Biophysical Chemistry 42 (1992) 1–6
Elsevier Science Publishers B.V., Amsterdam

BIOCHE 01617

Preservation of a kinetically originated folding of the *cis* antirepressor sequence for transport of HIV-1 viral RNA

Alejandro Belinky

Avenida Colón 254 (1º C), 8000 Bahía Blanca (Argentina)

and

Ariel Fernández *

Department of Chemistry, University of Miami, Coral Gables, FL 33124 and Department of Biochemistry and Molecular Biology, The Medical School, P.O. Box 016129, Miami, FL 33101-6129 (USA)

(Received 21 March 1991, accepted 13 June 1991)

Abstract

We compute a metastable secondary structure for the *cis* antirepressor sequence (CAR) in the viral RNA of human immunodeficiency virus 1 (HIV-1) whose lifetime is long enough to allow for further stabilization by interaction with the ribosomal machinery. The structure emerges as the viral genome RNA is being synthesized by RNA polymerase II and corresponds to the biologically active structure sustained between units 7364 and in *env* RNA. It is the most probable among the fast-formed structures which emerge during transcription. No tertiary interactions appear to influence the statistical weight of this metastable state. The structure is predicted by means of a Monte Carlo simulation which computes refolding events occurring as the CAR portion of viral RNA is being assembled. The final emerging structure is preserved for transportation of viral RNA and spliced *env* RNA from the nucleus to the cytoplasm of the host cell.

Keywords: Kinetically governed RNA folding, *Rev* protein; Human Immunodeficiency virus 1; *Cis* antirepressor sequence

1. Introduction

The *Rev* protein is known to play a decisive role in the immediate transportation of the transcribed viral genome RNA and spliced *env* RNA of HIV-1, from the nucleus to the cytoplasm of the host cell [1–3]. This justifies its importance as a positive regulator of the viral protein expression. The role of the *Rev* protein is determined by its interaction with viral RNA as well as with the spliced *env* gene. The target sequence for *Rev* has been well characterized and is known as CAR (*cis* antirepressor sequence), it is localized in the region sustained between units 7335 and 7627 of the *env* RNA [4]. The *Rev*–CAR interaction is mediated by whichever folded structure happens to be adopted by CAR. Such a structure

* Camille and Henry Dreyfus Teacher-Scholar Awardee, and to whom correspondence should be addressed at the Department of Biochemistry and Molecular Biology.

must have a lifetime long enough for further stabilization achieved by interaction with the ribosomal machinery. The aim of this paper is to prove that the biologically-active structure of CAR is the most probable among the fast-forming structures which emerge concomitantly with transcription of the integrated proviral DNA. Recent efforts have been devoted to elucidate the biologically active secondary structure of the CAR [4]. Such studies have made use of a free-energy minimization algorithm for RNA folding [5], combined with functional analysis of the different substructures based on mutagenetic techniques designed to alter one functional subunit at a time. It is worth emphasizing that the folding alternatives which have an active regulatory role vis-a-vis interaction with the *Rev* protein cannot be obtained from an analysis based exclusively on the relative stability of RNA intra-chain secondary structure. In fact, the thermodynamically most stable structure, the equilibrium structure, turns out to be biologically inert (cf. Dayton et al. [4]). Thus, the algorithms for equilibrium folding by themselves cannot aid us in deciding which folding alternative is biologically active. As shown by Dayton and coworkers [4], the functional role of each stem, loop and hairpin in the structure can be assessed by means of mutagenesis, introduced to dismantle or selectively alter specific regions of the secondary structure, followed by a test of the *Rev* response to the mutant. Thus, no *a priori* criterion can be used in folding algorithms based on free-energy minimization in order to decide which of the several local minima corresponds to the biologically active CAR structure. The key to the problem must be given by the computation of probable structures which emerge as transcription of the proviral DNA in the nucleus is taking place. Thus, we shall introduce and verify the following working hypothesis: The biologically relevant structure of CAR is the most probably secondary structure whose formation is kinetically controlled and emerges immediately after transcription by RNA polymerase II has been completed. This structure must be preserved at least until interaction of the *Rev* with the *env* portion of the genome has taken place. Thus, the folding pattern which occurs in the CAR portion of the *env* RNA concomitantly with is assembling is the pattern which prevails and it is the one which aids the regulatory function, the one which optimizes the *Rev* response.

In order to test the hypothesis, we shall compute the structure sustained by the portion of the *env* RNA between nts 7364 and 7559, whose primary sequence is given in Fig. 1. We shall concentrate only on the most probably structure that emerges as that portion of the *env* RNA is being synthesized and compare it with the one inferred by combining thermodynamic considerations with the mutagenetic experiments [4]. The thorough coincidence we shall encounter allows us to state that the "kinetic criterion" yields the correct structure.

2. Methods

In order to compute the most probable among the fast-formed structures, we shall implement a Monte Carlo simulation which predicts the most probable kinetically determined structure formed as the sequence indicated in Fig. 1 is being assembled [6,7]. For the sake of completeness, we shall briefly describe the algorithm:

The separation of timescales involved in the relaxation of metastable RNA secondary structures and in the polymerization events responsible for progressive elongation of the chain prompted the author to analyze refolding events occurring during chain formation as kinetically governed nonequilibrium events [7]. Our simulation is used to obtain the time-dependent probabilities for highly probable transient secondary structures. The simulation makes use of the fact

7364 7415
GUUCUUGGGAGCAGCAGGAAGCACUAUGGGCGCAGCGUCAAUGACGCUGACG

7416 7467
GUACAGGCCAGACAAUUAUUGUCUGGUAUAGUGCAGCAGCAGAACAAUUUGC

7468 7519
UGAGGGCUAUUGAGGCGCAACAGCAUCUGUUGCAACUCACAGUCUGGGGCAU

7520 7559
CAAGCAGCUCCAGGCAAGAAUCCUGGCUGUGGAAAGAUAC

Fig. 1. Primary sequence for the fragment of the *env* RNA between nts 7364–7559.

that, as the chain is being progressively elongated (either during transcription or replication) by sequential incorporation of nucleotides, new possibilities for folding arise and, concomitantly, previously existing metastable structures might be dismantled to allow for the formation of the emerging ones. Thus, the kinetically governed secondary structure formation must be taken into account jointly with chain growth.

The Monte Carlo simulation mimics a Markovian process. Thus, an underlying assumption in our treatment is that the set of polymerization and refolding events on the product of transcription, the viral RNA, constitutes a Markovian chain of events. Consequently, the simulation is comprised of three different kinds of elementary events: (a) intra-chain partial helix formation, (b) intra-chain helix decay and (c) chain-elongation by incorporation of one nucleotide. In addition, we have incorporation certain features absent in previous work: the possibility of G–T and A–C mispairs and the possibility of looped-out bases in the process of helix formation. The transition time for each of the events in the Markov process is a Poisson random variable. For instance, for chain growth by one nucleotide, the mean time is $t = k_{FP}^{-1}$, where k_{FP} is the rate constant for phosphodiester linkage formation which occurs when a new nucleotide is incorporated [8]. If, instead, another elementary event happens to be favored, for instance, an admissible helix formation, the inverse of the mean time for the transcription will be given by:

$$t^{-1} = f N \exp(-\Delta G_{loop}/RT) \tag{1}$$

where f is the kinetic constant for a single base-pair formation (estimated at $10^8 - 10^6$ s^{-1}, cf. Fernández [8], N is the number of base pairs comprising the helix and ΔG_{loop} is the change in free energy change of the set of all loops due to the folding which leads to the new intra-chain stem formation.

On the other hand, if the chosen elementary event happens to be the intra-chain helix decay, the inverse mean time is

$$t^{-1} = f N \exp(G_h/RT) \tag{2}$$

where G_h is the actual free energy of the helix which is being dismantled.

The entropic contribution of the intra-chain loops and the free energy terms for partial helices are taken from the Turner parametrization [9]. The parameters were extrapolated to 25 °C by the author. A comparison with previous compilations, such as that of Salser [10], reveals only minor weighting differences for the most probable transient structures and negligible differences in the probabilities of other members of the ensemble. In addition to the parametrization indicated, we shall impose a realistic cutoff value in the stimulation: the minimum admissible time-span of an intra-chain helix is taken to be 5 10^{-1} s. The cutoff adopted is not arbitrary but corresponds to the minimum lifetime for the most fragile helix which can be formed involving a G–C pair.

The Markov process is simulated by selecting one of the three possible elementary events at each stage. The effective transition time for the chosen elementary event is a Poissonian random variable with mean k^{-1} where the effective rate constant k is given by:

$$k = \sum_{i=1}^{F} k_1(j) + \sum_{j=1}^{D} k_2(j) + k_3 \tag{3}$$

The subindices 1, 2, 3 correspond to events of type (a), (b) and (c) respectively. The indices $= 1, \ldots, F$ label helices that can be formed so that they are topologically compatible with the pattern of existing ones. The latter ones are labelled by the dummy index $j = 1, \ldots, D$. In order to implement the simulation, we shall relabel the rate constants as follows:

$$\begin{aligned} &k = \sum_{m=1}^{M} k'_m, \qquad M = F + D + 1 \\ &k'_1 = k_1(1), \ldots, k'_F = k_1(F), \\ &k'_{F+1} = k_2(1), \ldots \\ &k'_{F+D} = k_2(D),\ k'_{F+D+1} = k_3. \end{aligned} \tag{4}$$

This is done in order to find the transition index m at each stage of the process. Thus, we consider a uniformly distributed random variable

R, $0 \leq R \leq k$, so that if the value r of R lies in the interval

$$\sum_{m=1}^{m'-1} k'_m \leqslant r \leqslant \sum_{m=1}^{m'} k'_m \qquad (5)$$

then, the index m' has been chosen. There is an underlying constraint which is built into the simulation in a combinatorial fashion: Each admissible helix must be topologically compatible with the set of preexisting ones in the sense that no knots are to be allowed.

The time-dependent probability $p = p(t)$ for the most probably secondary structure at time t is readily accessible from our simulations [8].

The results of the simulation, as applied to the fragment indicated in Fig. 1 will be discussed in the following section. The number of steps for all closely related mutants of the sequence given in Fig. 1 is 5×10^4, which corresponds to 22 minutes of Cray-1S computer calculation time. The assumption that the transcription of the integrated proviral DNA is a Markovian process appears to be *correct*, as well we shall presently show.

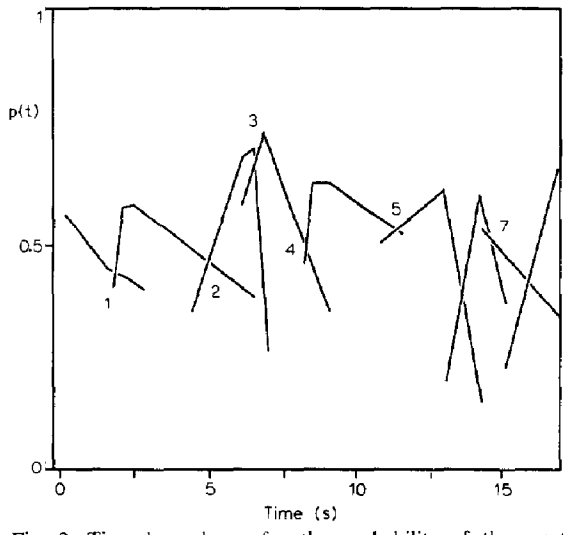

Fig. 2. Time-dependence for the probability of the most probable secondary structures formed as the RNA sequence given in Fig. 1 is being progressively generated. The curve-crossings correspond to re-folding events, where one structure is superseded by another one which emerges with a higher probability.

3. Results and discussion

Before specifying the structural details of the most probably structure for CAR emerging concomitantly with transcription of the viral RNA, we need to point out that no tertiary or long-range interactions have been taken into account in the computation of the statistical weight of the metastable secondary structures. In general, one could argue that the probability of a given secondary structure cannot be determined solely by examining the folding at the level of secondary structures since nonnegligible stabilization of a secondary structure might be achieved by long-range hydrogen bond interactions. However, it is highly unlikely that that might be the case when dealing with kinetically governed structures emerging during transcription. This is so since refolding events which occur as the RNA chain is being synthesized are *local* in nature: they involve regions of at most 40 nucleotides and are responsible for a local reduction of the enzyme footprint which enables the polymerase to move forward along the replication fork [7]. Moreover, long range interactions involving a portion of the RNA smothered by the enzyme environment are precluded.

Figure 2 displays the probability for the most probable secondary structure which emerges concomitantly with chain growth. The points of curve-crossing, indicated by digits, correspond to refolding events and the abscissas associated with them give the times when the refolding events occur. Thus, Fig. 3 displays the loci along the chain where chain growth is temporarily delayed (not only the nt position is given, but also the time when the pause starts) since a refolding event is taking place, the uncertainty in the refolding timespan is 10^{-1} s, and is due to the parametrization uncertainty. Thus, the system is chosing not to incorporate a new nucleotide to the chain until the refolding event has reached completion. The first such event occurs after two seconds and corresponds to the formation of the transient stem involving nts 7364–7367 and nts 7372–7375. All other pauses correspond to the formation of stem/loop substructures (indicated

in Fig. 3 by roman numerals II-VI), to the stem comprised of nts 7383–7389 and nts 7444–7449, and, finally, the stem formed by joining the regions I and I′. This last refolding event occurs after 15.4 s, counting from the starting point, and entails the dismantling of the previously formed partial helical structure formed during the first pause. This structure is dismantled in favor of a new stem formed by bringing together regions I and I′. Thus, the final emerging structure is the one shown in Fig. 3, in complete agreement with the one inferred by Dayton and coworkers [4].

One puzzling aspect of their functional analysis is the improvement of the *Rev* response when the so-called CAR 2/17 mutant was used as a target. The mutant was obtained by altering the subsequences involving nts 7377–7380 (GCAG → UUGU) and nts 7542–7545 (CUGG → ACAA). We ran the stimulation for the mutant and found that the initial transient hairpin, which was formed during the first pause, is not formed in this case, as revealed by the absence of one pause

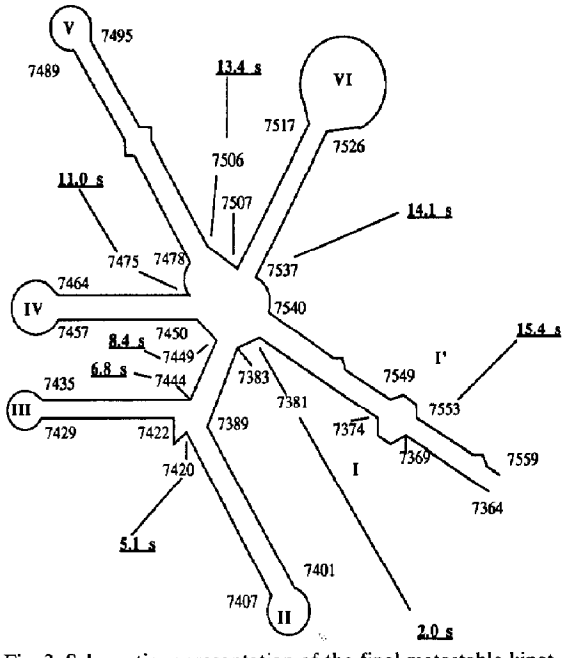

Fig. 3. Schematic representation of the final metastable kinetically governed CAR secondary structure which emerges once the simulation has been carried out to completion. The roman numerical notation is consistent with [7]. The sites labelled with nucleotide number and also with time denote the pause sites, corresponding to curve-crossing in Fig. 2.

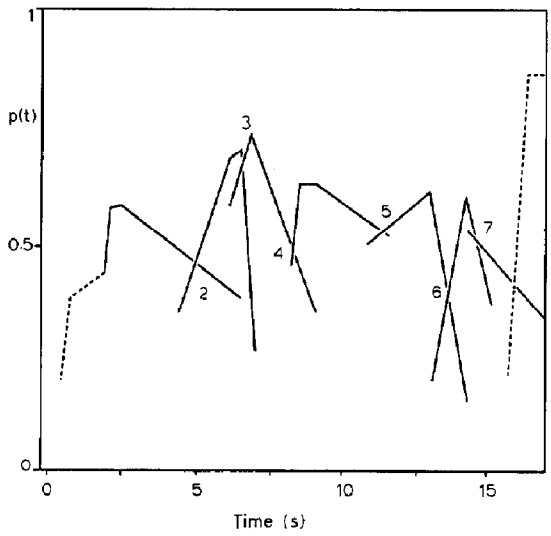

Fig. 4. Time dependence for the probability of the most probable secondary structures formed concomitantly with chain growth for the mutant species CAR 2/17.

which follows by close inspection of Fig. 4. The dashed lines represent the probability for the emerging of those structures in the mutant which are absent in the wild type). Moreover, the last refolding event leads to the I–I′ structure directly, that is, the formation of that structure does not require altering previously-existing ones, as it was the case with the wild type. That, in passing explains why the I–I′ structure emerges with a higher probability in the case of the mutant.

The helix I–I′ seems to play a crucial functional role in the regulation of the CAR–*Rev* interaction, although strikingly, the local recognition seems to be sequence-independent. Thus, in consistence with previous work by the author [6], a mutation like CAR 2/17 should enhance the *Rev* response since it destabilizes the I–I′ secondary structure. This destabilization is a consequence of the decrease in stacking energy due to the decrease in G–C pairing in the mutant with respect to the wild type. The situation is somewhat analogous to that presented in [6], where the replicative activity of a MDV-1 RNA species is improved by destabilizing the folding of the internal recognition site, thus favoring the initial interaction with the replicating enzyme Qβ-replicase. There is an additional feature which could

reinforce or contribute to the enhancement of the enzyme recognition: The kinetic pathway leading to the formation of the I–I′ helix is favored in the mutant when compared to the wild type. This fact follows from direct inspection of Figs. 2 and 4: No previously-existing metastable secondary structure (like the one occurring in refolding event, **1**, Fig. 2) needs to be dismantled to form the I–I′ hairpin for the mutant. Thus, the suppression of a competing refolding pathway might reinforce the effect of the destabilization of the I–I′ helix.

The previous analysis supports the idea that the highly probable fast-formed structure presented in Fig. 3 is the structure which occurs concomitantly with direct transcription of the provirus by RNA polymerase II. This folding must prevail in the fraction of full-length RNA transported from the nucleus to the cytoplasm, at least until the onset of the budding process which leads to virion formation. Moreover, the folding pattern given in Fig. 3 might even be preserved throughout the process of splicing of the portion of viral RNA which yields the mRNA for *env*. Although this needs to be confirmed, it is clear that if the translation of *env* RNA takes place the cytoplasm, the active structure of CAR shou have been preserved for transportation from tl nucleus to the cytoplasm. That must be so sin the *Rev* interaction with *env* RNA is mediat by the folding of CAR.

References

1 B.R. Cullen and W.C. Greene, Cell 58 (1989) 423–426.
2 B.K. Felber, M. Hadzopoulou-Cladaras, C. Cladaras, Copeland and G.N. Pavlakis, Proc. Natl. Acad. Sci. US 86 (1989) 1495–1499.
3 M.H. Malim, J. Hauber, S.Y. Le, J.V. Maizel and B. Cullen, Nature 338, (1989) 254–257.
4 E.T. Dayton, D.M. Powell and A.I. Dayton, Science 2 (1989) 1625–1629.
5 M. Zuker, Science 244, (1989) 48–52.
6 A. Fernández, Naturwissenschaften 76 (1989) 525–526.
7 A. Fernández, Eur. J. Biochem. 182 (1989) 161–163.
8 A. Fernández, Phys. Rev. Lett. 64 (1990) 2328.
9 S.M. Freier, R. Kierzek, J.A. Jaeger, N. Sugimoto, M. Caruthers, T. Neilson and D.H. Turner, Proc. Natl. Aca Sci. USA 83 (1986) 9373–9377.
10 W. Salser, Cold Spring Harbor Symp. Quant. Biol. ((1977) 985–1002.