Volume 297, number 1,2, 201–204 FEBS 10654 February 1992

On how hydrolysis at the 3′ end is prevented in the splicing of a sequentially folded group I intron

Ariel Fernández

Department of Chemistry, University of Miami, Coral Gables, FL 33124, USA, Department of Biochemistry and Molecular Biology, The Medical School, Miami, FL 33101, USA and *Max-Planck-Institut für Biophysikalische Chemie, Am Faßberg 11, W-3400 Göttingen, Germany*

Received 18 November 1991

We propose a dynamic model for the competition between exon–exon ligation and 3′-end hydrolysis valid for sequentially folded pre-mRNA introns of group I. This model accounts for the delay in the formation of conserved helix P10 until the 5′ exon has been cleaved, a requirement to prevent hydrolysis at the 3′ end of the intron. The model is rooted on computer simulations whereby the pre-mRNA searches for its structure as it is being transcribed. Thus, a competing interaction, engaging the internal guiding sequence, occurs initially and prevents P10 from forming until the 3′ end of the 5′ exon is habilitated as a nucleophilic agent. It is further shown that a destabilization of the competing interaction invariably leads to 3′ hydrolysis, crippling the splicing capability of the intron. The results may be probed by site-directed mutagenesis.

Ribozyme; 3′ Hydrolysis; In vivo RNA folding

1. INTRODUCTION

Perhaps the most important reaction competing with exon ligation in the splicing of group I introns is hydrolysis at the 3′ end of the intron. The prevention of this competing reaction poses a major difficulty to experimentalists attempting to engineer splicing conditions. Hydrolysis might be precluded if the formation of the conserved helix P10 is delayed until 5′ cleavage had taken place and the nucleophilic 3′ end of the already spliced 5′ exon is ready to attack the 3′ intron–exon junction [1]. Only then P10 should be allowed to form. We shall verify this scenario by simulating the folding of the fourth intron of the yeast apocytochrome b gene (YCOB4) [2,3], an intron that requires *trans*-acting factors to fold into an active structure and overcome its structural deficiencies.

The chronological order of events demands a **dynamical** model of folding where interaction I′ engages the internal guiding sequence (IGS) until 5′ cleavage disrupts it, precluding the **premature** formation of P10 (see Fig. 1). Thus, the postponement in the formation of P10 finds concrete meaning in a sequential model for the folding of the pre-mRNA under scrutiny. We advocate that the structure is searched concurrently with the assembling of the intron. In other words, folding is concomitant with transcription and results from the cumulative effect of partially preserving the upstream structure as the chain is progressively elongated [4,5].

An implementation of this approach reveals that refolding events during RNA synthesis must be **kinetically** rather than thermodynamically controlled. These events bias the ultimate search for active tertiary structures. Thus, the author has introduced a novel algorithm in the form of a Monte Carlo simulation which handles kinetically-controlled refolding events **together** with polymerization events.

2. METHODS

The simulation mimicks a Markov process such that if at a given stage a refolding event has a larger transition rate than a polymerization event, the former is chosen, whereas, if the reverse holds, the chain grows by incorporation of one nucleotide. The program has been vectorized and partially optimized to run on a Cray operating system. In particular, it may be adapted mutatis mutandis to a Cray Y-MP/24 supercomputer. For the sake of completion, we shall first sketch the general tenets of the simulation. The Markov process is comprised of three different kinds of kinetically-governed elementary events: (I) intrachain partial helix formation, (II) intra-chain helix decay and (III) chain growth by incorporation of a single nucleotide, with a fixed rate of phosphodiester linkage of 50 s^{-1}. The transition time for each event in the Markov process is a Poissonian random variable.

The interactive units are RNA structures accessible for any length N of the chain, starting at N=1. Only loop–stem systems are allowed, as discussed presently. Given two arbitrarily chosen structures '*i*' and '*j*', the rate for the interconversion i---->j is denoted k(i--->j) or, alternatively, k_{ij}, and is defined as:

$$k_{ij}^{-1} = k_{d(i)}^{-1} + k_{f(j)}^{-1} \quad (1)$$

where $k_{d(i)}^{-1}$ denotes the timespan for dismantling the minimal portion

Permanent and correspondence address: A. Fernández, Dept. Biochemistry and Molecular Biology, The Medical School, Miami, FL 33101, USA.

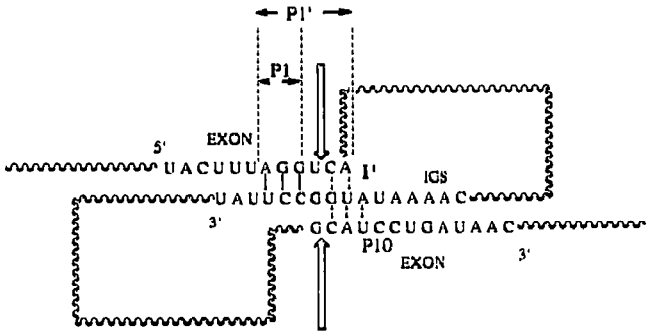

Fig. 1. Scheme of the **competing** base pairing involving the IGS for the YCOB4 intron. The arrows indicate the splicing sites. The IGS-5' intron helix, denoted I', is part of the most probable **but transient** structure P1' which forms during the initial stages of sequential folding. This structure will be dismantled when the extrinsic nucleophilic G-co-substrate attacks the 5'-splicing site, directed by the recently formed catalytic core. Once the IGS has been disengaged, it becomes available for the long-range P10 interaction which may form only when its associated 3' exon segment has been synthesized. This is compatible with the sequence of events in splicing, which dictates that 5' cleavage should **precede** recognition of the 3' terminal G.

of structure i which must be refolded to yield structure j and $k^{-1}_{f(ij)}$ is the timespan of formation of structure *j* starting from a partially- dismantled structure *i*. The only substructures whose disruption or formation we allow are single or multiple stem–loop systems. Thus, two structures whose interconversion requires two transformations (dismantling and refolding) of the type indicated might be connected and they are disconnected if their interconversion requires the occurrence of more than two such events.

The inverse mean time for intra-chain helix dismantling (an elementary event of type II) may be obtained from the expression for the kinetics for helix decay, obtained by Anshelevich et al. [6]. These authors give the equation:

$$k_{d(i)} = t^{-1} = f\,n \exp\,[G_h/RT] \qquad (2)$$

where f is the kinetic constant for a single base-pair formation (estimated at $10^6\ s^{-1}$, cf. [5,6], n is the number of base pairs in the helix and G_h is the (negative) free energy contribution of the set of base-pairs in the helix. If an admissible helix formation (an elementary event of type I) happens to be the event favored, the inverse of the mean time for the transition will be given by:

$$k_{f(ij)} = t^{-1} = f\,n \exp\,(-\Delta G_{loop}/RT) \qquad (3)$$

where ΔG_{loop} is the change in free energy due to the closure of the loop concurrent with helix formation. This contribution corresponds to a loss in conformational entropy. Since water is a relatively good solvent for RNA, excluded volume effects might be significant [3,7] and are accordingly incorporated by correcting the thermodynamic parameters, as discussed in [3].

We may see from Eqs. 1–3 that a kinetically-controlled simulation is dependent on a compilation of thermodynamic parameters. We have adopted the Turner compilation [8] extrapolated to favorable splicing conditions (5 mM Mg(II)) [9]. The robustness of the simulation was tested, especially considering that the compilation is far more suitable for small synthetic oligomers. No significant differences in structure prediction were found within the region of uncertainty in the statistical weights of the transient structures concurrent with RNA synthesis.

If the molecule would fold free from extrinsic or intrinsic perturbations, the structures formed at each stage would always be the most probable among fast-formed structures. The set of such structures is the **optimal** folding pathway. In general, this pathway need not coincide with the in vitro pathway. In the latter, perturbations caused by already-formed transient structures in the upstream region of the RNA transcript might bias the search for new folding alternatives which could arise when more nucleotides are incorporated.

Along the optimal pathway, if structure 'i_o' is chosen by the processor at time t, the next structure chosen would be 'j_o', the structure which realizes the maximum:

$$\mathrm{Max}_j\ k(i_o \text{---}> j) = k(i_o \text{---}> j_o) \qquad (4)$$

At this point, we are in a position to describe the implementation of a parallel realization of the kinetically-controlled simulation. In this extension of the algorithm, **each processor deals with a competing folding pathway**. Competing pathways result from perturbations of the optimal pathway due to occassional base-pair disruptions. However, in dealing with such base-pair disruptions of the transient structures, we shall **not** be able to distinguish or specify the agent causing the perturbation: the parallel algorithm only reveals the specific stages of sequential folding in which a perturbation is demanded to generate ultimately the most probable structure, that is, the structure with the highest statistical weight. The interaction of the growing chain with an upstream part of the ribozyme itself or with a trans-acting factor cannot be modelled explicitly as yet.

To assess the role of an in-vivo environment in perturbing the optimal pathway and in inducing the emergence of competing pathways, we need to introduce a connectivity matrix, as follows:

$$W = [w_{ij}]_{ij} \qquad (5)$$

$$w_{ij} = k[j \text{ ---> } i] \,/\, (\Sigma_n k[j \text{ ---> } n]) \qquad (6)$$

In Eq. 6, i, j and n denote transient structures such that i and any of the n's are accessible from j.

Two essential features in the architecture of the algorithm restrict and direct the computation and make the problem of branching of folding pathways tractable.

(1) The system consists of a set of hierarchically layered processing units (structures). Each layer corresponds to a fixed length N of the chain. Thus, if two structures i and j satisfy N(i)=N(j), then, they belong to the same layer $L_{N(i)}$. Each layer connects via excitatory links with the layer immediately above and receives an input from the layer immediately below.

(2) The output function $f_{N(i)}$ for layer $L_{N(i)}$ allows only certain units to produce an output signal. A unit i with state of activation a_i produces an output $O_i = f_{N(i)}\,(a_i)$ according to the following scheme: Let g be the most probable amongst fast-formed structures in $L_{N(g)}$ and i_o, the structure which realizes $\mathrm{Max}_i\ k(g \text{---}> i)$, then we define the output function $f_{N(i)}$ as follows:

$$f_{N(i)}\,(a_i) = \begin{cases} a_i & \text{if: } |a_i - a_{io}| < h[N(i)]; \text{ where} \\ & h[N(i)] \sim \exp\,[-\beta\Delta\ N(i)^{1/4}] \\ o & \text{otherwise.} \end{cases} \qquad (7)$$

The output function has been chosen in this way to incorporate structure fluctuations in the form of base-pairings and base-pair disruptions. Such fluctuations determine the branching of the optimal folding pathway. Thus, the constant $\Delta \approx 1.81\ k_B T$ (k_B = Boltzmann constant) from Eq. 7, is the scaling factor for the minimal activation energy of a refolding event [5]. The minimal activation energy between mutually-accessible foldings for a chain of length N is $E_a \sim \Delta N^{1/4}$. Thus, branching of a folding pathway becomes increasingly rarer as we approach higher layers.

At this point, we may define the input of structure i at time t as:

$$a_i\,(t) = \Sigma_j w_{ij} O_j(t) \quad ; \quad 0 \le a_i \le 1 \qquad (8)$$

Upon examination of Eqs. 5, 6 and 8, we may conclude that **the input a_i is to be intrepreted as the probability or statistical weight of structure** i.

3. RESULTS AND DISCUSSION

At this point, we may apply the parallel computation to the illustrative case of the YCOB4 intron. These results will contribute to establish a **dynamical** model for splicing based on sequential folding, allowing us to reproduce the required time ordering of splicing events.

In order to determine the stages of folding where perturbations are required, we first simulate the optimal folding pathway. This pathway is subsequently compared with the main folding pathway generated in the parallel extension of the algorithm. By 'main folding pathway' we mean the pathway which finally leads to the most probable structure. The results will reveal that there are two occasions where disruptions from the optimal pathway are required to generate the structure which carries the highest statistical weight: (a) a *trans*-acting factor is required to form the highly conserved helix P7, responsible for furnishing the G-binding site and (b) an already-formed upstream structure of the transcript itself must interact with the 5'-splicing site in order to make the IGS available for the long-range interaction P10.

This will be proven by establishing the following facts.

(a) The most probable structure which emerges immediately after the intron has been fully synthesized features a catalytic center while the 3' exon remains disengaged from P10 pairing. The shaping of the catalytic center requires the prior disruption of interactions involving the highly conserved segment R and the internal open reading frame (ORF).

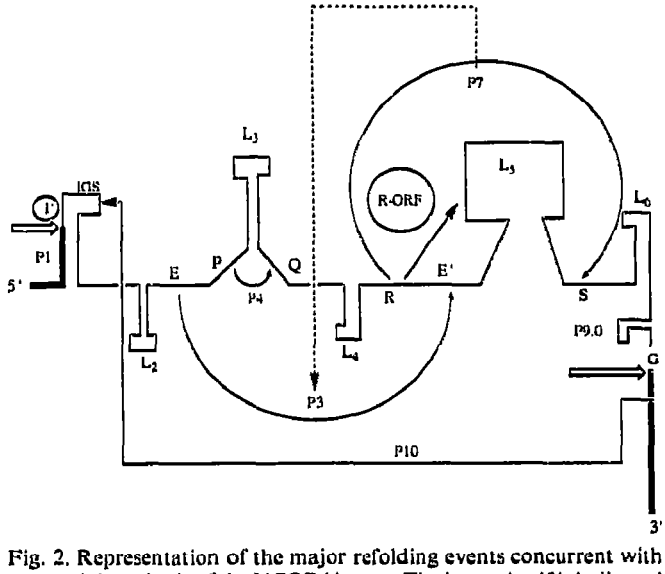

Fig. 2. Representation of the major refolding events concurrent with sequential synthesis of the YCOB4 intron. The intron itself is indicated by a thin line, thick lines represent exons. The splicing sites are indicated by thick arrows. The coding region for the maturase (ORF) is folded around the complex loop L_5. The major single loops formed concomitantly with polymerization are labelled L_2–L_6. The major conserved sequences are indicated by capital letters, following standard notation, revealing their relative location along the sequence. Structure-clustering occurs stepwise, by means of long-range interactions P3 and P7. **Conserved interaction P8, whose pairing segments are separated by the ORF, is entirely absent from the main folding pathway.** This circumstance facilitates the formation of the tertiary interaction indicated by the dashed arrow (see main text). The interactions denoted by circles are disrupted along the main folding pathway by a perturbing agent. The disruption of the R-ORF interaction promotes the formation of highly conserved helix P7 and the subsequent disruption of I' engages the IGS in the formation of the P10 helix.

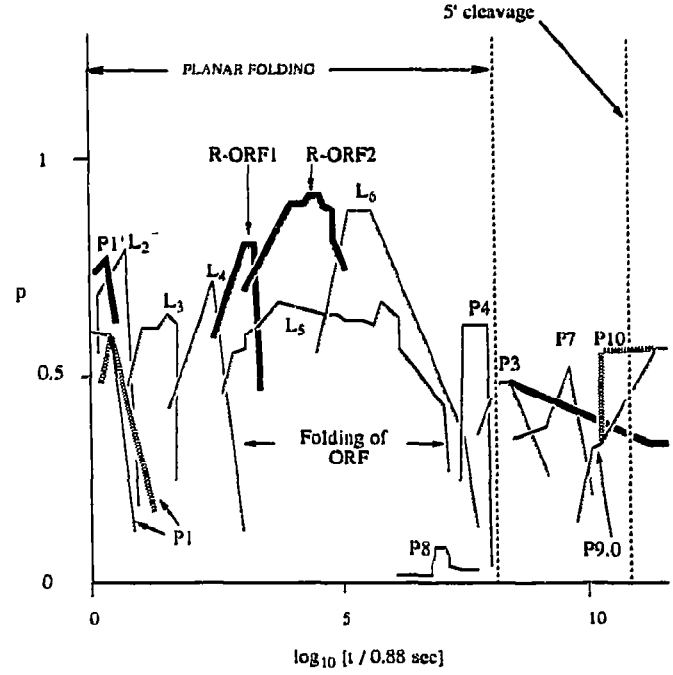

Fig. 3. Probability of transient structures along the main folding pathway. The thick lines represent structures which are not disrupted along the in vitro pathway. The thin solid line plot represents the main folding pathway occurring in vivo and the dashed lines correspond to the species resulting from the site-directed mutagenesis C---->A at the 5' end of the intron. A convenient logarithmic scale, monitoring the passing of time on the abscissas has been adopted. The conserved segment R pairs initially to the ORF forming R-ORF1 and this interaction is spontaneously dismantled to form the more probable helix R-ORF2. **Interactions P7 and P10 are absent in the in vitro pathway since they require prior disruption of existing structures. The terminal structure along the in vitro pathway is therefore inactive for splicing.** We may see that, in spite of the fact that disruptions lead to less probable structures at the time when they occur, their net consequence is to lead finally to a structure which carries a larger statistical weight (the active structure is more probable than the inactive one). The probability $P=P(i)$ of a given structure has been identified in the text with the input for that structure (Eq. 8). As shown in the figure, this probability increases as subsequent incorporation of nucleotides makes the structure more feasible and decreases when chain-elongation events lead to a better folding alternative. A salient feature distinguishing the folding of the mutant from that of YCOB4 is that formation of P10 is premature, since it does not require prior 5' cleavage. Thus, we may predict this species should be inefficient for exon–exon ligation.

(b) The helix P10 emerges when a subsequent disruption occurs at the 5′-splicing site. This disruption is attributed to an intrinsic factor (the catalytic core) which, in turn, required an extrinsic factor to be shaped.

Perturbations of the optimal folding pathway occurring during the initial stages of folding are conveniently studied starting at level L_{28}. This choice is made so as to allow the IGS to be engaged in the formation of the 5′-splicing site. The simulation is actually initiated at the seventh nucleotide of the 5′ exon, starting counting from its 3′ extremity. Thus, the choice of the initial layer is made so that all possible disruptions of the optimal structure P1′ are contemplated. The disruption of structure I′ is allowed to take place since this event belongs to the neighborhood of the first refolding event. In rigorous terms, the disrupted structures belong to the support of f_{28}. The disruption of structure I′, lowers the probability of the initial structure from 78 ± 2% to 64 ± 2% during the initial 0.8 s. Moreover, the disruption of I′ is the only alternative folding pathway which occurs when fluctuations are incorporated. The long-time effects are crucial, however, since they result, 108 s after synthesis started, in the formation of helix P10 which is responsible for activating the 3′-splicing site.

The folding pathway leading to the most probable structure is represented in Figs. 2 and 3. Direct inspection of Figs. 2 and 3 leads us to conclude that the separation between the conserved segments R and S by a long ORF makes the formation of the crucial G-binding-competent helix P7 possible only because, as we predict, an extrinsic factor disrupts the R-ORF interactions. This factor may be the translation machinery acting concomitantly with transcription or an intron-encoded maturase [10]. On the other hand, the occurrence of these interactions is intuitively obvious since chances of partial complementarity between R and the 1018 nucleotides-long ORF are solid. Moreover, the absence of conserved interaction P8 along the main folding pathway favors the stacking of helices P3 and P7, **thereby fulfilling a structural demand of the catalytic core.**

At this stage in the analysis, we ought to justify the following assumption: we have assigned the disruption of helix I′ to an interaction with part of the intron itself, acting as a perturbing agent. This is justified by following the in vitro pathway up until helix P4 is formed (see Fig. 3), and, thereafter, simulating in parallel the branching from the in vitro pathway. In this case, we have found that helix I′ prevails in all competing pathways. This indicates that helix I′ will not be disrupted unless the disruption of the R-ORF helices had occurred previously. Thus, we may conclude that the disruption of helix I′ is not due to an extrinsic or *trans*-acting factor but to an intrinsic factor involving the catalytic core itself. Moreover, **the competition between I′ and P10 is not accidental but necessary to prevent the premature formation of P10 which, as revealed in Fig. 3, forms strictly after 5′ cleavage.**

A natural probe of the proposed mechanism for preventing the premature formation of P10 is to perturb the interaction I′. This could be achieved by site-directed mutagenesis at the 5′ end of the intron (see Fig. 1). Thus, the mutation C--->A at the 5′ end would preclude the competing interaction I′ from occurring altogether. As shown in Fig. 3, this mutation should not alter 5′ cleavage. On the other hand, P10 now forms readily and independently of whether 5′ cleavage had occurred. This situation is responsible for the premature formation of P10 since this event is no longer associated to a prior shaping of the nucleophilic 3′ end of the 5′ exon. We predict that the mutation indicated should render the intron inefficient for exon–exon ligation, causing hydrolysis at the 3′ end to prevail.

Acknowledgement: The author is a Camille and Henry Dreyfus Teacher-Scholar.

REFERENCES

[1] Michel, F. and Westhof, E. (1990) J. Mol. Biol. 216, 585–610.
[2] Davies, R.W., Waring, R.B., Ray, J.A., Brown, T.A. and Scazzochio, C. (1982) Nature 300, 719–723.
[3] Fernández, A. (1991) Chem. Phys. Lett. 183, 499–507.
[4] Fernández, A. (1989) Eur. J. Biochem. 182, 161–164.
[5] Fernández, A. (1990) Phys. Rev. Lett. 64, 2328–2331.
[6] Anshelevich, V.V., Vologodskii, V.A., Lukashin, A.V. and Frank-Kamenetskii, M.D. (1984) Biopolymers 23, 39–58.
[7] Chan, H.S. and Dill, K.A. (1989) J. Chem. Phys. 90, 492–501.
[8] Turner, D.H., Sugimoto, N. and Freier, S.M. (1988) Annu. Rev. Biophys. Biophys. Chem. 17, 167–192.
[9] Puglisi, J.D., Wyatt, J.R. and Tinoco, I. (1988) Nature 331, 283–286.
[10] Kim, S.-H. and Cech, T.R. (1987) Proc. Natl. Acad. Sci. USA 84, 8788–8792.