Journal of Statistical Physics, Vol. 92, Nos. 1/2, 1998

The Lagrangian Structure of Long-Time Torsional Dynamics Leading to RNA Folding

Ariel Fernández[1]

Received December 3, 1997; final April 7, 1998

Within the context of biopolymer renaturation *in vitro*, a principle of maximization in the economy of the folding process has been previously formulated as the principle of sequential or stepwise minimization of conformational entropy loss (SMEL). When specialized to the RNA folding context, this principle leads to a predictive folding algorithm under the assumption that an "adiabatic approximation" is valid. This approximation requires that conformational microstates be lumped up into base-pairing patterns (BPPs) which are treated as quasiequilibrium states, while folding pathways are coarsely represented as sequences of BPP transitions. In this work, we develop a semiempirical microscopic treatment aimed at validating the adiabatic approximation and its underlying SMEL principle. We start by coarse-graining the conformation torsional space $X = 3N$-torus, with $N =$ length of the chain, representing it as the lattice $(\mathbf{Z}_2)^{3N}$, where $\mathbf{Z}_2 =$ integers modulo 2. This is done so that each point in the lattice represents a complete set of local torsional isomeric states coarsely specifying the chain conformation. Then, a coarse Lagrangian governing the long-time dynamics of chain torsions is identified as the variational counterpart of the SMEL principle. To prove this statement, the Lagrangian computation of the coarse Shannon information entropy σ associated to the specific partition of X into BPPs is performed at different times and contrasted with the adiabatic computation, revealing (a) the subordination of torsional microstate dynamics to BPP transitions within time scales relevant to folding and (b) the coincidence of both plots in the range of folding time scales.

KEY WORDS: RNA folding; Lagrangian mechanics; pattern recognition; long-time motion.

[1] Instituto de Matemática, Universidad Nacional del Sur, Consejo Nacional de Investigaciones Científicas y Técnicas, Bahía Blanca 8000, Argentina.

1. INTRODUCTION

This work is concerned with the theoretical underpinnings of the expedient by which natural biopolymers reach their active conformation under *in vitro* renaturation conditions within timescales incommensurably shorter than ergodic or thermodynamic times.[1-6] In this regard, we have shown in recent work specialized for an RNA context[2] that the amount of Shannon information, $-\sigma$, measured relative to a fixed coarse description of conformation space, reaches its *absolute* maximum within experimentally-relevant timescales. This result is not generic, and applies to RNA sequences which are targets of natural selection. Moreover, this result has been obtained making use of an "adiabatic ansatz," whereby microscopic conformations are lumped up or coarsely resolved as contact or base-pairing patterns (BPP's) treated as quasiequilibrium states,[7] while the rate of each elementary BPP transition is computed from within an Arrhenius-type scenario of activated processes. This kinetic adiabatic approach treats the transition probability between two BPP's as dependent on the kinetic barrier separating the respective valleys in the free energy landscape.[2, 4, 6, 7] From a purely combinatorial viewpoint, BPP's of an RNA chain folding onto itself are drawn upon the map of Watson–Crick base-pair complementarities (A-U, G-C), which in turn, is obtained for each sequence made up of the four units denoted A, U, G and C.[7]

In this regard the following question may be posed: *What is the microscopic origin of the expediency of the folding process which is revealed at the BPP-level*? In the present work we address this question by developing a semi-empirical microscopic model of folding in which the compact conformation manifold X for the flexible RNA chain (the cartesian product of as many circles as torsional degrees of freedom the chain possesses) is coarsely resolved modulo torsional conformational isomers as the lattice $(\mathbf{Z}_2)^{3N}$, where $\mathbf{Z}_2$ represents the cyclic field of integers modulo 2, and N is the length of chain. Thus, each point in the lattice represents a complete set of torsional isomeric states ("*cis* or *trans*" in the physical chemist language) which coarsely represent the chain conformation. The implementation of the model hinges upon the identification of the Lagrangian structure of the underlying dynamics in the lattice. In this work, the validity of this variational principle will be shown to be given by the fact that it is precisely the counterpart of the principle of sequential minimization of conformational entropy loss (SMEL).[4,5] The SMEL principle reflecting the maximization in the economy of means involved in each step of folding, is known to be valid at the BPP level and has been confirmed in previous predictive computations.[4,5] Thus, the Lagrangian dynamics defined over $(\mathbf{Z}_2)^{3N}$ yields the adiabatic Arrhenius-type dynamics when projected onto the BPP space, as shown in this work.

The variational principle introduced enables us to rigorously determine the time-dependence of the coarse information entropy, σ, with respect to a fixed partition of X into BPP's, Z, and compare the results with the adiabatic computation.(2) This analysis reveals that the adiabatic approximation is valid within biologically-relevant folding timescales, and thus provides a detailed semiempirical understanding of the expediency of the folding process.

A major stumbling block in the implementation of the variational treatment of longtime torsional dynamics of the flexible RNA chain and its bearing on the folding process is due to the parallel nature of the exploration in conformation space. The concurrence of folding events taking place at the same time in different portions of the same flexible chain precludes any meaningful isolation of a single "reaction coordinate" at any given time. In order to address this question, we shall first coarsely identify foldings as elements of $(\mathbf{Z}_2)^{3N}$, that is, as patterns of $3N$ locally-encoded binary signals representing sets of $3N$ torsional states, each defined within a two-well or *cis-trans* flipping activated process. Within this frame, an underlying coarse Lagrangian or least-action formulation will be introduced over trajectories defined as sequences of pattern transitions. This Lagrangian will be conceived to single out the folding pathway which at each step minimizes the conformational entropy cost while maximizing the decrease in enthalpy. This pathway yields the SMEL-pathway when projected onto the BPP space Z. These tenets warrant a folding collapse following preferably the most economic pathway.(1) However, this is not expected to be a generic feature, but rather a specific property of sequences which are targets of natural selection, as indicated in Section 6.

The definition of the Lagrangian demands that we first justify our coarse-graining of X upon which pathways are to be drawn. Our first problem becomes how to coarsely codify the local torsional states and local correlations of the flexible chain, and provide an effective dynamical picture introducing long-range correlations to account for its long-time behavior.(8) Accordingly, to solve this problem we shall regard foldings as coarse patterns of locally-encoded structural signals modelled as generated by two-state oscillators or spin flippers. Thus, we introduce a coarse description of the dynamics based on a topological representation of the chain backbone. This is done by providing a binary codification of the soft-mode or torsional dynamics based on the local conformational restrictions that basically lead to a two-well or *cis-trans* flipping between torsional isomers subject to local and long-range correlations. Thus, each torsional potential basin represents a local topological state representing a local constraint. The geometry itself is immaterial within this level of description, since the latitude in the torsional potential basins (30 to 60 degrees,(9)) yields vast conformational distortions which would make the conformations formed unrecognizable as BPP's.

As a first step, a binary coding of local topological constraints associated to each secondary and tertiary structural motif is introduced, with each local topological constraint corresponding to a local torsional state. Our treatment enables us to adopt a relatively large computation time step of 1 ns, a value far larger than typical hydrodynamic drag time scales, without sacrificing accuracy within our level of description. Accordingly, the solvent can no longer be treated as the hydrodynamic drag medium, instead we incorporate its capacity for forming local conformation-dependent domains of different dielectric constant. Each evaluation of the matrix of local topological constraints (LTM) depends on the conformation-dependent local dielectric domains that the confined solvent will produce: As shown in Section 4, these local solvent environments determine constraints on the torsional freedom due to the orientational demands imposed on the charged phosphate groups of the RNA backbone.[4]

Folding pathways are initially resolved as transitions between patterns of locally-encoded structural signals which change within the $1/10\,\mu$s–100 ms timescale range. These coarse folding pathways are generated by a parallel search for structural patterns in the oscillating LTM. Each pattern is evaluated, translated and finally recorded as a BPP, an operation which is subject to a renormalization feedback loop. The renormalization operation periodically introduces long-range correlations on the LTM according to the latest BPP generated by translation. Nucleation and cooperative effects are accounted for by means of the renormalization operation which warrants the persistence of seeding patterns or kernels upon successive LTM evaluations.

In consonance with the first goal, our working strategy may be sketched as follows: (a) First, we introduce an ensemble of $3N$ locally-correlated two-state oscillators or spin flippers to coarsely simulate torsional isomerizations, that is a flipping between the two torsional wells for each internal degree of freedom of the RNA backbone. Then, we search for *consensus* regions of torsional isomers along the chain. By consensus we simply mean regions of the chain where the local topological constraints associated to the formation of a particular folding of the chain are satisfied. In this way, a consensus window emerges as a pattern of structural signals encoded *locally* along the sequence. The broad latitude in local torsional coordinates, or local correlation maps of the chain,[9] and the vast structural distorsions it leads to implies that the binary codification cannot be implemented at the geometric level. Rather, the spin flippers are meant to mimick changes in the local *topological* constraints to which the flexible chain is subject in order to reach specific structural patterns. (b) We generate structural patterns as consensus regions within a matrix, the LTM, of local topological constraints (LTC's) of the chain. (c) We evaluate and translate such patterns into a contact matrix representing a BPP drawn upon the Watson–Crick map of compatible (A-U, G-C) base-pairing units along the chain evolving within the timescale range 10^{-4}–10^{2} s. Thus, the translation operation is actually a projection, henceforth denoted π, and becomes a pattern recognition and therefore, a *parallel* operation. Each pattern within an LTM emerges with a certain probability which is effectively computed as the number of evaluations of the LTM that yield the particular structural motif associated to the pattern divided by the total number of evaluations of the same LTM. (d) The translation operation is subject to a feedback loop, whereby a renormalization operation ρ readjusts the oscillator periods (or spin-flipping frequencies) according to the latest BPP translated, and the contour ranges of intrachain interactions and contour distances are renormalized relative to the latest CP formed. In other words, *the renormalization operation introduces long-range correlations on the LTM* by slowing down or speeding up specific oscillators, depending on whether new interactions are formed or dismantled. (e) Nucleation steps and the cooperativity in the formation of secondary structure are accounted for by means of the renormalization operation: Suppose the LTM is evaluated at a given time and a short consensus window is detected. Then, the oscillators which generated this initial consensus window become endowed with frequencies which are lower than those of the neighboring residues and, consequently, the consensus region initially formed has a chance to grow upon successive evaluations of the LTM.

The fact that we charaterize folding steps as BPP transitions does not imply that any "adiabatic assumption" has been introduced *a-priori*, in the sense that no enslavement or subordination of the fast-evolving microscopic degrees of freedom to BPP transitions has been imposed. Equivalently, BPP's are not treated *a-priori* as quasi-equilibrium states by integrating out the relatively fast torsions as conformational entropy. Thus, the BPP is generated by a parallel search for consensus windows in the LTM. Defined in this way, an LTM represents a coarse microscopic realization of a BPP such that the consensus windows reflect the fulfilment of the LTC's determined by the BPP. This sketch of the operational tenets reveals that, although advantage is taken of the fact that there exists a wide separation between characteristic timescales associated to folding events (typically in the range 10^{-4}–10^{3} s) and chain torsions (typically in the range 10^{-11}–10^{-7} s), no "adiabatic assumption" is introduced.

In this way, a computational strategy is devised to provide theoretical underpinnings of the folding dynamics emerging as the evolution of patterns of locally-encoded signals whose coherence reflects both cooperativity and nucleation effects. The key features of our approach are: (A) The two-state coarse codification of local topological constraints of the flexible chain; (B) The renormalization of timescales for torsional isomerizations and their correlation decays relative to the successive stages of folding, thus introducing long-range correlations due to large-scale motions; (C) The identification of structural patterns with consensus regions in which specific topological constraints are fulfilled; and (D) The vast range of timescales 10^{-11}–10^{2} s covered.

At this point we may return to the original problem posed. Our representation of foldings as patterns of structural signals encoded locally along the chain allows us to identified the most economic pathway-defined as a sequence of BPP transitions. This pathway entails at each step the minimal cost in conformational entropy while forming as many contacts as possible. This fact will enable us to readily define a Lagrangian underlying the generation of trajectories regarded as sequences of LTM transitions.

An outline of the work is as follows: In Section 2, we describe the interrelationships between the two levels of description of the RNA folding process: The semiempirical microscopic description of the torsional dynamics of the chain and the resolution of the folding process as BPP transitions. In Section 3, we describe the computation of the time-dependent σ for the fixed partition of conformation space into BPP's using the adiabatic ansatz. In Section 4, we introduce our semiempirical microscopic model in detail. Section 5 is devoted to identifying the variational principle at the semiempirical microscopic level. Finally, in Section 6, we compare the Lagrangian computation of σ with that obtained using the adiabatic approximation, thus revealing the validity of the latter simplified approach within the experimental timescales involved in detectable folding events.

2. THE LEVELS OF COARSENING OF THE RNA FOLDING DYNAMICS

The relationships and interplay between the different levels of coarseness with which the folding dynamics are simulated in this work require a clear representation scheme such as the one presented in Fig. 1. Our main concern in this section is to define what is meant by a consistent treatment of the folding problem in which different coarse levels of description of the chain dynamics are compatible with each other. Standard notation has been adopted,[1–9] thus the conformation of each unit along the chain is specified by three torsional variables, each of which taking values in a circle.

Let X denote the $3N$-torus (cartesian product of $3N$ circles) corresponding to the full-detailed or torsional conformation space for a chain of

$$\begin{tikzcd}
X \arrow[r, "\Omega"] \arrow[d, "\pi'"'] & TX \arrow[d, "T\pi'"] \\
X/\sim \arrow[r, "\Omega/\sim"] \arrow[d, "\pi"'] & TX/\sim \arrow[d, "T\pi"] \\
Z \arrow[r, "\Delta"] \arrow[u, bend left=70, "\rho"] & TZ
\end{tikzcd}$$

Fig. 1. Commutative diagram reflecting the compatibility of different coarsenings of the chain dynamics relevant to the RNA folding process.

length N. Then, at the richest level of description, the chain dynamics is given by a map Ω: $X \rightarrow TX$, where TX is the tangent space of X, that is, the space of all possible vector fields over the compact manifold X. This is actually a soft-mode description of the chain dynamics in its fullest detail and incorporates the solvent solely as a hydrodynamic drag term. This analysis requires an adiabatic averaging over the hard modes (stretching vibrations and planar angular vibrations) which oscillate on timescales three to four orders of magnitude shorter than the soft modes.[9] Very limited time ranges for the full-detail soft-mode dynamics (1-100 ps) have been explored by molecular dynamics (MD) simulations.[8-10] Since the time range effectively covered by MD is far shorter than the time ranges of interest to the folding context, we shall consistently introduce coarse descriptions of X, as indicated in Fig. 1.

Let $X/\sim$ be the quotient space consisting in the set of equivalence classes of torsional isomers modulo a relation "$\sim$," that is, conformations defined modulo their local torsional isomers, regardless of their difference at a finer level of detail. In accord with known stereochemical constraints,[9] there are only two distinctive torsional species or isomers for each torsional variable. Thus, two conformations of the entire chain are regarded as $\sim$-equivalent if their torsional values for each residue lie within the same regions of the circle corresponding to the same local torsional isomers. Since there are two torsional isomers defined for each dihedral variable, $X/\sim$ may be represented as a lattice of $2 \times 3N$ points drawn on the torus X. Actually, since each point represents one of the two basins of attraction in the unit circle, with each basin representing a torsional isomer, we get the isomorphism: $X/\sim \approx (\mathbf{Z}_2)^{3N}$, where $\mathbf{Z}_2$ represents the cyclic field of integers modulo 2. This space is also isomorphic to the space of all possible LTM's since local topological constraints are specified by regions within the circle for certain combinations of the canonical dihedral variables, as indicated in Section 4. The dynamics at the LTM level are given by the projected map $\Omega/\sim : (\mathbf{Z}_2)^{3N} \rightarrow T(\mathbf{Z}_2)^{3N}$, where, as usual, "$T$" denotes tangent space in this context.

Let Z denote an even coarser description of X. In this case, two conformations in X are regarded as equivalent if their respective BPPs are identical, regardless of finer detail. Thus, at the BPP level, the folding dynamics is given by a map Δ: $Z \rightarrow TZ$. Then, all three levels of description of the folding dynamics are compatible with each other if and only if the diagram given in Fig. 1 is commutative. That means that the following relations between map compositions must hold:

$$\Delta\pi = (T\pi)(\Omega/\sim) \tag{1a}$$

$$(\Omega/\sim)\pi' = (T\pi')\Omega \tag{1b}$$

where π is the canonical projection (referred to as the translation operation in the computational context already described) which associates to each LTM its equivalence class modulo CP's, π' is the canonical projection associating its LTM to each microscopic conformation belonging to X, and $T\pi$ and $T\pi'$ are their respective projections between tangent spaces, as indicated in Fig. 1.

To summarize, the commutativity of the diagram displayed in Fig. 1 defines the consistent theoretical framework dealt with in this work: All three descriptions of conformation space, the one with the richest detail, X, the LTM description, $X/\sim$, and the BPP-description Z will represent mutually-compatible frameworks within which the folding dynamics will be studied.

3. SIMULATING FOLDING DYNAMICS AT THE BPP LEVEL

This section is devoted to determining the expediency of the RNA folding process by determining the time evolution of the information content associated to the exploration in conformation space. Conformations have been resolved as BPP's, each of which regarded as a quasi-equilibrium state according to an adiabatic ansatz. In simple terms, this means that microscopic degrees of freedom are integrated out as conformational entropy. The partition of conformation space X relevant to the computation of σ is the collection Z, a family of mutually-disjoint classes. The coarse information entropy measures the spreading of the probability distribution vector $P(t) = (P_1(t),..., P_M(t))$, where $P_j(t)$, $j = 1,..., M$ indicates the probability that a chain is folded into the BPPj at time t, and M is the total number of *a-priori* possible BPP's for a fixed RNA sequence. These probabilities should be interpreted in a Gibbsian sense, as we have a statistically large number ($\sim 10^{20}$–10^{23} per unit volume) of replicas of our system given by actual RNA molecules which are folding onto themselves as soon as renaturation conditions are established or recovered in the environment. Thus, the information entropy associated to the folding process resolved at the BPP level is

$$\sigma(t) = - \sum_{j=1,\ldots,M} P_j(t) \ln P_j(t) \tag{2}$$

A stochastic process governs the flow of probability.[4, 5] This process is determined by the activation energy barriers required to produce or dismantle interactions that stabilize the BPP's. Thus, at each instant, the partially-folded chain undergoes a series of disjoint elementary events with transition probabilities dictated by the unimolecular rates of the events. The stochastic process is Markovian since the choice of the set of disjoint events at each stage of folding is independent of the history that led to that particular stage of the process.

In order to compute the probability distribution at any given time and the resulting behavior of σ, we first discretize time t by adopting an elementary time interval length u such that $t = t'u$, where t' is dimensionless and u is the shortest possible mean BPP transition time. Then, if $U = U(u)$ represents the stochastic transition matrix at the BPP level, we get:

$$P(t) = [U(u)]^{t'} P(u), \qquad \text{with} \qquad t = t'u \tag{3}$$

where the matrix element $[U(u)]_{ij}$ is given by:

$$[U(u)]_{ij} = \left[k_{ij} \Big/ \sum_{j' \in J(i)} k_{ij'} \right] \times p_{ij}(u) \tag{4}$$

In this equation, k_{ij} indicates the unimolecular rate constant for the BPP transition $i \rightarrow j$, $J(i)$ is the set of BPP's accessible from i through elementary transition steps involving surmounting a single kinetic barrier (see below), the factor $[k_{ij}/\sum_{j' \in J(i)} k_{ij'}]$ represents the probability for the transition $i \rightarrow j$ dictated by kinetic control within a timespan of the order of $\tau_{ij} = k_{ij}{}^{-1}$, and $p_{ij}(u)$ is given by

$$p_{ij}(u) = \int_o^u Y(t - \tau_{ij})\, dt \tag{5}$$

with $Y(t-\tau_{ij})$, a Gaussian distribution centered at the mean time τ_{ij} for the $i \rightarrow j$ transition with temperature-dependent dispersion. The dispersion parameter will be evaluated in Section 4 within a realistic physical context.

Explicit values of the unimolecular rate constants require an updated compilation of the thermodynamic parameters at renaturation conditions.(7) These parameters are used to generate the set of kinetic barriers associated to the formation and dismantling of stabilizing interactions, the elementary events in our context of interest. Thus, the activation energy barrier for the rate-determining step in the formation of a stabilizing interaction is known to be $-T\,\Delta S_{\text{loop}}$, where ΔS_{loop} indicates the loss of conformational entropy associated to closing a loop. Such a loop might be of any of four admissible classes: bulge, hairpin, internal or pseudo knotted. For a fixed number L of unpaired bases in the loop, we shall assume the kinetic barrier to be the same for any of the four possible types of loops.(2, 4) This assumption is warranted since the loss in conformational entropy is due to two overlapping effects of different magnitude: The excluded volume effect, meaningful for relatively large L ($L \geqslant 100$) and the orientational effect that tends to favor the exposure of phosphate moieties towards the bulk solvent domain for better solvation. Since both effects are independent of the type of loop, we may conclude in relatively good agreement with calorimetric measurements, that the kinetic barriers are independent of the type of loop for fixed L. On the other hand, the activation energy barrier associated with dismantling a stem is $-\Delta H(\text{stem})$, the amount of heat released due to base-pairing and stacking when forming all contacts in the stem.

For completion we shall display the analytic expressions for the unimolecular rate constants k's. For clarity of the notation we shall drop the subindexing, since we shall focus each time on a specific BPP transition. If the transition happens to be a helix decay process, we obtain:

$$k = fn \exp[G_h/RT] \tag{6}$$

where n is the number of base pairs in the helix formed in the jth step, $f \approx 10^6\ \text{s}^{-1}$ is the fixed effective frequency of successful collisions(2, 4) 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. On the other hand, if the transition happens to be formation of a stabilizing interaction, the inverse of the mean time for the transition will be given by:

$$k = fn \exp[-\Delta\, G_{\text{loop}}/RT] \tag{7}$$

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.

Equations (2)–(7) should be regarded as the working equations of the adiabatic approximation. This approximation will be tested in Section 6 by comparing the adiabatic computation of the information entropy with a more rigorous computation obtained from a more profound level of description of the long-time dynamics of the chain.

4. RNA FOLDING AT A SEMIEMPIRICAL MICROSCOPIC LEVEL

The vast gap between the timescales accessible to molecular dynamics computations, typically in the range 1 ps–10 ns, and those inherent to transitions between contact patterns BPP's, typically in the range 1 μs–10^3 s, suggests the need for a semiempirical model judiciously simplifying the soft-mode or torsional dynamics. Thus, the problem becomes how to incorporate effective internal degrees of freedom of the chain whose dynamics translates or projects onto sequences of BPP transitions. The aim of this section is to introduce a matrix where local torsional states of the chain are codified in a simplified binary fashion, so that patterns of locally-encoded structural signals may be recognized and translated as BPP's.[8]

Since conflicting possibilities may arise yielding different evaluations or pattern recognitions, a probabilistic approach appears to be necessary. A realization of this concept is introduced in this work and materializes in a semiempirical model which deals mechanistically with the rich dynamic hierarchy of timescales determined by the different levels of structural resolution of the folding process and their interplay.

Thus, the exploration of conformation space results from the parallel occurrence of trails of folding events. The base-pairing matrix (BPM) evolving in the 1/10 ms–10^3 s timescale and built upon the map of Watson–Crick (W–C) antiparallel complementarities is generated by a search for consensus windows in the LTM which records the phases of $3N$ two-state oscillators or spin flippers evolving within a period range 10^{-11} s–10^{-5} s. Each oscillator represents the flipping between the two potential basins (or "*cis-trans*" isomers) for each dihedral torsional degree of freedom. The LTM represents a coarse microscopic realization of the BPP represented by the BPM such that a consensus window reflects the fulfilment of the local topological constraints determined by the putative intrachain contact: Specific dihedral torsions must be in the "correct state" required for contact formation. As emphasized in Section 1, the local geometry itself is immaterial since the latitude of the torsional potential basins (30 to 60 degrees,[9]) yields vast conformational distortions which would make the conformations formed unrecognizable if translated as BPPs.

The local constraints are themselves imposed by the conformation-dependent confined-versus-bulk solvent environments of different dielectric, and by the steric restrictions determined by loop closure. In this way, a coarse long-time torsional dynamics of the chain becomes computationally accessible. The codification will be taken to be binary since each torsional degree of freedom flips between two potential wells representing two local torsional isomers subject to local correlations and constraints. As shown in this section, the computational time step at the typical folding temperature of 303 °K is 1 ns, a value far larger than the hydrodynamic timescale of 15 ps used in the continuum soft mode analysis.[10] Accordingly, the solvent can no longer be treated as a hydrodynamic drag medium: Its capacity to form local conformation-dependent dielectric domains must be incorporated.

To illustrate how our theory works, suppose a putative contact involves W-C complementary regions of the chain which flank the consensus window and requires the closure of intra-chain loops which define different dielectric environments determined by the confined cluster-like versus bulk solvent. The formation of such environments imposes constraints on the dihedral torsional states of the units forming the loop which must be fulfilled if the contact is to be formed and as such, registered in the BPM: Not only the backbone two-state "vertebrae" (see Fig. 2) must be correctly positioned for the loop to form, but the charged phosphate groups of the RNA backbone should face the best dielectric environment available for better solvation.[4] Now let us cast this situation within our the computational context: If all spin flippers or two-state oscillators are in the "correct" state specified above at the time of the reading, a contact is recorded

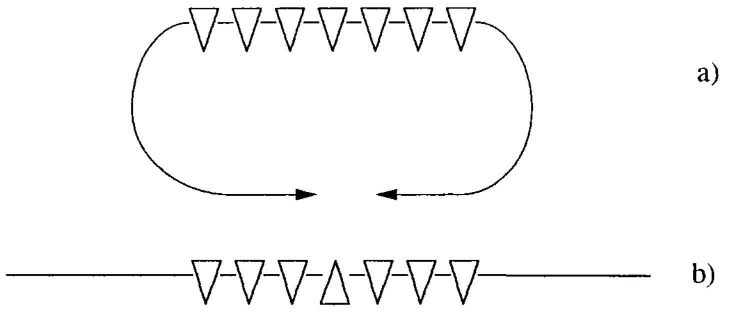

Fig. 2. Schematic representation of vertebral or row (1) consensus: When the vertebrae are correctly positioned, as in a), loop closure becomes readily feasible, while it is precluded except for large loops ($N(\text{loop}) \gg 17$) if vertebral consensus is not reached, as in situation b).

in the BPM. In this way the coarse soft-mode dynamics of the chain evolves as a sequence of pattern recognitions.

These dynamical aspects are incorporated in our computations since they determine the way in which the LTM is translated into the BPM by means of a pattern recognition. In turn, this parallel operation must be re-prescribed—technically renormalized—after each BPP transition, since consensus evaluation depends on the last BPP generated: Once foldings have formed, the distance between any two specific units is no longer the contour distance along the chain, and the loops which must be closed to form a contact are of different length relative to a those formed upon the random coil. These facts imply that the dielectric constraints are different and thus, the patterns in the LTM are recognized differently with each folding step. This sketch reveals that long-range correlations are introduced in the LTM by means of the renormalization operation.

To summarize, the approach put forth in this section may be best described as a coarse-grained analysis of soft-mode (torsional angular motion) dynamics defined by the recognition of evolving patterns of local dihedral torsional constraints consistent with a simplified topological model for the RNA backbone. Long-range intramolecular interactions are then induced by the fulfillment of local constraints to which the chain dynamics is subject due to the capacity of the solvent to determine domains of different dielectric. The essential premise in this analysis is that the local dihedral torsions within an intrachain loop must be constrained to remain in one potential basin ("*cis or trans*") if the concurrent contact is to be formed, so that physically, charged phosphates within the loop face the highest dielectric environment. Thus, the fact that the solvent defines conformation-dependent dielectric environments is computationally assimilated to the fact that a specified pattern of fulfilled topological constraints at the time of a reading of the LTM gets translated into a BPP.

This sketch of the operational tenets reveals that, although advantage is taken of the fact that there exists a wide separation between characteristic timescales associated to folding events (typically in the range 1 μs–10^3 s) and internal backbone motions, essentially realized as dihedral torsions (typically in the range 10^{-11}–10^{-5} s),[9-11] no "adiabatic assumption" subordinating or enslaving microscopic degrees of freedom to BPP transitions is introduced.

The basic representational and operational tenets of the computational design for the semiempirical microscopic model sketched above are:

(a) An LTM matrix simultaneously recording the state of each "vertebral" and phosphate orientation dihedral. Each effective torsional state is generated by a two-state oscillator whose period is chosen from Gaussian temperature-dependent distributions which are different depending on whether the nucleotide is free or engaged in an intrachain contact and its concurrent loops. A new period is chosen for an individual oscillator after completion of the previous period, thus incorporating thermal fluctuations. The broad latitude (up to 30° to 60°,[9]) in local torsional coordinates within local correlation maps of the RNA chain, and the vast structural distorsion it leads to, implies that the binary codification cannot be implemented at the geometric level. Rather, the spin flippers or oscillators are meant to mimick changes in the local topological constraints to which the flexible like chain is subject in order to reach specific structural patterns.

(b) A Watson-Crick map (WCM) of antiparallel windows upon which the BPM's are built by translation of the information encoded in the LTM.

(c) A built-in internal clock (C), incorporated in order to synchronize the LTM generation timing and its reading and subsequent evaluation at regular intervals. In accord with Shannon's information theory, the fixed beat period of C must be at most half of the shortest dihedral period.

(d) The folder (F) which evaluates the LTM and identifies consensus windows. This operation is matched with the WCM to generate BPM's and ponders the interrelationship between "vertebral" and phosphate-orientation consensus. In addition, F may dismantle or relax contact regions in the BPM whenever a significant consensus bubble arises within a previously-formed consensus window.

(e) The operation of F is feedbacked within a renormalization loop into the LTM generator, and the new evaluation of consensus windows is renormalized or prescribed according to the last BPM generated.

Taking into account that folding materializes in a statistical ensemble of RNA molecules and that conflicting consensus evaluations demand a probabilistic approach, we shall adopt an appropriate output representation: The statistical dynamics of consensus search is defined by the time evolution of a base-pairing probability matrix (BPPM), representing the weighted overlap of different consensus evaluations.

The essential operation in our parallel algorithm, the folding operation, lies within a renormalization feedback loop and consists in the translation of information casted in terms of the state of internal microscopic degrees of freedom into a coarser representation, a BPP defined by a BPM. The latter matrix is in turn built upon the WCM. Thus, the algorithm and its underlying semiempirical model identify the BPP class to which an LTM belongs at certain time intervals according to a prescribed set of rules. The generation of the LTM is in turn affected by the last BPP transition that has taken place, since a new set of constraints arises with each new BPP formed, and the prescription for the evaluation operation itself is renormalized and thus depends on the last BPP which has occurred. Prior to defining the folding process explicitly within our model, we must specify the basic representational elements and their interrelationships with regards to the basic operations:

4a. The LTM

This $3 \times N$ matrix is a coarse representation of a microscopic realization of a BPP. Each entry adopts the values 1 or 0, representing two significant states of an RNA backbone torsion localized in a specific nucleotide (*nt*) or unit. The entry values are generated by two-state oscillators, one for each entry, whose period τ is automatically adjusted after one whole period has been completed from a fixed T-dependent distribution $w(\tau)$ according to renormalization specifications detailed below. Each column in the LTM represents a different *nt*, with the ith column ($1 \leqslant i \leqslant N$) corresponding to the *nt* with contour value i along the chain. Each of the three rows represents a different reading space and their interrelationships are pondered each time the LTM is translated into the BPM, according to the size of the consensus window.

Each entry in the first row, denoted (1), indicates the dihedral spin state for a backbone "vertebral" torsion, as schematized in Fig. 2. A consensus window of consecutive spins in, say, state 1 in row (1) is a necessary condition for closure of a loop comprised of the sequence of *nt*'s within the associated contour window. The physical interpretation of this vertebral consensus as a necessary constraint for loop closure is schematized in Fig. 2. Notice that vertebral consensus is not directly related to any geometric curvature condition for loop closure, which would make it impossible to have consensus subwindows flanked by W-C complementary regions (it would be impossible to satisfy at the same time and in the same region of the chain geometric constraints for loop formation involving the whole region and those for any smaller loop involving a subregion). On the other hand, the possibility of different conflictive consensus evaluations is perfectly compatible with the probabilistic nature of our model. Depending on the size of the window, the existence of vertebral consensus at the time of a reading with flanking regions within the WCM may lead to the BPM-recording of an intra-chain long-range contact formation between the *nt*'s flanking the consensus window. If N(loop) indicates the size of the consensus window, we may state that the necessary condition becomes sufficient if and only if $N(\text{loop}) > N_c = 17$ (cf. ref. [12]), as shown below.

The entries in the second and third rows, denoted (2) and (3), of the LTM indicate the dihedral spin states of those backbone torsions engaged in orienting the negatively-charged phosphate group. Thus, for the sake of convention, spin state 1 in rows (2) and (3) for column i indicates that the phosphate of the ith *nt* faces bulk solvent whenever this *nt* is part of an intramolecular loop.

4b. The Interrelationship Among the Rows of the LTM

To specify the interrelationships between these different reading spaces in consensus evaluation, we first define an intramolecular (i, j) contact with $i < j$ as the W-C base-pairing engaging the two *nt*'s with contour values i and j. The occurrence of the (i, j) contact is marked by a 1 in the ij-entry of the triangular BPM. Suppose the initial BPP corresponds to the random coil, that is, there are no intramolecular contacts, and $N(\text{loop}) = |j-(i+1)| > 17$ (cf. ref. [12]). Then, once a reading of the LTM takes place, an (i, j) contact will be produced after evaluation and recorded as such in the new BPM if all dihedral "vertebral" spin states for the segment of row (1) flanked by entries i and j are in the correct torsional conformation for folding, that is, they are in state 1 (cf. Fig. 2). The fact that the range of the putative (i, j) interaction must be larger than 17 to materialize with vertebral consensus *only* can be justified as follows: Loop closure defines two solvent domains, an innerlow-dielectric domain confined by the loop of rod-like dimensions and an outer high-dielectric or bulk-like domain. If the loop is sufficiently large, encompassing more than 17 *nt*'s, as demonstrated in ref. [12], the dielectric difference between the inner and outer domain becomes negligible: Each charged phosphate group pointing to the inside of the loop admits four water-solvation layers for a loop of size 17 or larger. However, for smaller loops the consensus demands are higher and vertebral consensus is no longer sufficient: The phosphate groups must be oriented towards the bulk for better solvation. This argument leads us to qualitatively distinguish rows (2) and (3) from row (1): As more orientational constraints are associated to loop closure, more consensus is needed for it to materialize, so that for a putative loop smaller or equal to 17, that is $|j-(i+1)| \leqslant 17$, consensus windows in rows (1), (2) and (3), flanked by columns i and j becomes the necessary and sufficient condition.

4c. The Dihedral Frequency Distributor (DFD)

In attempting to project dihedral torsions into the BPP space, where the folding process is conventionally recorded, the timescale limitations of molecular dynamics simulations must be circumvented. This explains the need to introduce models, such as the one presented in this work, in which activated molecular motions in the ns to μs range are considered.[9–11] Thus, faster diffusional-like unhindered torsions, such as the torsion around the glycosidic sugar-base bond in an unpaired *nt* are integrated out as conformational entropy of the state defined by the LTM representation. Such motions, well into the ps range, determine the rod-like shape of the RNA molecule when viewed within the timescale window between two LTM states. Precisely this interrelation between shape and timescale justifies the concept of inner and outer solvent domain defined by an intramolecular loop, as put forth in our consensus analysis of the entropic cost of loop closure. Thus, the microscopic mean time range relevant to LTM transitions is 1 ns–1 μs, covering the timescale for internal motions (1 ns–10 ns), of the order of the calculated diffusional displacements of flexible hinged domains,[10] and, at the other end of the spectrum, the limiting value (1/10 to 1 μs) for a localized helix-unwinding event leading to a bubble within a helix.[11]

These considerations lead us to define a temperature-dependent normalized distribution of periods, $w = w(\tau)$. In particular, the periods of unhindered dihedral oscillators are assigned from this distribution in such a way that the effect of thermal fluctuations on the formation of consensus and thus, on structural transitions is incorporated. The distribution has three Gaussian peaks, each with dispersion $\sigma^2 = gT$, where the constant g depends on the actual denaturation temperature T(denat.) and on the consensus interpretation of denaturation, as shown below. The peaks occur, respectively, at mean periods 10 ps, 10 ns and 1 μs. This distribution allows us to classify *nt*'s in two classes: To class I belong all *nt*'s with mean dihedral period 10 ps, while class II contains all *nt*'s whose mean dihedral period is either 10 ns or 1 μs. The first class corresponds to internal dihedral torsions of the RNA chain of the type probed by fluorescence depolarization.[10] These torsions occur in free *nt*'s, that is, unpaired *nt*'s not belonging to a loop. Accordingly, the DFD in the folding machine establishes a lottery from which periods of dihedral spin oscillators for free *nt*'s are assigned from within the period range centered at 1 ns. A new period is assigned from the lottery to each oscillator each time a whole previously-assigned period has been completed. The frequency $f = 1/\tau$ of a dihedral spin in any row of the DSSM corresponding to a *nt* not engaged in an intra-chain interaction or loop satisfies the inequality

$$|f^{-1} - 1 \text{ ns}| \leqslant |\tau' - 1 \text{ ns}|$$

with τ' satisfying:

$$w(\tau') = \text{Infimum}_{\tau}\{w(\tau) \geqslant 1/[3N]\} \qquad (8)$$

The condition yielding the extreme period τ' arises from the fact that there are at most $3N$ free oscillators in the chain. At the typical folding temperature $T = 303$ °K, we get $\tau' \approx 2$ ps.

The other two peaks in the distribution correspond, respectively, to mean periods for *nt*'s engaged in an A-U or G-C Watson–Crick base-pair within a helix, or to *nt*'s within loops. In the latter case the period range covers the entire bimodal distribution. Again, the same considerations apply in regards to the period assignment to oscillators for class II *nt*'s. These rules imply that the *nt*'s in loops concurrently formed with an intrachain helix adopt the same cadence as the helix *nt*'s themselves. This is so, since the rate of helix dismantling is exclusively dependent on the size of the helix taken by itself,(4, 13) and determined by the formation of a significant consensus bubble amongst the class II *nt*'s engaged in the helix. Furthermore, this local limiting event is fairly independent of concurrent microscopic events in the associated loops.

The mechanistic aspects of period distribution, as performed by the DFD imply that this operation is subject to renormalization with each BPP transition: A BPP determines which columns in the LTM correspond to free or class I *nt*'s in the chain and which correspond to *nt*'s engaged in an intrachain interaction, or, equivalently, they belong to class II. Thus, since a BPP transition reclassifies the *nt*'s, it also dictates the range from which new period assignments are drawn. The period range for a specific *nt* remains the same as before the BPP transition if the transition does not alter the class of the *nt*, and changes if the BPP transition transfers the *nt* to a different class.

4d. The Built-In Internal Clock (C)

In order to satisfy the basic tenets of Shannon's information transmission while making all operations C-synchronized, the intrinsic beat period of C should be taken to represent $1/2\tau'$, one half of the shortest possible period to be assigned to a dihedral spin oscillator. Since τ' depends on $w(\tau)$, it is itself T-dependent. The time interval between two consecutive readings of the LTM, $\tau(\text{read})$, is taken as constant and, in order to lower the operational cost, it is fixed at the shortest time that a BPP transition could possibly take. Thus, $\tau(\text{read}) = 2^{3\times 3}\tau' \approx 1$ ns, the shortest time to form the 3-loop, the smallest possible loop engaging the fastest oscillators.

4e. The Folder (F)

The actual translation of the consensus evaluation of the LTM into the BPM is performed by the folder (F), with the aid of an updated version of the reading instructions manual (RIM). An updating takes place with each BPP transition marked by a change in the BPM. The folder might form or dismantle (relax) several intrachain helices in parallel and records in the BPM the consensus evaluation performed with the aid of the RIM, as indicated in the schemes displayed in Fig. 3. The consensus search or evaluation under unconstrained conditions, that is, with a RIM defined by the random coil BPP, has been partially delineated in a) for helix formation. The flow chart for this parallel operation is displayed in Fig. 4.

On the other hand, helix dismantling materializes and is recorded as such by deletion in the BPM whenever a consensus bubble forms amongst class II *nt*'s engaged in basepairing. By "consensus bubble" we mean that in any of the three rows, a consecutive sequence within the set of dihedral

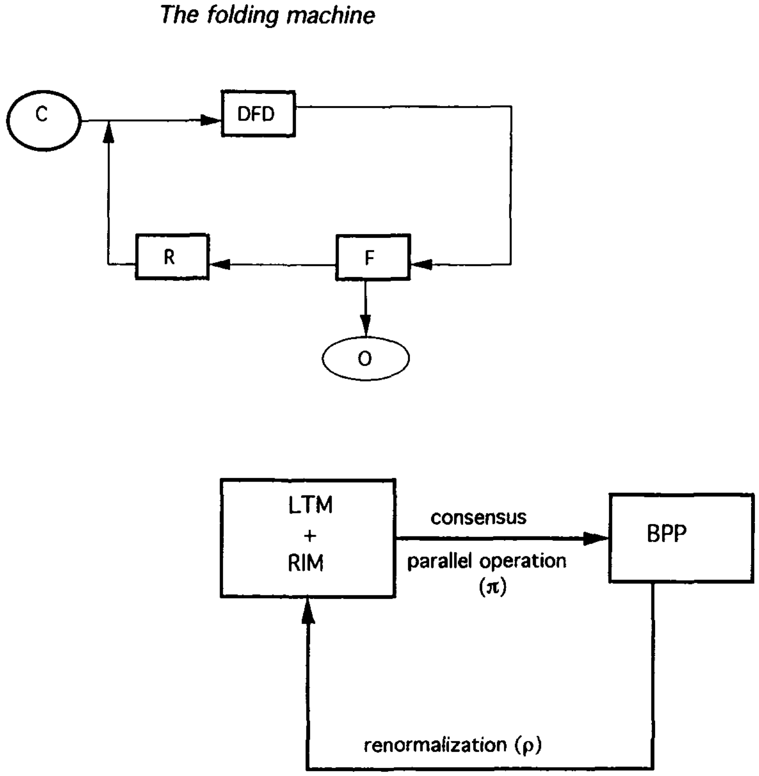

Fig. 3. A) General scheme of the folding machine featuring its basic components: The inherent clock (C), the dihedral frequency distributor (DFD), the folder (F), the output displayer (O) and the renormalizer (R). B) General scheme of the interrelation between the different representational elements: The Local topological constraints matrix (LTM), the reading instructions manual (RIM) and the base pairing matrix (BPM).

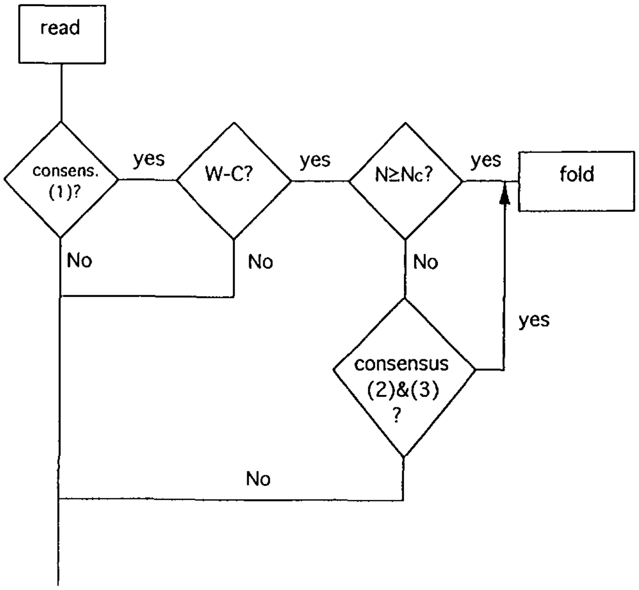

Fig. 4. The flow chart of the folding operation as a consensus evaluation after reading the three rows (1), (2) and (3) of the LTM. The following notation has been adopted: "consens. (1)?" and "consens. (2)&(3)?" refer respectively to finding a consensus region on row (1) or finding it on rows (2) and (3) on the LTM; "W-C?" refers to finding Watson–Crick complementarity in the regions flanking the consensus region and "$N \geqslant Nc$?" refers to deciding whether the size N of the consensus region is larger than the critical size Nc.

spins of helix *nt*'s of length 30% of the total helix length must be out of phase with the consensus value 1. In other words, the sequence within the helical region must adopt the state 0 at the time when the reading of the LTM takes place. Because of the renormalization loop, this transition at the BPM level immediately transfers a new set of constraints for the generation of new LTM's: The *nt*'s previously engaged in the helix and in the concurrent loops are reclassified, being transferred from class II to the higher frequency class I.

Stacking effects[14] reflect themselves mechanistically in the formation of the consensus bubble: The larger the helix, the more improbable to find a 30% out-of-phase subsequence of oscillators from class II. Furthermore, these considerations enable us to estimate the constant g which determines the effect of thermal fluctuations on the period distribution: At the denaturation temperature T(denat.), every helix formed must develop a consensus bubble evaluated and recorded with the next reading of the LTM. Thus, if σ is "large enough," the period distribution in the helix is broad enough so that consensus cannot be preserved: The period range, of the order of σ, is such that a helix consensus cannot survive two consecutive readings. From these considerations, and taking into account our empirical estimate of the denaturation dispersion fixed at $\sigma = 0.6\,\mu$s, and the typical experimental T(denat.) = 312 °K for ribozymes such as the ones studied in this work,[14] we get $g \approx 1.2 \times 10^{-15}$ s^2/°K.

4f. The Renormalizer (R)

The renormalizer has two simultaneous roles: It updates the RIM by determining contour distances *relative* to the latest BPP generated, and, by readjusting frequencies, it places a new set of constraints on the generation of new LTM's based on the latest BPP translated. Thus, *renormalization accounts for long-range correlations that develop on the LTM as a consequence of folding events recorded as BPP transitions.*

4g. Actual Computational Implementation

The computation performed by $(\pi - \rho)$ loop iteration constitutes a parallel algorithm designed to predict the active structure of biomolecules reached within times incommensurably shorter than those required for thermodynamic equilibration. Such computations require a parallel evaluation of folding possibilities at regular intervals which, in turn, define the set of constraints to which the system is subject when undertaking the next folding stage. A sequential machine with adequate memory capacity ($\sim$10 GB) is suitable to perform such computations.

The dynamics can be generated sequentially provided the LTM is quenched throughout each $(\pi - \rho)$ iteration. Within a single $(\pi - \rho)$ loop, the sequential computational bottleneck is the π operation since it is inherently a parallel operation consisting of a pattern recognition in the LTM. If N is not too long ($N \approx 100$–300) this operation may be accessible to a sequential machine engaged in column by column reading with concurrent memory storage. A state-of-the-art microVAX cluster takes approximately 1.08 ms of real time to sequentially read each LTM for $N = 220$. On the other hand, the renormalization operation which redistributes frequencies according to the pattern generated by π using a Monte Carlo lottery routine, is inherently sequential and takes about 0.3 ms. The overall computation time involved in $10^7(\pi - \rho)$ iterations is of the order of 10^4 s for the system size and machine specifications mentioned.

In practice, each dihedral oscillator or flipper may be modelled as a two-state spin coupled to an external rapidly oscillating magnetic field of invariant frequency conveniently fixed at 10^{12} Hz. Each coupling constant defines the flipping response to the oscillations of the external field. Thus, different coupling constants entail different degrees of entrainment by the external field. Then, a distribution of coupling constants begets a distribution of flipping frequencies. Good advantage may be taken of spin glass simulation technology to actually generate the LTM's in this way. This is the actual computational ansatz which has been used in this work to generate LTM's after a sequential renormalization taking place during the quenching concurrent with each $(\pi - \rho)$ iteration.

5. A LAGRANGIAN DEFINED AT THE SEMIEMPIRICAL MICROSCOPIC LEVEL

Foldings of the RNA chain have been identified as patterns of locally-encoded signals whose generation, translation and renormalization have been studied in the previous section. Since structural patterns have been identified with consensus windows within which local topological constraints are fulfilled, it follows that the preferred folding pathway resolved as a sequence of BPP-transitions is the one for which each transition entails the minimal cost in conformational entropy, while maximizing the enthalpy decrease. This BPP-pathway is the easiest to form, and therefore the most probable as it entails the maximum economy of means for each step. This is so since consensus as identified in a LTM, implies the fulfillment of local topological constraints which, in turn, determine the BPP and, on the other hand, the enthalpic content is exclusively dependent on the contacts which have materialized. Thus, our semiempirical approach is perfectly compatible with the SMEL principle.[4, 13]

In view of these facts, the following problem arises: Is it possible to define a Lagrangian L at the coarse microscopic level of LTM-transitions which reproduces the SMEL behavior when an LTM-trajectory is translated as a sequence of BPP-transitions? According to our treatment, the solution to this central problem involves determining $L = K - H$, where H is the enthalpy content, or potential in our context, and K is the "kinetic energy" of an LTM, $K = K(\text{LTM})$, which must satisfy the following conditions:

(A) It must be uniquely defined for all LTM's associated to a specific BPP, since otherwise many LTM-trajectories, amongst which is the one singled out by L, would translate into the SMEL pathway.

(B) A BPP-transition is favored if it entails a minimal entropic cost, therefore its corresponding LTM-transition must entail a minimization of the kinetic energy loss.

In accord with conditions (A) and (B), we define

$$K(\mathrm{LTM}) = 1/3 \sum_{i=1,\ldots,3N} f_i^2 \tag{9}$$

where the f_i's, given in Hz, are the mean frequencies of the oscillators of the LTM and hence H must be given in Hz^2 units.

The choice given in Eq. (9) is a good one, as we shall demonstrate making use of the fact that the oscillator frequencies of the LTM are determined by the renormalization operation: Let us suppose that a pattern of local signals has been identified and recorded in the BPM at a given time by the translation operation. Then, the renormalization operation reclassifies the residues engaged in the latest BPP generated, from class I into class II. The reader is reminded that nts. belonging to loops whose closure is concurrent with interactions adopt the same cadence of those nts. that form the respective contacts. Thus, the loss in conformational entropy associated to the CP transition is corresponded with a loss in kinetic energy of the LTM in such a way that a minimization of the entropy loss is corresponded with a minimization of the loss in kinetic energy. This implies that the trajectory which maximizes the L-action is the one that, projected onto BPP-space, represents the SMEL behavior.

On the other hand, the renormalization operation warrants the uniqueness of K for every LTM which translates into the same BPP, since the renormalization operation establishes the new frequencies exclusively on the basis of the latest BPP formed. By virtue of the renormalization operation, the occurrence of intrachain contacts is reflected as a lowering of oscillator mean frequencies. This is the reason why the Lagrangian L underlies the coarse microscopic behavior which shows up at the level of BPP transitions as the SMEL principle.

In regards to the T-dependence of K, we must take into account that at the denaturation temperature $T = T(\text{denat.})$, consensus bubbles are created which dismantle every BPP, as indicated in Section 4. This is the result of a broad spreading of frequencies beyond the critical spreading value and the concurrent impossibility of preserving phase in consensus windows, as indicated in Section 4. Thus, the renormalization operation near criticality will speed up all the oscillators for consensus windows where critical size bubbles occur, while such bubbles will become more and more ubiquitous as T approaches $T(\text{denat.})$. Then, K increases with T and the following relation holds as T approaches $T(\text{denat.})$ from below:

$$\operatorname*{limit}_{T \to T(\text{denat.})^-} K = \text{Maximum } K = N \times 10^{22}\ \mathrm{Hz}^2 \tag{10}$$

To conclude we may state that the most economic folding pathway resolved at the coarse microscopic level, y^*, satisfies the following classical dynamics result:

$$L\text{-action}(y^*) = \underset{\text{all LTM-trajectories } y}{\text{Maximum}} \; L\text{-action}(y) \qquad (11)$$

Where "L-action(y)" denotes the sum of all $L(\text{LTM}(i))$'s for all the LTM(i)'s, with i = discretized time index, along the trajectory y defined by mapping the previously-alloted time interval over $(\mathbf{Z}_2)^{3N}$.

Without loss of generality we may assume that there exists a positive constant q so that $q + L = L' > 0$. We may define a stochastic process Ξ at the semiempirical microscopic level of DSSM transitions by defining transitional probabilities with respect to L':

$$p(V \rightarrow W) = [L'(W) - L'(V)] \Big/ \left[\sum_{W'} L'(W') - L'(V) \right] \qquad (12)$$

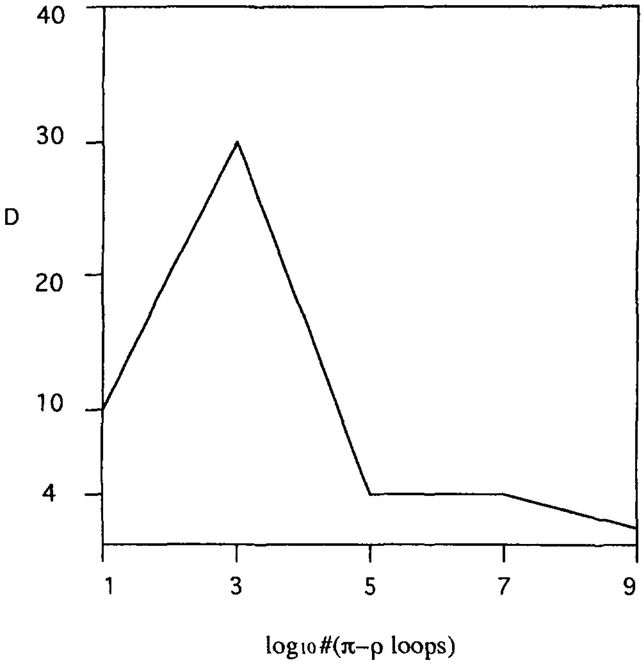

Fig. 5. Plot of the Hamming distance (× 100) between consecutively-evaluated LTM's along the folding pathway generated by iterating $(\pi - \rho)$ loops for the Tt LSU RNA.

where V, W, W' are DSSM's, and the sum is extended over all W' states accessible from V within a pre-determined Hamming distance $d = d(\text{max.})$ between LTM matrices (see Fig. 5). This distance is defined as:

$$\begin{aligned} d(V, W) &= (1/3N) \sum_{i,j} \Delta_{ij}(V, W), \\ \Delta_{ij}(V, W) &= 1 \qquad \text{if} \quad V_{ij} \neq W_{ij}, \qquad \text{and} \qquad = 0 \text{ otherwise} \end{aligned} \tag{13}$$

Prior to carrying out the stochastic computation using the Lagrangian, we determine the time-dependence of $d(\text{max.})$ using the basic operational tenets defined in Section 4. A $D - t$ plot ($D = 100\,\text{d(max.)}$) has been generated by performing 10^7 ($\pi - \rho$ loops) of the type indicated in Fig. 2, where π denotes translation and ρ renormalization. Thus, D is determined by computing after every $\tau(\text{read})$ interval the distance between two consecutive LTM's evaluated through two consecutive π-operations. A computation for the functional RNA TtLSU (long splicing unit of *Tetrahymena*), is displayed in Fig. 5. This intensively-studied RNA is a prototype ribozyme or catalytic RNA of the so-called group I.[14, 15]

6. RESULTS

The aim of this section is to establish the Lagrangian structure of our coarse version of the soft-mode chain dynamics described in Sections 4–5. In order to reach this goal, we first consider the fixed partition Z of conformation space X into BPP's. As indicated in Section 2, we already know that the dynamics over Z may be defined by Arrhenius-type elementary activated processes between BPP's. This treatment involves the adiabatic assumption that BPP's may be regarded as quasi-equilibrium states. Now, we shall establish the Lagrangian structure of the LTM dynamics by comparing the time-dependent spreading of the adiabatic probability over Z measured by Shannon information entropy with the time-dependent spreading of the probability over Z determined by the Lagrangian.

The identified Lagrangian L enables us to account for the microscopic origin of the phenomenon of saturation of Shannon's informational content due to the folding process.[2] This thorough saturation of information content has been inferred previously[2] using the adiabatic ansatz. The microscopic treatment requires a comparison between the coarse informational entropy, $\sigma_L(t)$, determined by the Lagrangian-induced stochastic process Ξ and that determined by the adiabatic process ξ. The former is identified as follows:

$$\sigma_L(t) = -\sum_{i=1,\ldots,M} \pi_t\gamma\pi^{-1}\mathrm{BPP}(i)[\ln(\pi_t\gamma\pi^{-1}\mathrm{BPP}(i))] \tag{14}$$

where π is the translation or projection of an LTM into its BPP class and γ is the probability measure determined by Ξ over the space of LTM pathways. Thus, for a given BPP, indicated BPP(i), the quantity $\pi_t\gamma\pi^{-1}\mathrm{BPP}(i)$ is computed as follows:

$$\pi_t\gamma\pi^{-1}\mathrm{BPP}(i) = \frac{\#\{\Xi\text{-generated LTM-trajectories intersecting } \pi^{-1}\mathrm{BPP}(i) \text{ at time } t\}}{\#\{\Xi\text{-generated LTM-trajectories}\}} \tag{15}$$

An understanding of Fig. 6 requires the previous interpretation of Fig. 5. The analysis of Fig. 5 is carried out taking into account the time evolution of patterns of locally-encoded structural signals determined by the time dependence of the DSSM. Within the range 10–10^3 $(\pi-\rho)$ loops we observe a steady growth in the Hamming distance until the maximum

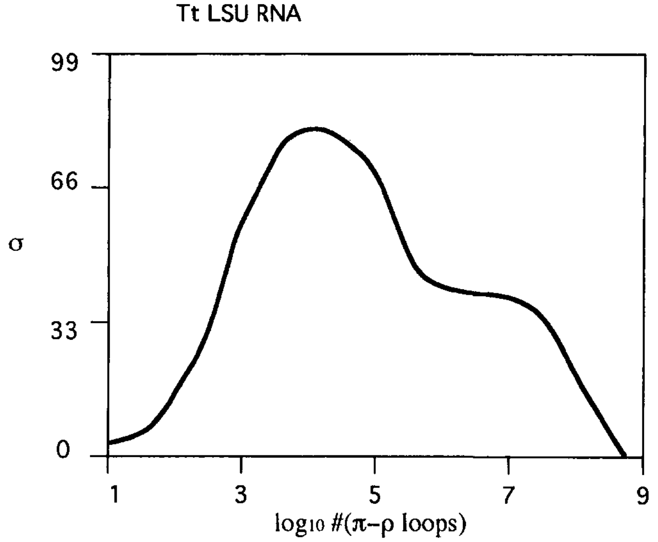

Fig. 6. Time dependence of the coarse information entropy σ_L relative to the partition Z for Tt LSU RNA. The abscissas indicate time in $\#(\pi-\rho)$ loops units. At $T=303$ °K, each LTM $(\pi-\rho)$-evaluation takes place at $\tau(\text{read}) \approx 10^{-9}$ s.

value $D = 30$ is reached. This steady increase in pattern fluctuations is due primarily to the large initial kinetic energy of the patterns, implying that the phase of up to 30 % of the oscillators does not survive two consecutive evaluations (π-translations) of the LTM. The maximum in D is due to the initial formation of "misfolded" short-range structures which are easily dismantled within the μs timescale, as indicated below.

Incipient helices formed in structure-nucleation events in the range 10^{-8}–10^{-5} s are easily dismantled (a 30 % consensus bubble is formed more easily than in fully-developed structures) with concurrent increase in kinetic energy of the LTM pattern due to class II → class I transitions in all nts. formerly belonging to the helical stem as well as its concurrent loops. This fragility of incipient structures causes the large pattern fluctuations marked by a large D in the range 10^{-8}–10^{-5} s, or 10–10^4 $(\pi-\rho)$ loops. A large plateau starting at 10^{-4} s (10^5 $(\pi-\rho)$ loops) marks the formation of a relatively stable kinetic intermediate which contains all structural motifs which may form *noncooperatively*. That is, those motifs whose associated N(loop) lies within the favorable ranges of low conformational entropy cost: $3 \leqslant N(\text{loop}) \leqslant 14$ or $18 \leqslant N(\text{loop}) < 100$.[2, 12] At 10^{-4} s, cooperative events lead to other helices whose loops have favorable *renormalized* sizes, while their sizes relative to the random coil are unfavorable. On the other hand, the increase in class II nts. beyond the formation of the kinetic intermediate stabilizes the patterns determining the survival of the oscillator phase in consecutive LTM evaluations. This determines the relatively-low fluctuations beyond 10^{-4} s.

At this point, we may examine Fig. 6. For short timescales 10^{-8}–10^{-5} s, fast-evolving internal degrees of freedom simulated as torsional oscillators are not yet *enslaved* or entrained by BPP transitions which evolve within typical timescales 10^{-4}–10^{3} s. For this reason, within the range 10^{-8}–10^{-5} s, the level of exploration of conformation space due to uncorrelated or short-range correlated torsional excitations must be vastly larger than that resulting from an adiabatic process. However, an actual adiabatic computation of σ using working Eqs. (2)–(7) is clearly beyond present day capabilities since the number M of a-priori possible BPP's is of the order of e^N.

As soon as the stabilized kinetic intermediate is formed,[16] the long-range correlations coupling distant-row oscillators in the LTM, begin to develop, as cooperative effects occur upon short-range nucleating interactions. This long-range correlations are in turn induced by BPP transitions, in consonance with the nature of the renormalization operation. Thus, initial structure-nucleating steps involving uncorrelated or locally-correlated motions do not demand as much enslavement of fast-evolving torsions as cooperative events, which entail long-range correlations. For this reason, we expect the adiabatic approximation to fit the rigorous results as soon as long-range correlations governed by BPP transitions occur.

The information content as derived from the evolution of patterns at the semiempirical microscopic level reaches its *absolute* maximum within experimentally-relevant timescales,[2, 6] as indicated in Fig. 6. This result is to be expected, as it reflects the expediency and reproducibility of the folding process which yield the active structure within biologically-relevant timescales, thus concentrating the probability distribution in the time alloted to complete renaturation.

In order to test the range of validity of the adiabatic approximation, we compared the L-based computation of the informational entropy, as determined by Eqs. (9)–(15), with the adiabatic computation which stems from Eqs. (2)–(7). Due to the exhorbitant cost in computing time required to iteratively multiply a matrix of order $M \times M$, with $M \sim e^N$, we have confined ourselves to a limiting chain length $N = 18$. Both the adiabatic and the L-based plots are contrasted in Figs. 7 and 8, revealing an almost perfect coincidence with higher than 98 % agreement beyond 8×10^2 $(\pi - \rho)$ interations $\approx 8 \times 10^{-7}$ s. As before, the L-based dynamics requires the previous computation of the size of LTM fluctuations due to $(\pi - \rho)$ iterations. This is displayed in Fig. 9.

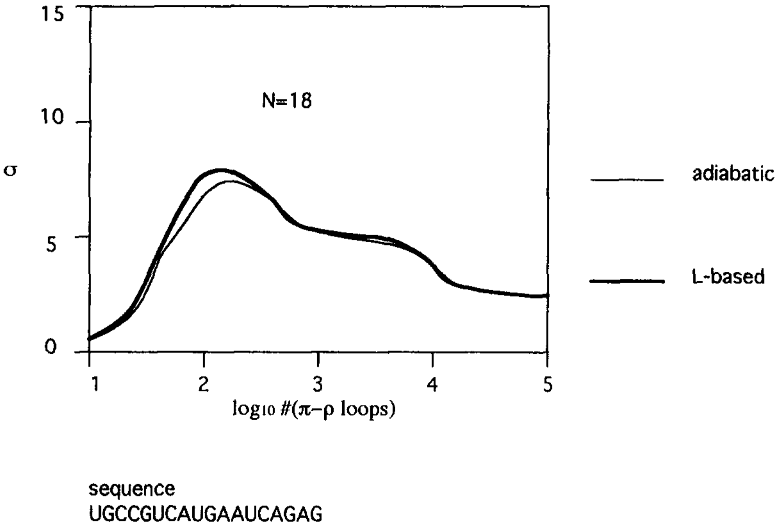

Fig. 7. Time dependence of the coarse information entropies $\sigma(t)$ and $\sigma_L(t)$ relative to the partition Z for a specific randomly generated sequence of length $N = 18$. The abscissas indicate time in $\sharp(\pi - \rho)$ loops units. Each LTM $(\pi - \rho)$-evaluation takes place at $\tau(\text{read}) \approx 10^{-9}$ s.

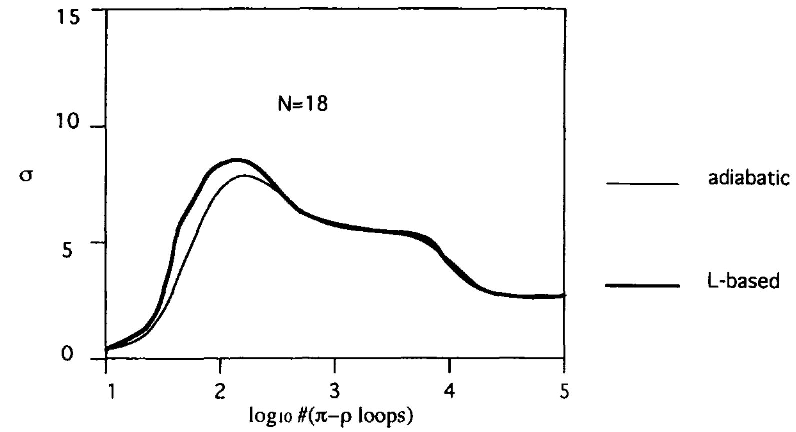

Fig. 8. Time dependence of the coarse information entropies $\sigma(t)$ and $\sigma_L(t)$ relative to the partition Z for an ensemble average over 100 randomly generated sequences of length $N = 18$. The abscissas indicate time in $\#(\pi - \rho)$ loops units. Each LTM $(\pi - \rho)$-evaluation takes place at $\tau(\text{read}) \approx 10^{-9}$ s.

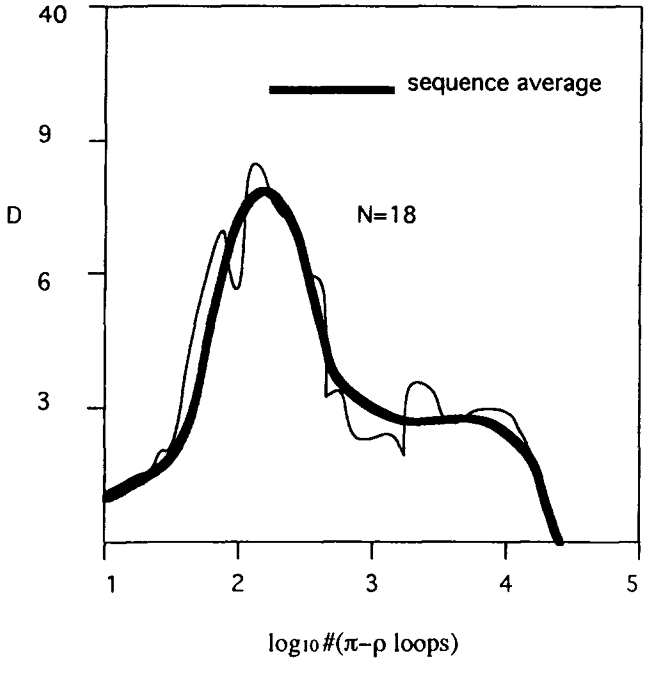

Fig. 9. Plot of the Hamming distance (× 100) between consecutively-evaluated LTM's along the folding pathway generated by iterating $(\pi - \rho)$ loops for a specific randomly-generated sequence and for an ensemble average over random sequences of length $N = 18$.

The same quantitative agreement is found both in the randomly generated sequence of Fig. 7 as well as in the ensemble average over 100 random sequences (Fig. 8). The discrepancy between $\sigma(t)$ and $\sigma_L(t)$ raises to an upper bound of 8 % within the timescale range 10^{-7} s–10^{-6} s. This is clearly due to the microscopic origin of fluctuations which becomes apparent at shorter timescales and is therefore only effectively captured by the L-based dynamics over $(\mathbf{Z}_2)^{3N}$. An inspection of Figs. 7–8 reveals that the dynamics become entrained over the longer timescales (1 μs–1 s) relevant to folding. The initially large fluctuations observed in both computations of the Shannon entropy correspond to noncooperative misfolded structures(16) formed in the 1/10 μs to 1 μs range, most of which are later dismantled to yield a fairly stable cluster of kinetically-related structures.(16) The existence of such a dynamic intermediate state is confirmed by the existence of a plateau sustained within the 1 μs–1/10 ms timescale range.

As observed in Figs. 7 and 8, σ does not tend to zero in the long-time dynamics relevant to the folding timescale frame, as is the case with naturally-selected sequences (Fig. 6). Rather, the coarse entropy decreases asymptotically to a plateau value $\sigma = 2.4$ valid for $N = 18$. This reflects the fact that folding into a unique structure, reaching a sharply-peaked probability distribution within biologically-relevant timescales is not a generic feature of the long-time chain dynamics. However, the coincidence between the adiabatic and the projected Lagrangian behavior is indeed a generic feature of RNA folding because it was obtained irrespective of natural selection, revealing the inherently Lagrangian structure of the coarse microscopic dynamics.

ACKNOWLEDGMENTS

A. F. is principal investigator of the National Research Council of Argentina (CONICET). Partial financial support for this work was provided by the J. S. Guggenheim Memorial Foundation through a fellowship awarded to A. F.. The author thanks Prof. Jerome Percus for valuable ideas and his hospitality during visits to the Courant Institute of Mathematical Sciences (New York University), where this research was first conceived.

REFERENCES

1. T. E. Creighton, Understanding protein folding pathways and mechanisms, in *Protein Folding*, L. M. Gierasch and J. King, eds. (American Association for the Advancement of Science, Washington, 1990), pp. 157–170.
2. A. Fernández, *Physica A-Statistical & Theoretical Physics* **233**:226 (1996).

3. R. L. Baldwin, *Proc. Natl. Acad. Sci. USA* **93**:2627 (1996).
4. (a) A. Fernández, H. Arias, and D. Guerín, *Phys. Rev. E* **52**:R1299 (1995); (b) A. Fernández, H. Arias, and D. Guerín, *Phys. Rev. E* **54**:1005 (1996).
5. K. A. Dill, K. M. Fiebig, and H. S. Chan, *Proc. Natl. Acad. Sci. USA* **90**:1942 (1993).
6. P. Zarrinkar and J. Williamson, *Science* **265**:918 (1994).
7. J. A. Jaeger, D. H. Turner, and M. Zuker, *Proc. Natl. Acad. Sci. USA* **86**:7706 (1989).
8. M. Guenza and K. F. Freed, *J. Chem. Phys.* **105**: 3823 (1996).
9. C. Cantor and P. Schimmel, *Biophysical Chemistry*, Vols. I–III (W. H. Freeman & Co., New York, 1980).
10. C. Brooks III, M. Karplus, and B. Montgomery Pettitt, Proteins: A theoretical perspective of dynamics, structure and thermodynamics, *Advances in Chemical Physics*, Vol. LXXI (J. Wiley & Sons, New York, 1988).
11. A. Fernández, *Zeit. Physik B (Condensed Matter)* **79**:255 (1990).
12. A. Fernández, G. Appignanesi, and H. Cendra, *Chem. Phys. Lett.* **242**:460 (1995).
13. A. Fernández and G. Appignanesi, *Phys. Rev. Lett.* **78**:2668 (1997).
14. R. F. Gesteland and J. F. Atkins, eds., *The RNA World* (Cold Spring Harbor Press, New York, 1993).
15. F. Michel and E. Westhof, *J. Mol. Biol.* **216**:585 (1990).
16. A. Fernández and G. Appignanesi, *J. Phys. A: Math. Gen.* **29**:6265 (1996).