International Journal of Theoretical Physics, Vol. 31, No. 6, 1992

Structural Organization of an RNA Catalyst with the Random Energy Model As a Reference Frame

Ariel Fernández[1]

Received March 5, 1991

We investigate the kinetics of refolding of a catalytic RNA species along a folding pathway which starts at a biologically active structure and ends in the most stable structure, or global free energy minimum. The RNA species studied is not a purely informational molecule, since it is spliced *in vivo*. We show that the kinetic barriers involved in the relaxation of the active structure are considerably larger than those of a purely informational RNA. Our results attempt at establishing a kinetic criterion to distinguish between a catalytic and a coding RNA species on the basis of their folding.

1. INTRODUCTION

Intron self-splicing (Cech, 1988; Inoue *et al.*, 1985) constitutes the epitome of RNA catalytic activity, a phenomenon whose very existence challenges the division of biomolecules into informational and functional categories. An RNA molecule capable of self-splicing acts autocatalytically, excising the intron and ligating the sequences that are functional in the mature RNA (exons). The overall reaction, be it self-splicing or simply splicing (requiring an *in vivo* environment), comprises two trans-esterification reactions along the RNA backbone and could be adequately described as a cutting-and-pasting process by which the two exons are joined together. This is done in order to properly instruct the subsequent synthesis of a protein, so that the coding RNA is available for translation.

It is only natural that there must be essential differences in the way of folding of an RNA species which exerts a function, or has a catalytic role, and an RNA species which is purely informational. This poses the problem of how to define those differences in a convenient fashion. The present

[1]Department of Chemistry, University of Miami, Coral Gables, Florida 33124, and Department of Biochemistry and Molecular Biology, The Medical School, Miami, Florida 33101-6129.

0020-7748/92/0600-0983$06.50/0

work suggests that there is a definite trend in the way in which introns fold and that the resulting structures are far more organized when compared to those of flanking exons or internal open reading frames of similar length. This common motif must become apparent irrespective of whether the intron is capable of self-splicing or whether it requires an *in vivo* environment for the shaping of the catalytic site. In this paper we shall concentrate on a representative example of the YC4 intron (Davies *et al.*, 1982) and examine its refolding events vis-à-vis those of an open reading frame. Ultimately, we aim at introducing an operating hypothesis which would allow us to distinguish between the folding of an autocatalytic RNA and a purely informational RNA.

The search for optimal or active RNA structures starts with the very first refolding event during the synthesis of the molecule (Fernández, 1990). Given the relatively short biological timescales involved [approximately 15 sec for the synthesis by transcription of a fragment 220 nucleotides long (Fernández, 1989)] the exploration of configuration space concomitant with chain growth is heavily time-constrained. We proposed earlier a Monte Carlo simulation which handles kinetically-controlled refolding events together with sequential polymerization events, in an attempt to account for biological time constraints. The grounds for implementing the simulation can be found in the fact that the actual search for structures appears to be determined by how fast they are able to form while chain growth takes place.

2. METHODS

The simulation mimics 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 X-MP/24 supercomputer. For the sake of completion, we shall first sketch the general tenets of the simulation. The Markov process comprises three different kinds of *kinetically-governed* elementary events: (I) intrachain partial helix formation, (II) intrachain helix decay, and (III) chain growth by incorporation of a single nucleotide. The transition time for each event in the Markov process is a Poissonian random variable. If an admissible helix formation happens to be the event favored, the inverse of the mean time for the transition will be given by

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

where f is the kinetic constant for a single-base-pair formation [estimated at 10^7 sec (Anshelevich *et al.*, 1984)] n is the number of base pairs comprising

the helix, and ΔG_{loop} is the change in free energy of the set of all loops due to the folding which leads to the new intrachain stem formation. The formation of new helices should always be topologically compatible with the pattern of existing ones in the sense that no knots can be allowed in the core (planar) secondary structure. This condition has been given proper combinatorial form and as such is incorporated in the algorithm in a standard manner.

If intrachain helix decay is the chosen event, the inverse mean time can be obtained from an improved version of the expression for the kinetics for helix decay, obtained by Anshelevich *et al.* (1984). These authors give the equation

$$t^{-1} = fnS(\text{eq.})^{-n} \tag{2}$$

where $S(\text{eq.})$ is the equilibrium constant for base-pair formation. However, their treatment does not properly distinguish between stacking and initiation of the base-pairing process. Thus, we shall use instead the improved equation

$$t^{-1} = fn(KS^{n-1})^{-1} \tag{3}$$

where S is the geometrical mean of the base-stacking equilibrium constants (adequate for a random uncorrelated primary sequence) and K is the equilibrium constant for base-pairing initiation (nucleation equilibrium constant); $K(\text{A-U}) \approx 4 \times 10^{-5}\ \text{M}^{-1}$ and $K(\text{G-C}) \approx 2.5 \times 10^{-4}\ \text{M}^{-1}$.

Finally, if a polymerization event happens to be favored, the rate constant for phosphodiester linkage formation $t^{-1} \approx 50\ \text{sec}^{-1}$ (Fernández, 1989) should be adopted as transition rate.

The Markov process is simulated by selecting one out 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

$$k = \sum_{i=1}^{F} k_1(j) + \sum_{j=1}^{D} k_2(j) + k_3 \tag{4}$$

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

$$\begin{aligned} k &= \sum_{m=1}^{M} k'_m, \qquad M = F + D + 1 \\ k'_1 &= k_1(1), \ldots, k'_F = k_1(F) \\ k'_{F+1} &= k_2(1), \ldots \\ k'_{F+D} &= k_2(D),\ k'_{F+D+1} = k_3. \end{aligned} \tag{5}$$

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

$$\sum_{m=1}^{m'-1} k'_m \leq r \leq \sum_{m=1}^{m'} k'_m \tag{6}$$

then the index m' is chosen.

The only tertiary interaction incorporated in the simulation is the pseudoknot motif (Pleij *et al.*, 1985). By doing so, we do not introduce a breakdown of the additivity of free energy contributions for separate loop-stem systems and therefore our computation times do not grow beyond the order of minutes in a Cray X-MP/24. However, severe restrictions on the type of tertiary pseudoknotted interactions allowed must be incorporated since our present knowledge of the associated enthalpies of stacking and of the entropic contributions of loops is scant. These restrictions are summarized as follows:

(a) No steric constraints imposed by loop size are allowed for any of the two loops in a pseudoknot. Neither the loop which crosses the deep groove nor the one which crosses the shallow groove is allowed to have less than three nucleotides for a quasicontinuous pseudoknot stem of six consecutive base pairs (Pleij *et al.*, 1985).

(b) The size of the loops of hairpins which could potentially compete with pseudoknots is sufficiently large so that a pseudoknot is always stabilized over any alternative hairpin of equal or smaller number of base pairs. In concrete terms, the loop of a hairpin which might be an alternative to a pseudoknot must have more than six unpaired bases. This constraint is rooted in the well-established fact that hairpins with up to six nucleotides are thermodynamically more favorable than those with larger loops (Groebe and Uhlenbeck, 1988).

(c) The number of base pairs in each of the stems whose partial stacking produces the pseudoknot stem is three and the partially-stacked helix resulting from the coaxial joining of the two separated stems has six base pairs. This restriction is fulfilled, for instance, when the intron posesses an internal guiding sequence (IGS) which functions as intramolecular adaptor (see below). The restriction implies that we shall only admit a quasicontinuous helix resulting from partial stacking of the two stems having a degree of base-pairing well below one full helical turn, thus avoiding formation of a knot. In turn, this last condition is essential to make the problem tractable: If a knot were to occur, the premise of additivity of free energy of separate regions would be violated.

These three restrictions enable us to treat the entropic contribution for loop formation leading to a pseudoknot as a convex function of the number

of unpaired bases in the loop, precisely as in standard secondary structure. In order to allow for the possibility of large loops forming part of pseudoknots [of the order of several hundreds of unpaired bases (Pleij *et al.*, 1985)], we shall adopt an accurate convex function for the entropic contribution, $S \sim \ln u$, where u is the number of unpaired bases. This substantially increases the timespan of the simulation as well as the order of the solution algorithm. Nevertheless, it is entirely justified given that we are interested in eventually computing long-range interactions which take place once polymerization steps cease to occur and are responsible for the creation of a catalytic mode. The enthalpy of stacking for pseudoknotted quasi-continuous A-RNA helices will be taken to be 66% (≈two-thirds) of the enthalpy of formation of a continuous helix made up of the same base pairs. These enthalpy parameters are in turn obtained from the Turner *et al.* (1988) compilation. The empirical fit seems suitable for synthetic oligomers capable of forming pseudoknots at 5 mM Mg(II), for which detailed thermodynamic studies have been performed in Tinoco's laboratory (Puglisi *et al.*, 1988).

The key quantity accessible from the simulation is $p(n, t)$, the probability of a structure n at time t. This probability is given by

$$p(n, t) = \left\{ \sum_{\mu} k(n-1 \to \mu, t) \right\}^{-1} \left\{ k(n-1 \to n, t) - \sum_{\beta} k(n \to \beta, t) \right\} \qquad (7)$$

provided the following inequality holds:

$$k(n-1 \to n, t) \geq \sum_{\beta} k(n \to \beta, t) \qquad (8)$$

where α, β, μ, n, and $n+1$ denote folding patterns occurring during replication and $k(\alpha \to \beta, t)$ is the time-dependent rate of refolding of structure α to yield structure β. These rates depend solely on the transition rates for whichever elementary events are required to refold the first structure into the second one. Obviously, all rates $k(\alpha \to \beta, t)$ will vary as more nucleotides are incorporated to the growing chain and, in this implicit sense, the rates are time-dependent. The probabilities $p(n, t)$ are kinetically-determined and obviously path-dependent, since they are defined inductively. The inequality given in (8) simply implies that the only structures considered are those with a nonnegligible lifetime, whose rate of formation is larger than the overall rate of decay.

3. RESULTS AND DISCUSSION

Our simulations reveal that the YC4 intron adopts initially a planar structure as it emerges from the transcriptional machinery. This structure is obtained monitoring refolding events concomitant with chain growth

until the 20th nucleotide on the exon following the 3′ end of the intron has been reached. We shall denote it the core structure. The next task consists in monitoring subsequent refolding events after we have precluded polymerization events any further than the 20th exon nucleotide. This is done in order to predict subsequent long-range interactions for the YC4 intron, which are partly responsible for the maturation of the mRNA (that is, for the splicing and joining of the sensical portions of the molecule). We show that in a first stage, a core *metastable* secondary structure emerges after the full YC4 intron, including a subsequence coding for the enzyme maturase (Lazowska *et al.*, 1980), has been synthesized. The core structure is the most probable short-ranged planar structure among those which can form within the timespan of transcription of the YC4 intron (see Figures 1–3). It contains

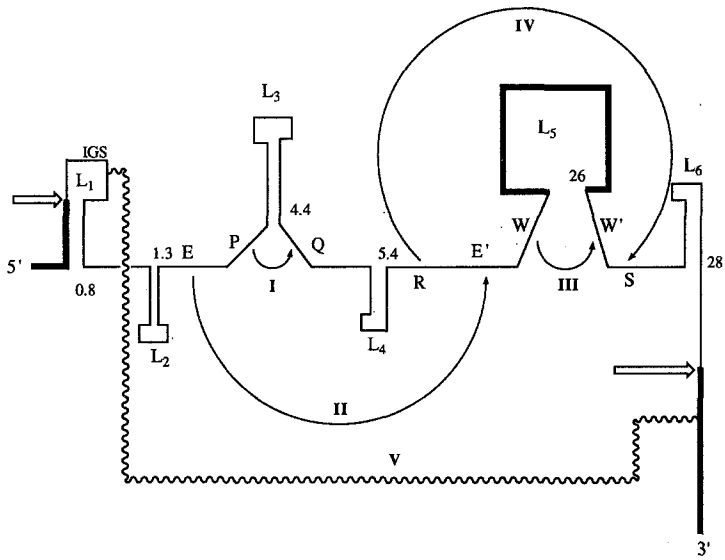

Fig. 1. Representation of the core secondary structure (planar) for the YC4 intron and subsequent clustering folding events during maturation of fungal mitochondrial mRNA, as obtained from the Monte Carlo simulation. The intron itself is indicated by a thin line, thicker lines represent exons. The coding region for maturase is folded around the complex loop L_5. The major loops formed concomitantly with polymerization in the core planar secondary structure are labeled L_1–L_5. The major conserved sequences are indicated by capital letters, revealing their relative location along the sequence. The notation is consistent with Davies *et al.* (1982). The numbers along the chain indicate the instants in seconds when polymerization does not progress further until a refolding event upstream has taken place. The clustering folding occurs stepwise, by means of long-range interactions denoted I–V. Three of these interactions, II, IV, and V, are tertiary. The latter, represented by a wiggly line, is responsible for the alignment of the splicing sites marked by thick arrows. The core structure becomes nonplanar ever since the occurrence of clustering step II.

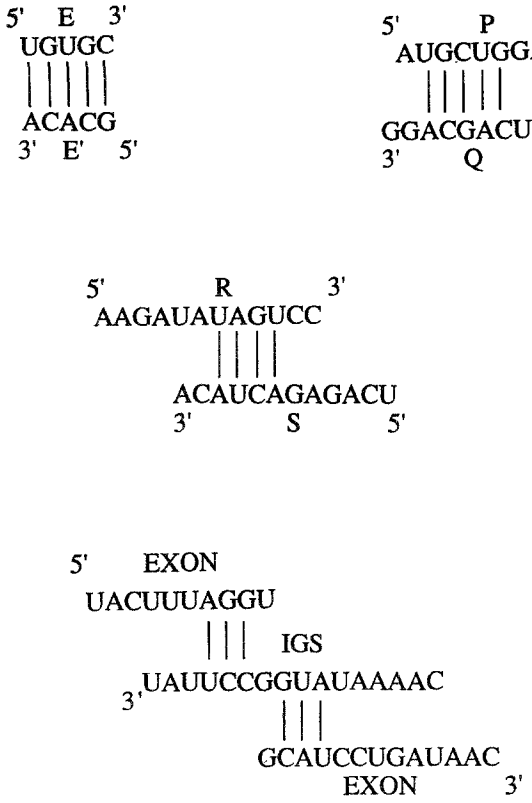

Fig. 2. Base-pairing in the major interactions among conserved sequences, responsible for subsequent folding of the core secondary structure. All interactions are long-ranged secondary or tertiary.

the two-base-pair interaction denoted I' and involves also the planar folding of the coding region for the protein maturase. We also observe that the nucleation and clustering of the core structure occur progressively in four kinetically-determined steps which must be completed within the maturation timespan.

The numbers in Figure 1 indicate the instants in seconds when polymerization stops in order to allow for a refolding event to occur. The overall timespan for the Markov process comprising polymerization and refolding events up to the downstream splicing site is 28 sec. Five major loops $L_1, \ldots, L_5$ are generated. The number of unpaired bases in each loop is, respectively, 6, 8, 23, 8, and 108. The last loop is actually a complex unknotted loop involving the 1120-nucleotide-long coding region. Thereafter, polymerization elementary events are assigned zero probability,

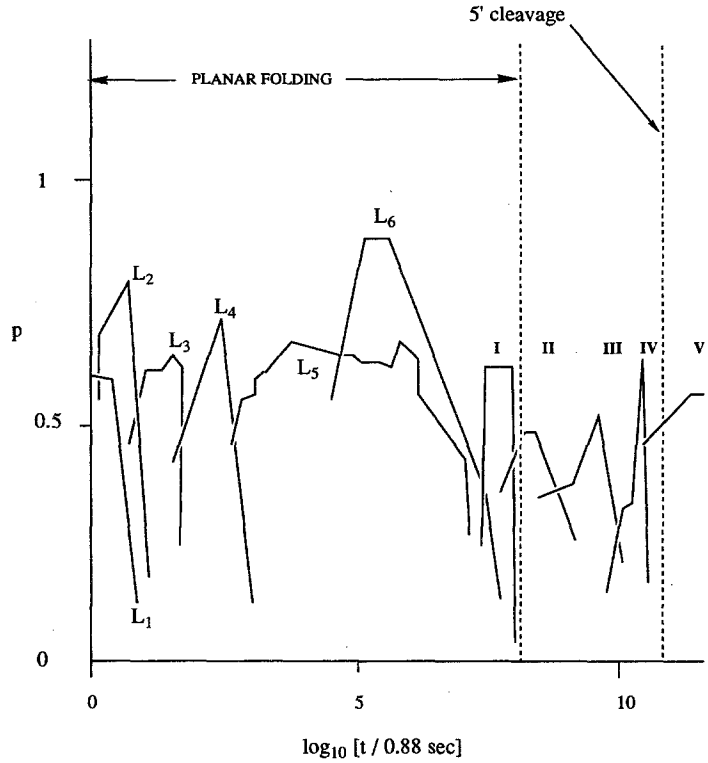

Fig. 3. Probability of the most probable structures along the folding pathway. The logarithmic scale for the abscissas, chosen to display the passing of time, was adopted for convenience.

the no-knot constraint is relaxed, and the clustering by means of long-range interactions of the core metastable structure into a higher organization is monitored. The overall timespan of the simulation is 14 min of Cray X-MP/24 time. The clustering process leading to a nonplanar structure comprises four steps denoted I–IV in Figure 1.

The steps, with the exception of closure of the largest loop, involve long-range interactions among conserved sequences (Burke, 1988) in mitochondrial introns. The explicit interactions are specified in Figure 2. Steps I and III entail secondary interactions, whereas the remaining steps involve tertiary interactions. The instants when initiation or nucleation of helix formation takes place are 30, 38, 41, and 56 sec for steps I–IV, respectively. Step V is the one that brings the downstream exon in contact with the IGS and thus activates the molecule by bringing the splicing sites together.

Time evolution along the folding pathway is revealed in Figure 4. The lines represent the probabilities for the most probable structure at each given instant t. The first sequentially-formed structures L_1–L_6 and I, yield the planar structure. A higher degree of structural organization is subsequently achieved with long-range interactions II–V. The final interaction V becomes now feasible, leading to the partial stacking of the helices IGS-5′ exon and IGS-3′ exon. This coaxial alignment is responsible for the shaping of the catalytic site, bringing the two splicing sites to close spatial proximity. However, since it is a pseudoknotted stacking interaction, its enthalpy of

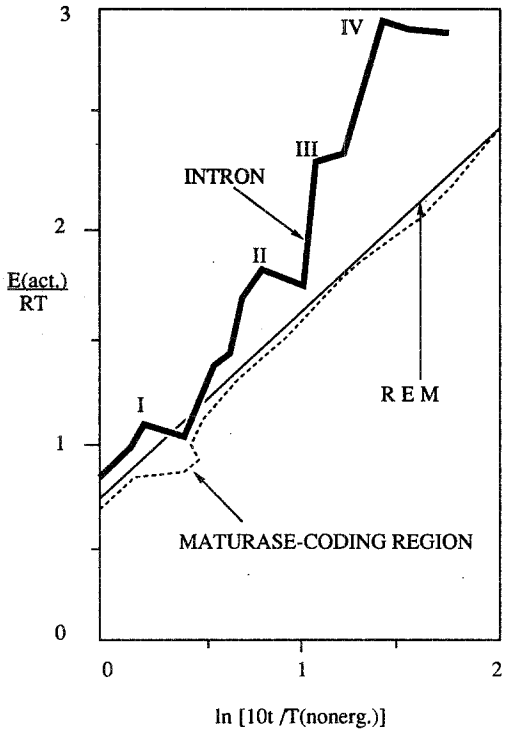

Fig. 4. Time-dependent behavior (thick line) of the activation energy E(act.) for the first four refolding events (I–IV) involved in the shaping of the catalytic site for the YC4 intron. The ordinates are proportional to the logarithm of a properly-scaled timespan of refolding events, while the abscissas give real time in a logarithmic scale. The scaling constant T(nonerg.) indicates the timespan for the whole clustering process which starts immediately after the core planar structure has been formed. The solid thin line corresponds to the REM and reflects the initial stages in the search for the global energy minimum for a random uncorrelated sequence. The dashed-line plot corresponds to the refolding events of an internal open reading frame (coding for the maturase).

formation has not been extracted from the Turner compilation. The pseudoknot motif must be included from a separate compilation, taking the enthalpies of stacking to be 66% of the Turner enthalpies, that is, of those which would correspond to a normal helix made up of the same base pairs as in the pseudoknotted interaction.

Our aim is to substantiate the conjecture that introns fold in a sense "better" than coding RNAs (exons) and we intend to provide a criterion altogether consistent with the fact that our folding algorithm is determined by the *kinetics* of admissible events. The example of the already studied YC4 intron proves to be inspiring in determining the sense in which introns might fold better: The initial configuration is specified as a planar core structure which results from an accumulation of successive refolding events during transcription. Thereafter, the computer code is designed so as to preclude further polymerization and monitor exclusively refolding events and long-range interactions leading to pseudoknotted catalytically-active structures (Davies *et al.*, 1982). These refolding events are responsible for clustering the core planar structure, thus shaping the catalytic site.

At this point, we are in a position to examine the plot of the timespan of each refolding event versus real time in a log-log scale, as displayed in Figure 4, and compare it with the plot for an internal coding frame. This is *guided* by the behavior of a random uncorrelated sequence; that is, regarding the relaxation of the structure of a random chain as a reference frame (Fernández and Shakhnovich, 1990). The choice of this reference frame is motivated by the fact that a random chain posesses the minimal degree of structure and that the relaxation of this structure follows a precise law.

The folding pathway which starts with the core planar structure and leads to the global free energy minimum in configuration space has very different activation-energy demands in the case of introns when compared with the activation-energy landscape (Fernández and Shakhnovich, 1990) for exons. The active intron structure is closer to the free energy minimum. In the case of exons, the relaxation (series of refolding events leading to the global minimum) appears to follow a random energy model (REM) (Fernández and Shakhnovich, 1990), where the kinetic barriers along the pathway grow logarithmically with time. This behavior is familiar from quenched disordered systems of the type encountered in condensed matter physics. Introns, on the other hand, have much higher kinetic barriers which do not grow logarithmically as we approach the free energy minimum. *It is as if, precisely because of the need to splice, pre-mRNA introns must fold "better" than exons.* Thus, since a random uncorrelated RNA chain refolds following the REM (Fernández and Shakhnovich, 1990), we may assess the structural organization adopting the refolding kinetics of a random copoly-

mer as the reference frame. This would lead us to the conclusion that *splicing introns have a higher degree of structural organization* (see Figure 4). This conclusion is validated by the fact that all catalytic RNA molecules belonging to a certain group known as group I (Davies *et al.*, 1982; Pleij *et al.*, 1988) appear to share the same structural motif as the case treated in this work. Since more than 80 species in that class have been identified so far, it appears that our conclusions are substantiated.

ACKNOWLEDGMENT

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

REFERENCES

Cech, T. R. (1988). *Gene*, **73**, 259.

Inoue, T., Sullivan, F. X., and Cech, T. R. (1985). *Cell*, **43**, 431.

Davies, R. W., Waring, R. B., Ray, J. A., Brown, T. A., and Scazzocchio, C. (1982). *Nature*, **300**, 719.

Fernández, A. (1990). *Physical Review Letters*, **64**, 2328.

Fernández, A. (1989). *European Journal of Biochemistry*, **182**, 161.

Anshelevich, V. V., Vologodskii, V. A., Lukashin, V., and Frank-Kamenetskii, M. D. (1984). *Biopolymers*, **23**, 39.

Pleij, C. W., Rietveld, K., and Bosch, L. (1985). *Nucleic Acids Research*, **13**, 1717.

Groebe, D. R., and Uhlenbeck, O. C. (1988). *Nucleic Acids Research*, **16**, 11725.

Turner, D. H., Sugimoto, N., and Freier, F. M. (1988). *Annual Reviews of Biophysics and Biophysical Chemistry*, **17**, 167.

Puglisi, J. D., Wyatt, J. R., and Tinoco, I. (1988). *Nature*, **331**, 283.

Lazowska, J., Jacq, C., and Slonimski, P. P. (1980). *Cell*, **22**, 333.

Burke, J. M. (1988). *Gene*, **73**, 273.

Fernández, A., and Shakhnovich, E. I. (1990). *Physical Review A-Rapid Communications*, **42**, 3657.