

Searching for the underlying action principle that governs biopolymer folding

Ariel Fernández

Instituto de Matemática - INMABB and Instituto de Investigaciones Bioquímicas - INIBIBB, Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional del Sur, Bahía Blanca 8000, Argentina,

The Frick Laboratory, Princeton University, Princeton, NJ 08544, USA.

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.

1- 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 algorithms 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.

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:

$$\Pr(A) = \int_A d\eta(\vartheta) \quad (1)$$

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 residue, 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 $M(N)$:

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

A folding pathway becomes a trajectory on X defined by a map $\vartheta: I \rightarrow X$, 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 $\xi: I \rightarrow \{\text{Automorphisms on } X\}$, (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 $\mathfrak{S}(\Theta)$, and finally, endow it with a measure η induced by the stochastic process ξ which generates the trajectories.

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

$$\mathfrak{S}(X) = \{A \cap X; A \in \mathfrak{S}(\mathfrak{R}^{M(N)})\}. \quad (3)$$

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 :

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

Thus, $Y \supset \Theta$, where $\Theta = C(I \rightarrow X)$ is the space of continuous maps of the interval I on X . This space Θ is endowed with the topology $\mathfrak{S}(\Theta)$ inherited from the product topology $\prod_{t \in I} \mathfrak{S}(X_t)$ of Y . Moreover, Θ is naturally endowed with a measure μ induced by the product Boltzmann measure $\text{Pr}_B = \prod_{t \in I} \mu_{B,t}$ defined on $\wp(\prod_{t \in I} \mathfrak{S}(X_t))$, 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 $\xi: X \times I \rightarrow X$. 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]. 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 $t \rightarrow J(x, t) = \{j: 1 \leq j \leq n(x, t)\}$, 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 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: $k_j(x, t)^{-1} \leq \|\cdot\|$.

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

$$\sum_{j=0}^{j^* - 1} k_j(x, t) < r^* \leq \sum_{j=0}^{j^*} k_j(x, t), \quad (5)$$

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

then the event $j^* = j^*(x, t)$ is chosen at time t for the folding process that starts at conformation x . Thus, the map $t \rightarrow j^*(x, t)$ 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**:

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

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

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 $\mathfrak{X}(X)$, 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:

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

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:

$$F(h) = \int_X \langle h(\xi_x) \rangle_x d\mu_B(x) = \int_X \left[\sum_{\xi_x} h(\xi_x) p_X(\xi_x) \right] d\mu_B(x) \quad (8)$$

In Eq. 8, the symbol " $\langle \dots \rangle_x$ " 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:

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

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 $A \in \mathfrak{S}(\Theta)$, then we define its measure as:

$$\eta A = \text{Sup} \{F(h), 0 \leq h \leq 1, h \in C(\Theta), A \supset \text{support}(h)\} \quad (10)$$

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

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

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

Where $\pi_t : \Theta \rightarrow X_t$ is the canonical projection; $\beta = 1/k_B T$ (T =temperature, k_B =Boltzmann constant); $0 \leq f(t) \leq 1$ 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 $G(A) = \{0 \leq h \leq 1, h \in C(\Theta), A \supset \text{support}(h)\}$ with respect to the norm determined by the measure μ . Therefore we have:

$$\eta A = \text{Sup} \{F(h), h \in D(A)\} \quad (12)$$

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

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

This completes the proof of the theorem. QED.

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

$$\underline{L}(\Phi_t, \Phi'_t) = \int_D L(\Phi_t(y), \Phi'_t(y)) d^M y = \lim_{\Delta \rightarrow 0} -\Delta^{-1} [\eta_{t+\Delta} A_\Phi(t+\Delta) - \eta_t A_\Phi(t)] \quad (14)$$

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

$$\text{Min}_{\{\xi_x\}} \int_I L(\xi_x(t), \xi'_x(t)) dt = \int_I L(\xi_x^*(t), \xi_x'^*(t)) dt, \quad (15)$$

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

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

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

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

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 $u'(t) \geq 0$;

$B_i = u(t_{i+1}) - u(t_i) = i$ th barrier to be surmounted along the pathway $x(t)$. Thus, the action along a generic pathway $x(t)$ is given by

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

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

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:

$$\int_I L(x(t), x'(t)) dt = 2en\Delta + n\Delta^2 + O(\Delta^3) < \int_I L(X(t), X'(t)) dt = 2en\Delta + n^2\Delta^2 + O(\Delta^3) \quad (19)$$

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- Concluding remarks

Stability criteria appears to be inherent to current thinking about predictive algorithms of biologically-relevant biopolymer structure (see, for example [1,2,12,13]). 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 [12,13]. 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 as SV-11 RNA [14] and Q β MDV-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:

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

for any $h: X \times I \rightarrow \mathfrak{R}$; $h \in C(X \times I)$ (20)

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:

$$\int_X \chi_E(x) d\rho(x,t) = \int_{\Theta} \chi_E(\pi_t^{\vartheta}) d\eta(\vartheta), \quad (21)$$

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 $t \rightarrow \infty$.

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 [4,14].

Acknowledgements

A. F. is a principal investigator of CONICET, the National Research Council of Argentina. Insightful discussions with Profs. Edward Nelson and Eliot Lieb (Princeton) are gratefully acknowledged. The author acknowledges support from the Air Force Office of Scientific Research, USA.

References

- [1] R. Jaenicke, *Angew. Chem. Intl. Ed. Engl.* **23** (1984) 295
- [2] a) T. E. Creighton, *Bioessays* **8** (1988) 57; *Proc. Natl. Acad. Sci. USA* **85** (1988) 5082
b) 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) A. Fernández, *Eur. J. Biochem.* **182** (1989)161 ; b) A. Fernández, *Phys. Rev. Lett.* **64** (1990) 2328
- [5] A. Fernández, *Phys. Rev. A-Rapid Comm.* **45** (1992) R8348
- [6] A. Fernández, *Physica A* **201** (1993) 557
- [7] E. Nelson, *Annals of Mathematics* **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] J. A. Jaeger, D. H. Turner, and M. Zuker, *Proc. Natl. Acad. Sci. USA* **86** (1989) 7706
- [13] M. Zuker, *Methods in Enzymology* **180** (1989) 262
- [14] C. K. Biebricher and R. Luce, *EMBO J.* **11** (1992) 5129.