Information generation and the loss of conformational entropy during RNA folding

1996 J. Phys. A: Math. Gen. 29 L433

(http://iopscience.iop.org/0305-4470/29/17/003)

J. Phys. A: Math. Gen. **29** (1996) L433–L438. Printed in the UK

LETTER TO THE EDITOR

Information generation and the loss of conformational entropy during RNA folding

Ariel Fernández†‡ and Alejandro Belinky§

† Instituto de Matemática de Bahía Blanca, Universidad Nacional del Sur, Consejo Nacional de Investigaciones Científicas y Técnicas, Avenida Alem 1253, Bahía Blanca 8000, Argentina
‡ The Frick Laboratory, Princeton University, Princeton, NJ 08544, USA
§ Departamento de Economía, Universidad Nacional del Sur, Avenida Alem 1253, Bahía Blanca 8000, Argentina

Received 2 January 1996, in final form 29 April 1996

Abstract. A stochastic process generating biologically-relevant RNA folding pathways induces a time-dependent probability measure in conformation space. This measure determines the information content which is used here as a marker for the steps that make folding a robust process. For real RNA molecules, the amount of information generated is shown to reach its *absolute* maximum within the experimentally-relevant timescale, well below the thermodynamic limit of ergodic time.

Intramolecular biopolymer folding is a central problem in molecular biophysics. The search for the active structure starting from a denatured random-coil conformation is a robust and efficient process [1, 2] demanding theoretical underpinnings.

This work is partly devoted to determining the degree of efficiency of a biopolymer folding process by measuring the generation of information concurrent with an expeditious search for the biologically-active structure [1, 2]. In this regard, our result specialized for RNA may be anticipated: adopting a random coil as initial condition in a renaturing environment, and choosing a nontrivial coarse description of the atomic aggregate, we observe that the amount of Shannon information generated as a result of folding reaches the *absolute* maximum within experimentally-relevant timescales (10^2 s) for biological RNA molecules [2]. In simple terms, this means that the fate of the system is determined by its evolution relative to the specified coarse description long before the thermodynamic limit of ergodic times is reached. In contrast, the coarse dynamical system entropy [3] for a random sequence of the same length remains relatively large within comparable timescales.

The fact that the coarse system entropy reaches its absolute minimum or, equivalently, that the information content within the coarse description is absolutely maximized within realistic timescales implies that the sequence of folding events warranting the robustness of the folding process is appropriately describable within the coarse representation of the atomic aggregate, and that the completion of this sequence is adequately indicated. This is why we introduce the coarse dynamical system entropy as a marker for the sequence of events which determines the robustness of the folding process. The folding scenario referred to in this treatment is consistent with a rapid collapse of the random coil followed by a slow refinement of the structure with the latter demanding a finer description of the atomic aggregate than the coarse representation used to characterize the robust directing steps.

0305-4470/96/170433+06$19.50

Since it accounts for the expeditiousness of the folding process under stringent time constraints, our main result validates our means of generating RNA folding pathways based on a stochastic kinetically-controlled algorithm that mimics the search in conformation space with sequential minimization of the conformational entropy loss [4]. The predictive value of such an algorithm has been established elsewhere [4, 5] and hinges upon the premise that the biologically-relevant base-pairing pattern (BPP) is the destination structure of the pathway that minimizes at each step the loss in conformational entropy.

To be precise, we shall start by specifying the degree of coarse graining of conformation space X. The RNA chain folds intramolecularly by base-pair association of complementary residues or nucleotides, forming antiparallel stems with a concurrent loss in conformational entropy due to loop formation. In our coarse description, two RNA conformations are regarded as equivalent if they share the same BPP. This equivalence relation determines a partition Z of X in mutually-disjoint equivalence classes. By BPP we not only mean secondary structure, incorporating all planar motifs resulting from hairpin, bulge or internal loops, as displayed in figure 1, but we also incorporate the pseudoknot motif [6]. A pseudoknot forms when the residues in a hairpin loop engage in base pairing with residues outside the hairpin, forming an additional stem and loop region.

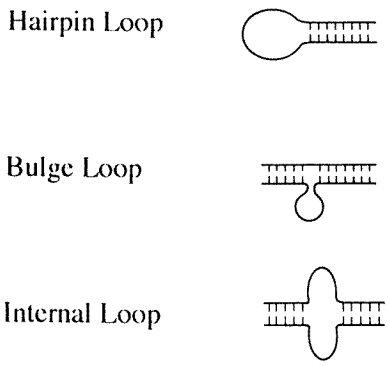

Figure 1. The three kinds of planar loops in RNA structure. The dashed lines represent weak hydrogen bonds between non-adjacent residues. The contact regions form double-stranded helices in the three-dimensional realization of the structure. The bulge loop and the internal loop allow for the coaxial stacking of the two side helices, which may be piled up along a single axis, forming a single entity.

Given the partition Z and a stochastic process ξ defining transition probabilities between elements of Z at each given time, we may define a coarse dynamical system entropy $\sigma(t)$ associated with the partition Z in the following way,

$$\sigma(t) = -\sum_{A \in Z} \pi_t \eta(A) \ln[\pi_t \eta(A)] \tag{1}$$

where A is a generic BPP, η is the probability measure in the space Θ of folding pathways ϑ's($\vartheta \in \Theta$) determined by our stochastic process [7, 8], and $\pi_t \eta$ is the projection of measure η at the instant t. This projection actually yields the weight distribution among BPPs at time t and can be easily computed using our folding algorithm [4, 5],

$$\pi_t \eta(A) = \eta\{\vartheta \in \Theta : \pi_t \vartheta = \vartheta(t) \in A\} = \Omega(A, t)/\Omega \tag{2}$$

where $\Omega(A, t)$ is the number of generated folding pathways passing through A at time t and Ω is the total number of generated folding pathways.

Direct inspection of equation (1) shows that our coarse entropy $\sigma(t)$ is minus a Shannon information content, since $\pi_t\eta(A)$ may be interpreted as the probability of finding a single molecule folded in BPP A at time t.

To obtain a range for σ we first note that $\exp(N)$ is the minimal upper bound for the number of *a priori* plausible BPPs for any chain length N. Since the maximum coarse entropy corresponds to a uniform distribution over all BPPs, and the minimum, to a measure concentrated on a single BPP, we get $0 \leqslant \sigma \leqslant N$.

In order to determine the behaviour of $\sigma(t)$ during the folding of specific RNA molecules, we first specify the process ξ. To implement the process at the computational level, we first make use of current combinatorial algorithms (see, for example, [9]) to predict all plausible BPPs. Such algorithms incorporate the pseudoknot as a tertiary interaction motif and consider only base pairing and stacking as stabilizing interactions in intramolecular structure. We must point out that, although other forms of interaction such as coaxial helix–helix stacking [10] might be crucial to stabilize the BPP by piling up along the same axis two non-adjacent stacked double-stranded regions (see caption for figure 1), the fate of the system is already determined regardless of whether or not such interactions play a significant role. The computations displayed in this work reveal that the entropy σ has virtually reached its *absolute* minimum (0) within biologically-relevant timescales regardless of the fact that we have not incorporated other modes of relaxation such as coaxial helix–helix stacking.

The stochastic process is determined by the activation energy barriers required to produce or dismantle stabilizing interactions. 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 [4, 5]. The process is mechanistically constructed as follows.

For each time $t \in I$, we define a map $t \to J(x,t) = \{j : 1 \leqslant j \leqslant n(x,t)\}$, where $J(x,t) =$ collection of *elementary* events representing conformational changes which are feasible at time t given that the initial conformation x has been chosen at time $t = 0$, and $n(x,t) =$ number of possible elementary events at time t. Associated with each event, there is a unimolecular rate constant $k_j(x,t) =$ rate constant for the jth event [4] which may take place at time t for a process that starts with conformation x. The mean time for an elementary refolding event is the reciprocal of its unimolecular rate constant. Thus, for a fixed time interval I, the only elementary events allowed are elementary refolding events that satisfy $k_j(x,t)^{-1} \leqslant |I|$. We now introduce a random variable $r \in [0, \sum_{j=1}^{n(x,t)} k_j(x,t)]$, uniformly distributed over the interval. Let r^* be a particular realization of r arising in a simulation of the process, then there exists an index j^* such that

$$\sum_{j=0}^{j^*-1} k_j(x,t) < r^* \leqslant \sum_{j=0}^{j^*} k_j(x,t) \qquad (k_0(x,t) = 0 \text{ for any } x,t). \tag{3}$$

This implies that the event $j^* = j^*(x,t)$ is chosen at time t for the folding process that starts at conformation x. Thus, the map $t \to j^*(x,t)$ for fixed initial condition x constitutes a realization of the Markov process which determines the folding pathway ξ_x. In turn, the probability that the j^*-event is chosen at time t is $k_{j^*}(x,t)/\sum_{j' \in J(x,t)} k_{j'}(x,t)$.

Explicit values of the unimolecular rate constants require an updated compilation of the thermodynamic parameters at renaturation conditions [11]. These parameters are used to generate the set of kinetic barriers associated with 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 with closing a loop. Such a loop might be of any of four admissible classes: bulge, hairpin, internal or pseudoknotted. 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 [11]. 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 favour 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 [11] 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 $-\Delta H(\text{stem})$, which is the amount of heat released due to base pairing and stacking when forming all contracts in the stem.

For completion we shall display the analytic expressions for the unimolecular rate constants [4, 12]. If the jth step or event happens to be a helix decay process, we obtain

$$k_j = fn \exp[G_{\text{h}}/RT] \tag{4}$$

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

$$k_j = fn \exp(-\Delta G_{\text{loop}}/RT) \tag{5}$$

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.

At this point we can present computations of the coarse entropy $\sigma(t)$ for specific RNA sequences in order to study its convergence to its absolute minimum $\sigma = 0$, and thus quantify the efficiency of the folding process outside the thermodynamic limit. The behaviour was determined by monitoring σ concurrently with the running of Monte Carlo (MC) simulations making use of working equations (1)–(5). The folding of all 87 catalytic RNAs (ribozymes) of the so-called group I intron family [13, 14] was examined using the compilation of thermodynamic parameters [11] specified by the renaturation conditions of the kinetic experiments presented in [2]. In all cases, the results reveal the convergence of $\sigma(t)$ to within the range of values 0–0.2. This convergence took place 100 s after σ having reached its maximum in the 100 μs timescale. Computational capabilities (100 s–36 min CRAY Y/MP time for a chain length 400) make it almost forbidding to explore the behaviour of σ beyond the 100 s $= 10^6$ MC steps range (in consistence with equations (4) and (5), the order of magnitude in real time is correlated logarithmically with the number of MC steps), nevertheless the trend towards convergence to the absolute minimum is, in all cases, unambiguously apparent. The predicted structure after 100 s was in all cases identical [15] to the philogenetically-inferred secondary structure believed to be catalytically active [13].

The main source of uncertainty in the results undoubtedly stems from the uncertainty in the thermodynamic parameters used to estimate the entropic cost involved in loop closure [11]. This uncertainty grows to up to 20% with loop size and is due to the fact that, unlike

enthalpic contributions, entropic parameters are not determined directly from calorimetric measurements. For this reason, the results using the thermodynamic compilation given in [11] were compared with those obtained from the numerical values resulting from rigorous expressions for the entropic costs associated with loop closure [15]. Within the uncertainty margin, no quantifiable difference in the behaviour of the coarse dynamic entropy has been obtained in any of the systems studied.

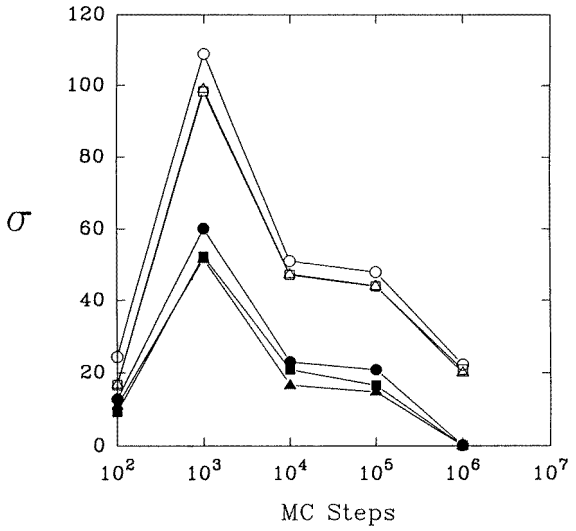

Figure 2. Time-dependent behaviour of the coarse dynamic entropy σ for three specific ribozymes of group I [13, 14]. The solid circles, squares and triangles correspond to TtLSU, T4sunY and TtL21-ScaI, respectively. The open circles, squares and triangles correspond to random sequences of the same respective lengths. The lengths for TtLSU, T4sunY and TtL-21ScaI are 414, 401 and 404, respectively [13].

For the sake of illustration, the results are displayed in figure 2 for three selected ribozymes of the group I family. Standard notation has been adopted and the primary sequences have been obtained from [13]. In order to assess their folding efficiency, a comparison with random sequences of the same length was established. Figure 2 allows for the following interpretation. Within the 100 μs timescale (10^3 MC iterations) a vast exploration of conformation space takes place, with σ reaching its maximum, albeit always strictly below the upper bound N. A drastic reduction of the coarse entropy (slightly higher than 50%) takes place within the 10 ms $= 10^4$ MC steps timescale. This reduction corresponds to the simultaneous occurrence of disjoint non-cooperative folding events leading to stabilizing interactions with low or moderately low kinetic barriers. Such interactions cover the ranges $4 \leqslant L \leqslant 14$ and $24 \leqslant L \leqslant 100$. For biological sequences, the second major and decisive reduction of coarse entropy leads to its absolute minimum. It is achieved in the 100 s timescale, or 10^6 MC steps, and corresponds to the occurrence of cooperative events. Such events result in the formation of interactions of an unfavourable range when viewed starting from the random coil. These interactions require closure of pseudoknotted and complex internal loops which, in turn, require the prior occurrence of nucleating interactions of favourable ranges. The net effect of such nucleating interactions is the shortening of the length of loops for the initially-unfavourable interactions. Nucleating interactions have already occurred in the 10 ms timescale and determine a predictable and

detectable [2, 14] kinetic folding intermediate.

Since the coarse information content has reached its maximum in the 100 s timescale, we can infer that future events demand a finer description of the atomic aggregate, corresponding to further refinement of the structure beyond the robust sequence of steps whose end is marked by σ reaching its absolute minimum.

Upon examination of figure 2 one sees that the fate of the system is determined by the occurrence of interactions of base pairing and stacking type—the only ones included in our simulations—which essentially confine the system to a fixed BPP within biologically-relevant timescales. This observation is supported by the experimental finding that the philogenetically-conserved and active BPP of specific group I ribozymes is achieved in the 1.3 min timescale [2]. In the light of these findings we infer that coaxial helix–helix stacking [10] and other tertiary interactions represent subsequent stages of relaxation to a lower free energy minimum, thus validating a *hierarchical* approach to folding, entirely outside the thermodynamic limit of exhaustive exploration of conformation space.

The results in figure 2 reveal that random as well as biologically-significant RNA sequences both display *qualitatively* the same expeditive folding behaviour up to the stage when the first kinetic intermediate forms within the 10 ms timescale. However, while the real sequence saturates the level of folding involving the occurrence of disjoint non-cooperative folding events in 10 ms, the random sequence, being uncorrelated, takes approximately 100 s (10^6 versus 10^4 MC steps) to reach the same level of structural complexity. Furthermore, the long-time behaviour is radically different. While real RNA strands have minimized their entropic content—or maximized their loss in conformational freedom—in 100 s, random sequences of comparable length take at least four orders of magnitude of that time to reach their entropy minimum as computations starting at the saturated non-cooperative structure reveal.

AF is a principal investigator for 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 AF.

References

[1] Jaenicke R 1984 *Angew. Chem. Intl. Ed. Engl.* **23** 295
Creighton T E 1988 *Bioessays* **8** 57; 1988 *Proc. Natl. Acad. Sci. USA* **85** 5082
[2] Zarrinkar P and Williamson J 1994 *Science* **265** 918
[3] The notion of coarse entropy defined in this work is inspired by Kolmogorov's entropy, first expounded in Kolmogorov A 1933 *Grundbegriffe der Wahrscheinlichkeitsrechnung (Erg. der Math. Bd. 2)* (Berlin: Springer) (in German)
[4] Fernández A 1989 *Eur. J. Biochem.* **182** 161; 1990 *Phys. Rev. Lett.* **64** 2328
[5] Fernández A 1992 *Phys. Rev.* A **45** R8348; 1993 *Physica* A **201** 557
[6] Pleij C W, Rietveld K and Bosch L 1985 *Nucleic Acids Res.* **13** 1717
[7] Kakutani S 1943 *Proc. Imp. Acad. Japan* **29** 184
[8] Fernández A 1994 *J. Stat. Phys.* **77** 1079
[9] Gautheret D, Major F and Cedergren R 1989 *Methods Enzymology* **183** 318
[10] Walter A E, Turner D H, Kim J, Lyttle M H, Müller P, Mathews D H and Zuker M 1994 *Proc. Natl. Acad. Sci. USA* **91** 9218
[11] Jaeger J A, Turner D H and Zuker M 1989 *Proc. Natl. Acad. Sci. USA* **86** 7706
[12] Anshelevich V V, Vologodskii V A, Lukashin A V and Frank-Kamenetskii M D 1984 *Biopolymers* **23** 39
[13] Michel F and Westhof E 1990 *J. Mol. Biol.* **216** 585
[14] Gesteland R F and Atkins J F (eds) 1993 *The RNA World* (New York: Cold Spring Harbor Laboratory)
[15] Fernández A, Appignanesi G and Cendra H 1995 *Chem. Phys. Lett.* **242** 460