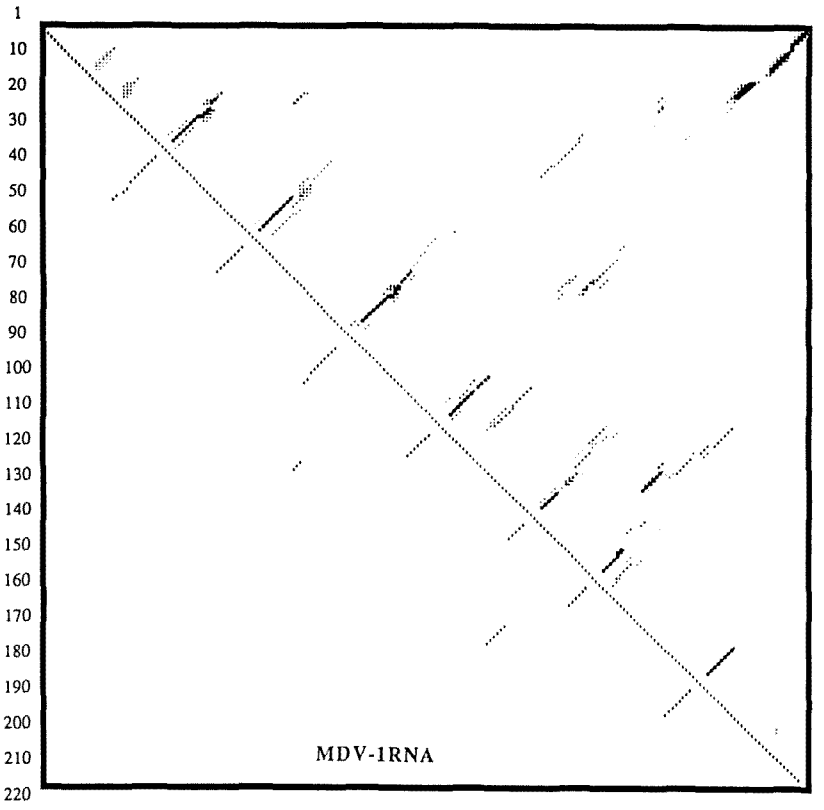

Fig. 2. Same as Fig. 1, but BPPM at real time $t = 10$ s.

Just like MDV-1 and MNV-11 RNA, SV-11 is able to serve as the template, that is, it is able to instruct its own replication, whereby Qβ-replicase assembles a complementary copy (replica) of the, say, (+) strand of the molecule by progressive incorporation of nucleotides. Thus, at each position along the (−) strand read from the so-called 5′ end (the ppp extreme in Figs. 4–6) towards the 3′ end (the OH extreme), there is a base complementary to the base that exists at the same position in the (+) strand. The latter strand, which in our argument serves as the template, is read from the 3′ extreme to the 5′ extreme.

The 115-nucleotides-long species SV-11 has been shown to have evolved *in vitro* from the 87-nucleotides-long species MNV-11 RNA, whose biologically-active secondary structure coincides with the global free energy minimum [16]. The active secondary structure of MNV-11 RNA is displayed in Fig. 6 and, within the regions of homology with SV-11, it is identical to the metastable active structure of SV-11 (Fig. 5).

Our simulations reveal that the metastable structure of SV-11 as well as the most stable structure of MNV-11 are the destination structures of pathways that constitute the extremes of the action integral defined over the space of sequential folding

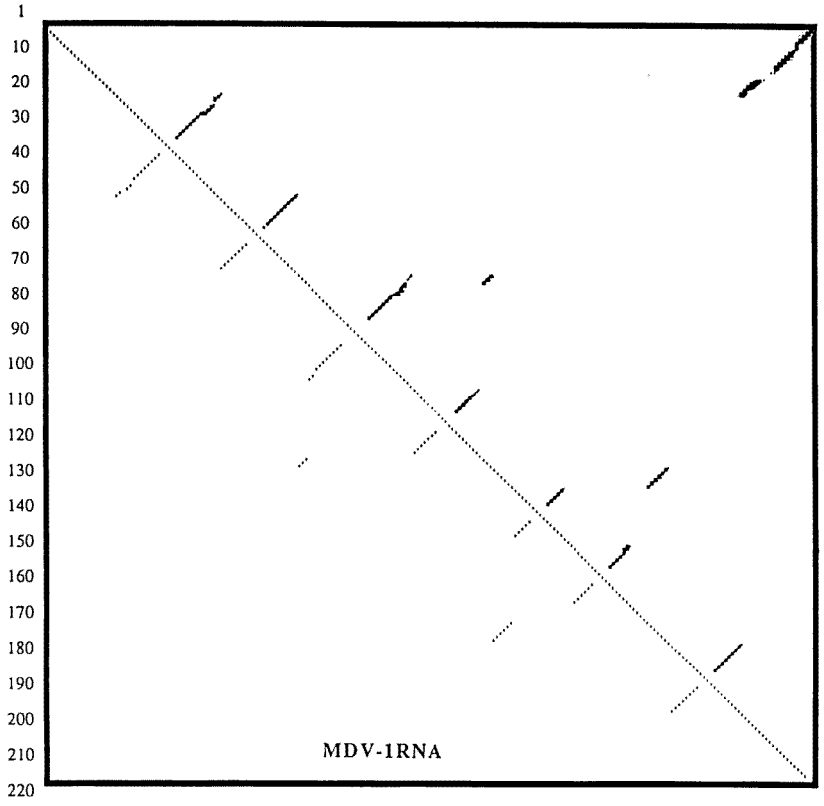

Fig. 3. Same as Fig. 1, but BPPM at real time $t = 15$ s.

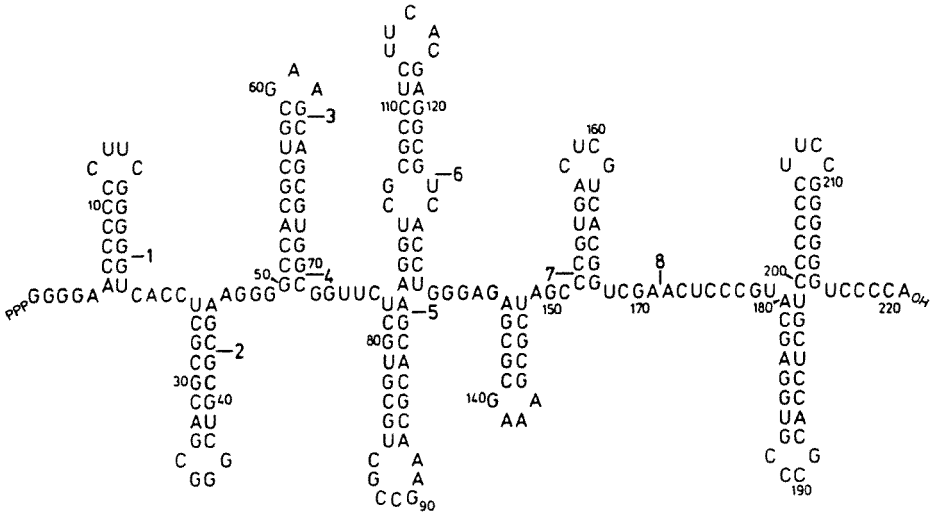

Fig. 4. The active secondary structure for QβMDV-1RNA [4,6].

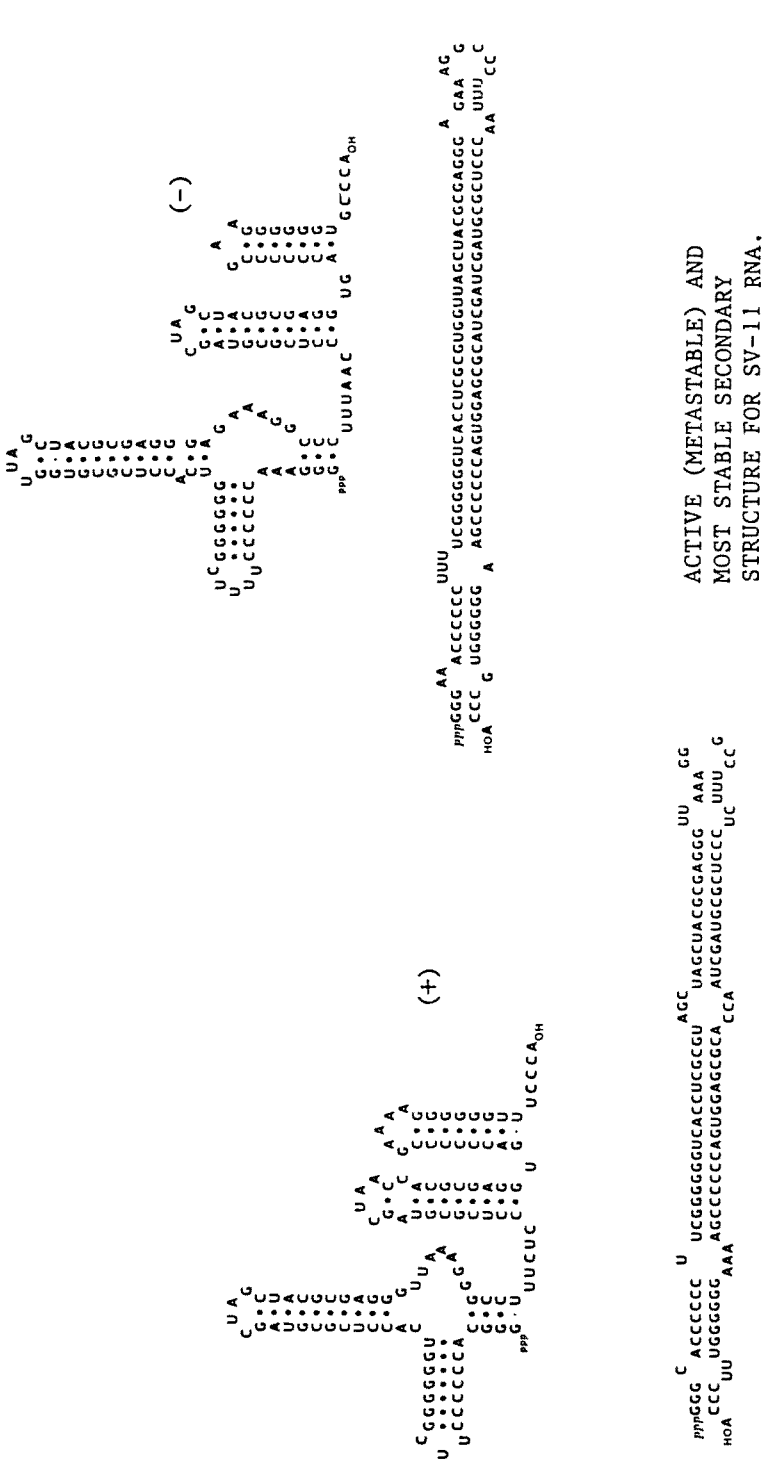

Fig. 5. Active (metastable) and most stable secondary structures for the (+) and (−) strand of SV-11RNA. The most stable structures are hairpin-like.

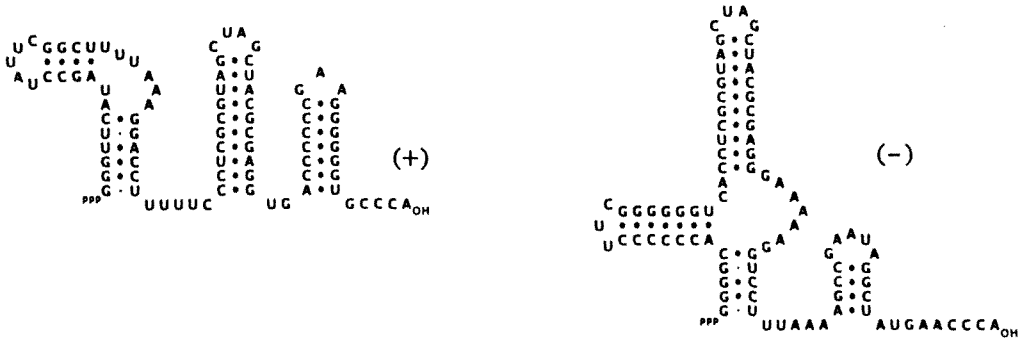

Fig. 6. The active, which is also the most stable, structure for the (+) and (−) strand of MNV-11 RNA.

pathways. In specific terms, such pathways are generated by incorporating a new type of elementary step to the Markov chain generator defined in section 6: We incorporate chain growth by addition of a single nucleotide as elementary event and fix the unimolecular rate constant for such an elementary step at 50 s^{-1}, the rate of linkage formation. This appears to be reasonable since, according to experimental estimates, Qβ-replicase synthesizes RNA at a rate of 50 nucleotides per second [4]. Thus, chain refolding will be favored over chain growth at a given stage in our simulation only if its unimolecular rate constant surpasses the value of 50 s^{-1}. Within this framework, the active structure for SV-11 or MNV-11 RNA is the only secondary structure with significant weight that results from sequential folding. In fact, no sequential pathway other than the extreme, denoted by θ, of the action integral is obtained for both molecules. Thus to any perceptible degree, the measure η must actually be $\eta(\vartheta) = \delta(\vartheta - \theta)$. A similar conclusion had been reached for MDV-1RNA [4].

Our results are supported by experimental evidence: Biebricher and Luce [16] have shown by means of kinetic (pulse-chase) experiments that the active folding of SV-11 must be metastable and must result from sequential folding: The activity of SV-11 as a template is progressively lost after the molecule emerged from the replication complex since the structure relaxes to the global free energy minimum displayed at the bottom of Fig. 5.

Concerning the potential implications of the action principle in molecular design, we first examine how SV-11 is selected and evolves from MNV-1 [16] in order to learn how recombinant products arise. The next stage is to assess the potential of each recombinant species to fold expeditiously into its active secondary structure.

The recombination process involves a copy-shifting mechanism, as direct inspection of the homologies between MNV-11 and SV-11 reveals [16]. This copy-shifting mechanism might be switched on at high ionic strength, as it requires the destabilization of the replication complex (template-replicase-replica) and reattachment of

a different template. Thus, say, the (+) strand of MNV-11 serves *initially* as template to Qβ-replicase and a budding (−) strand starts to form of what will turn out to be not a replica of (+) MNV-1RNA but the (−) strand of a new species. Provided the replication complex is dismantled at some point in the process of replica assembling, there is a possibility that the replica might bind through its copy-shift region to a complementary portion of a (−) strand of MNV-11. If replication using this new template is now carried to completion, the resulting (−) strand of the new species will be a recombinant of the (+) and (−) strand of MNV-11. This inevitably results in a palindromic sequence (see Figs. 5 and 7), which explains the great propensity to form in the long run the highly stable hairpin structure (global free energy minimum) displayed in Fig. 5.

Thus, once palindromic portions are recombined with part of the original MNV-11, the active structure of the resulting species is no longer *the most* stable, as it is the case with MNV-11 RNA (Fig. 6), but rather a *metastable* one, since a highly stable hairpin structure emerges in which the complementary portions of the sequence are bound to each other.

Since this copy-shifting mechanism by which a template is abandoned by a budding replica and a new one is adopted, could lead to various degrees of recombination, it is essential to examine the existence of extremes of the action integral for various degrees of recombination. Taking into account the template-shifting mechanism represented in Fig. 7, a parameter is adopted to indicate the chain length increase

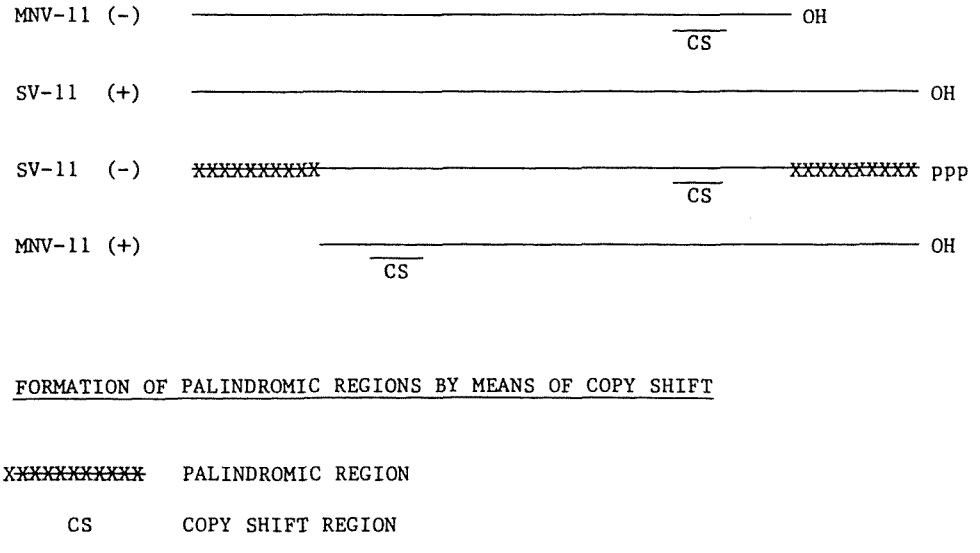

Fig. 7. Scheme of the copy-shifting mechanism, whereby a (+) strand of SV-11 would start being assembled instructed by a (−) strand of MNV-11 and would end up being assembled based on a (+) strand of MNV-11, thus shifting to a new template.

Table 4
Percentage of the statistical weight of the extreme of the action integral as a function of the length increase of the chain of the recombined product. The increase is measured with respect to the length of MNV-11 RNA.

# Recombined bases	Percentage weight
0	100
2	100
4	100
6	100
8	100
10	88
12	88
14	86
16	86
18	86
20	86
22	80
24	80
26	80
28 – – – – metastability threshold – – – – –	76
30	42
32	42
34	22
36	22
38	0
40	0

due to the recombination process beyond $N = 87$, the length of the MNV-11 RNA. The statistical weight of the bunch of sequential folding pathways that lead to the active structure is determined for each value of the recombination parameter and the results are displayed in Table 4. *One sees a single dominant bunch leading to the active structure up to a length increase of 28, precisely the difference in lengths between SV-11 and MNV-11 RNA! Thus, our computations predict that SV-11 is the maximal-length recombinant product which folds expeditiously into the active structure.* Detectable replication activity would still be possible with longer recombination products, although their populations would be easily outgrown by that of MNV-11, since the statistical weight of its active structure is 100%, versus 42–22% for the longer recombinants.

There is no extreme for the action integral as the chain increases beyond 36 nucleotides. Thus, the resulting recombinants should not be capable of instructing their own replication since they do not fold appreciably into an active structure. The biologically-competent folding that would enable such species to act as templates has been completely superseded by the inert hairpin structure.

8. Concluding remarks

Stability criteria appears to be inherent to current thinking about predictive algorithms of biologically-relevant biopolymer structure (see, for example, [1,2,14,15]). In this context, the concept of *suboptimal* folding, that is, a conformation realizing a *local* – rather than the global – free energy minimum, has been introduced as a means of accounting for the biologically-relevant conformation [14,15]. This tenet implies that, although stability control might constitute a valuable aid to structure prediction, a complementary principle must be introduced if we intend to predict biologically-active structures known to be metastable. In this regard, we pose the following question: How could we implement a useful structure-prediction algorithm which can deal with species such as SV-11 RNA [16] and QβMDV-1RNA [4], which are experimentally known to adopt biologically-active conformations whose free energy is far above the global minimum?

We have dealt with this question in this work by first recognizing that the stringent schedule under which biopolymers fold is incompatible with the long-time limit that warrants thermodynamic control [1,2,4–6,9,10]. Thus, the possibility of active conformations which are metastable arises naturally. This context leads us to replace the potential in conformation space by an action principle in the form of a path integral. This variational principle begets a nonequilibrium statistical mechanics in which we weight folding histories or pathways, incorporating time as a dimension.

Since the measure η defined on the space of folding pathways might prove difficult to visualize, we may alternatively adopt another measure $\rho = \rho(x, t)$ which weights conformations but, in contrast with the Boltzmann measure, is time-dependent. The measure ρ is related to η according to the following equation:

$$\int_I \int_X \langle h(\pi_t \xi_x, t)\rangle_x \, d\mu_B(x)\, dt/|I| = \int_{X\times I} h(x,t)\, d\rho(x,t)$$

$$= \int_I \int_\Theta h(\pi_t \vartheta, t)\, d\eta(\vartheta) dt/|I|$$

$$\text{for } \textit{any } h\colon X \times I \to \mathcal{R}; h \in C(X \times I)\,. \tag{22}$$

Thus, if we wish to compute the measure of a measurable set E contained in X at time t, we need to compute the following integral:

$$\int_X \chi_E(x)\, d\rho(x,t) = \int_\Theta \chi_E(\pi_t \vartheta)\, d\eta(\vartheta)\,, \tag{23}$$

where χ_E is the characteristic function of the measurable set E ($\chi_E(x) = 1$ if x belongs to E and $\chi_E(x) = 0$ otherwise).

Thus, in this work the Boltzmann measure μ_B over X has been effectively replaced by a time-dependent measure ρ which tends to μ_B in the thermodynamic limit $t \to \infty$.

The illustrative example worked out in this paper suggests the need for a departure from the classical picture in which a Boltzmann weight is assigned to each conformation: We believe that future algorithms for structure prediction will incorporate action principles in the form of path integrals as a means of accounting for the time dependence. Such algorithms will prove especially useful in those contexts where active conformations are known to be metastable [4,16], since they integrate the stringent schedule within which the folding process takes place. The implications of such algorithms in computer-aided design might prove an exciting new area of research whose possibilities have been assessed in a preliminary fashion in this work.

Obviously, extensions of this kinetic approach to other biopolymers, especially proteins, should be attempted. After all, experimental evidence on kinetic intermediates was first reported for proteins [2,17]. However, in order to implement an action-based approach in this context, the kinetic barriers associated to elementary events should be computationally accessible. This is not the case for proteins because the conformational entropy loss associated to loop closure is far more difficult to derive analytically than in RNA. This is due to the variable composition of polar and nonpolar groups in loops, to the diversity in size for the different residues and to the secondary structure elements (α-helices and β-sheets) which are superimposed to the pattern of native intra-chain contacts that we intend to predict. Thus, the coarse-graining of conformation space that enables to model the dynamics of the system is not readily feasible for proteins, as it is for RNA.

Besides these limitations, theoretical attempts have been made to address experimentally-probed folding pathways for proteins [18–20]. However, these attempts, based mainly on spin glass models [18,19], or lattice-casted caricatures of the peptide chain [19,20], do not incorporate a realistic landscape of kinetic barriers. At best, they suggest an ad-hoc distribution of barriers. We believe such speculation to be premature given that even the spectrum of energies has been itself modelled based on ad-hoc assumptions, such as the random energy model [18]. Thus, in so far as no realistic barrier landscape and energy spectrum has been effectively incorporated, we believe that the protein folding problem, addressed as a dynamic or variational problem has not been dealt with so far.

Acknowledgements

A.F. is a principal investigator of CONICET, the National Research Council of Argentina. Insightful discussions with Prof. Edward Nelson (Princeton) are gratefully acknowledged. The authors acknowledge support from the Air Force Office of Scientific Research. A.F.'s research has been partially supported by a fellowship from the J.S. Guggenheim Memorial Foundation.

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; A. Fernández, Phys. Rev. Lett. 64 (1990) 2328.
[5] A. Fernández, Phys. Rev. A 45 (1992) R8348.
[6] A. Fernández, Physica A201 (1993) 557.
[7] E. Nelson, Ann. Math. 69 (1959) 630.
[8] J.A. Monforte, J.D. Kahn and J.E. Hearst, Biochemistry 29 (1990) 7882.
[9] S. Partono and A. Lewin, Mol. Cell. Biol. 8 (1988) 2562.
[10] A. Fernández, J. Theor. Biol. 157 (1992) 487.
[11] A. Fernández, A. Lewin and H. Rabitz, J. Theor. Biol. 164 (1993) 121.
[12] V.V. Anshelevich, V.A. Vologodskii, A.V. Lukashin and M.D. Frank-Kamenetskii, Biopolymers 23 (1984) 39.
[13] D.H. Turner, N. Sugimoto and S.M. Freier, Ann. Rev. Biophys. Biophys. Chem. 17 (1988) 167.
[14] J.A. Jaeger, D.H. Turner and M. Zuker, Proc. Natl. Acad. Sci. USA 86 (1989) 7706.
[15] M. Zuker, Methods in Enzymology 180 (1989) 262.
[16] C.K. Biebricher and R. Luce, EMBO J. 11 (1992) 5129.
[17] O. Ptitsyn and G. Semisotnov, The mechanism of protein folding, in: *Conformations and Forces in Protein Folding*, eds. B. Nall and K. Dill (Amer. Ass. for the Advancement of Science, Washington, DC, 1991).
[18] J.D. Bryngelson and P.G. Wolynes, Proc. Natl. Acad. Sci. USA 84 (1987) 7524.
[19] E.I. Shakhnovich and A.M. Gutin, Proc. Natl. Acad. Sci. USA 90 (1993) 7195.
[20] K. Dill, Biochemistry 29 (1990) 7133.