Stability domains of actin genes and genomic evolution

E. Carlon
Interdisciplinary Research Institute, Cité Scientifique, Boîte Postale 60069, F-59652 Villeneuve d'Ascq, France; Ecole Polytechnique Universitaire de Lille, Cité Scientifique, F-59655 Villeneuve d'Ascq, France; and Institute for Theoretical Physics, Katholieke Universiteit Leuven, Celestijnenlaan 200D, B-3000 Leuven, Belgium

A. Dkhissi, M. Lejard Malki, and R. Blossey
Interdisciplinary Research Institute, Cité Scientifique, Boîte Postale 60069, F-59652 Villeneuve d'Ascq, France
(Received 6 August 2007; published 21 November 2007)

In eukaryotic genes, the protein coding sequence is split into several fragments, the exons, separated by noncoding DNA stretches, the introns. Prokaryotes do not have introns in their genomes. We report calculations of the stability domains of actin genes for various organisms in the animal, plant, and fungi kingdoms. Actin genes have been chosen because they have been highly conserved during evolution. In these genes, all introns were removed so as to mimic ancient genes at the time of the early eukaryotic development, i.e., before intron insertion. Common stability boundaries are found in evolutionarily distant organisms, which implies that these boundaries date from the early origin of eukaryotes. In general, the boundaries correspond with intron positions in the actins of vertebrates and other animals, but not much for plants and fungi. The sharpest boundary is found in a locus where fungi, algae, and animals have introns in positions separated by one nucleotide only, which identifies a hot spot for insertion. These results suggest that some introns may have been incorporated into the genomes through a thermodynamically driven mechanism, in agreement with previous observations on human genes. They also suggest a different mechanism for intron insertion in plants and animals.

DOI: 10.1103/PhysRevE.76.051916 PACS number(s): 87.15.−v, 82.39.Pj

I. INTRODUCTION

Unlike their prokaryotic counterparts, the large majority of eukaryotic genes are split. The parts of the gene that carry the genetic code from which the proteins are synthesized, the exons, are interrupted by long stretches of "junk DNA," the introns [1]. Much is still uncertain about introns and in general about junk DNA. There is, however, a clear advantage for a gene of hosting introns: different mRNAs and henceforth different proteins can be synthesized from the same gene through a mechanism known as *alternative splicing* (see Fig. 1). In different tissues of a multicellular organism, the mRNAs are synthesized by placing the exons in a different order or by skipping some of them. This produces quite similar, but not identical, proteins. Alternative splicing is responsible for the appearance of slightly different proteins, say, in brain and in liver, both encoded by the same gene.

The origin of introns has triggered quite some debate in recent years. The discussion was polarized into two different viewpoints: the "introns early" [2] and the "introns late" [3] theories. The introns late viewpoint states that introns came "late" in evolution, say after the separation between the eukaryotic and prokaryotic kingdoms. Ancient genomes, like today's bacteria, had no introns. During evolution introns were inserted at some positions in the coding sequence of eukaryotes. Bacteria did not get introns in order to keep their genome short. According to the introns early perspective, introns were already present in ancient genomes. In these genomes minigenes were separated by junk DNA sequences. Complex genes appeared during evolution when the minigenes were assembled together.

Although the issue is not completely settled yet, there is widespread agreement about the fact that most introns were inserted late in the genome, except for a few which could have a very old origin [4]. There are questions remaining unanswered: Through what mechanism were introns inserted into the genes? Did they target some specific stretches of sequences or was their insertion a random process?

In a previous paper [5], we suggested that some introns may have targeted and got inserted in specific regions of the gene because of some physical stability properties of these regions. DNA is an inhomogeneous polymer: Sequences richer in CG nucleotides are more stable than AT-rich regions since CG pairs form three hydrogen bonds, while AT only two. Using a random set of 80 human genes, we found that there is a strong correlation between intron positions and stability boundaries [5], to be defined more precisely in the next section.

The aim of this paper is to investigate this issue further. We consider here a single gene, actin, and analyze its stabil-

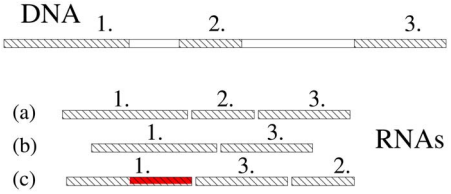

FIG. 1. (Color online) Through the mechanism of alternative splicing different RNAs and thus different proteins can be obtained from the same intron-containing DNA sequence. Typically, these different RNAs are synthesized in different tissues. In the example shown here, three different RNAs are formed from the same DNA sequence: exons are colored, while introns are white. (a) Exons are assembled following the same order as in the DNA sequence. (b) Exon 2 is skipped. (c) Exons 2 and 3 are in reversed order compared to the DNA sequence.

ity in animals, plants, and fungi. Although originating from a single gene in a common ancestor, actin genes have diversified during evolution. Several different actins are present in the genome of a given eukaryote. By analyzing the stability properties of genes belonging to a common family, we gain insight into mechanisms of intron insertion. We will show that common stability boundaries are found in actin sequences of species belonging to different kingdoms. This implies that the boundaries observed in this and in previous work [5] have a very remote origin, dating back to the development of early eukaryotes, and supports previous observations that stability boundaries may have influenced the insertion of at least some introns. An extensive discussion of the consequences of our findings is given in the final section of this paper.

II. THERMODYNAMIC STABILITY

When a double-helical DNA in solution is brought to a sufficiently high temperature, the two strands dissociate, or melt. DNA oligonucleotides of 20–30 base pairs melt at a single temperature. This temperature can be estimated using the nearest-neighbor model from which one computes Gibbs free energies, enthalpies, and entropies of melting. Quite some effort has been dedicated in recent years to an accurate determination of these thermodynamic parameters (see, e.g., [6] and references therein), due to the importance of DNA melting and the reverse transition, DNA hybridization, in many biotechnological processes. The melting temperature depends, as well as on the sequence composition, on salt concentration and the *p*H of the solution.

If the sequences are sufficiently long, DNA melting becomes a multistep process [7]. Regions of the sequence which are richer in GC will melt at higher temperatures compared to GC-poorer regions. The interesting quantities to calculate in this case are the multiple partial melting temperatures, and at the same time one needs to determine the regions of the sequence that melt at those temperatures. In order to perform this type of calculation, various statistical mechanical models have been developed [7,8]. The calculations presented in this paper are based on the MELTSIM algorithm [9], in which a DNA configuration is approximated by a sequence of noninteracting loops and helical segments according to the Poland-Scheraga model [10,11]. In this approach, each base pair is in one of two possible states either open ($\theta_i=0$) or closed ($\theta_i=1$), where θ defines the order parameter and i is an index running over all base pairs of a sequence ($i=1,2,\ldots,N$). In the MELTSIM algorithm, recursion relations [12] and an approximation for the closed loop entropy [13] allow a rapid computation of the opening and closing probability at any given temperature for chains of several thousand base pairs.

Computations based on the Poland-Scheraga model have been quite popular in recent years [14–20]. Yeramian *et al.* analyzed the genomes of *Saccharomyces cerevisiae* (yeast) [14] and of *Plasmodium falciparum* [15] and identified genes on the basis of thermodynamic signals obtained from the melting analysis. The effects of mismatches [16] and of disorder [19] on DNA melting have also been discussed. A recent study has produced the melting map of the whole human genome [21].

The other popular model for studies of the thermodynamics of the DNA is the Peyrard-Bishop model [8,22], which has attracted quite some attention in recent years [23–28]. This model is probably more accurate on shorter length scales, as a configuration is identified by the distances between complementary bases and not by a simple Boolean variable ($\theta=0,1$) as in the Poland-Scheraga picture. However, for the purposes of calculating stability properties which involve melting domains of about 100 base pairs the Poland-Scheraga model is good enough. Programs like MELTSIM have been fine tuned to fit experimental data [9,29]. Interestingly, the thermodynamic boundaries found in the MELTSIM approach in a previous paper [5] have also been found in an analysis of the Peyrard-Bishop model [26]. This shows that the properties discussed here are robust and model independent.

In this paper we have used the same set of thermodynamic parameters as in Ref. [30]. In order to estimate the thermal stability boundaries, we proceed as follows. Starting from sufficiently low temperatures at which the whole chain is in a helical state, we increase the temperature at a constant small step ΔT (=0.01 °C in the calculation). At each point the configuration of the chain is calculated and the boundaries between helix and coil regions recorded. To discriminate between a helical and a coiled region, we calculate the average value of θ_i at a given temperature and define the boundary as the point separating a $\theta>1/2$ domain from a $\theta<1/2$ domain.

Typical outputs of the calculations are shown in the graphs of Figs. 2–4 and 6. In these graphs the x axis is the temperature, while the y axis represents the position along the sequence. For each gene we considered only the coding sequence (CDS) with all introns removed. This is the so-called complementary DNA (cDNA), which is a double stranded copy of the mRNA and can be obtained from it in the laboratory through reverse transcription. For the purposes of inferring information on genome evolution, we can look at cDNA as an old gene before introns were inserted. In order to avoid boundary effects, i.e., dissociation dominated by the opening of forks at the edges, we have enclosed the sequences by two stretches of poly(G) of 200 nucleotides each [a poly(G) sequence is a stretch of DNA composed only of nucleotides G; in this case the sequence referred to is double stranded with one strand containing only G's while the other strand contains only C's]. These stretches have high melting temperature; hence they dissociate well beyond the melting temperatures of the CDS. The solid thick lines in Figs. 2–4 and 6 separate the coiled from the helical regions (to the right and to the left of the curve, respectively). Due to strong cooperativity [31], the DNA melts through few sharp transitions involving the dissociation of hundreds of base pairs simultaneously. Hence, only a few stability domains are found in the analysis of a sequence of 1000 base pairs.

III. ACTIN

Actin proteins play a central role in eukaryotes. Actin filaments constitute the cytoskeleton of all eukaryotic cells,

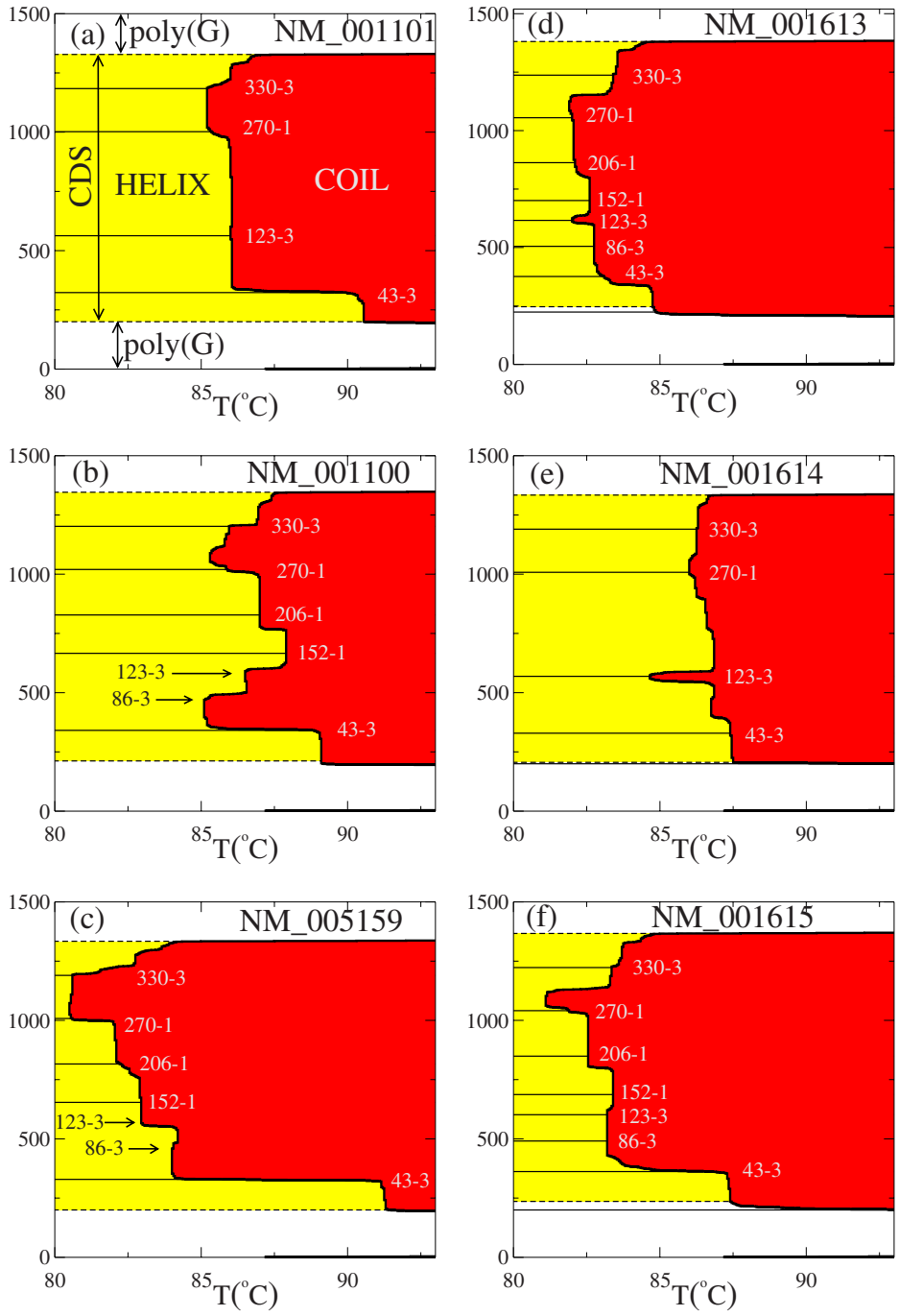

FIG. 2. (Color online) Melting domains for *H. sapiens* actins. GenBank entries: (a) NM_001101 (actin β, ACTB), (b) NM_001100 (actin α_1, skeletal muscle ACTA1), (c) NM_005159 (actin α, cardiac muscle, ACTC1) (d) NM_001613 (actin α_2, ACTA2) (e) NM_001614 (actin γ_1, ACTG1) (f) NM_001615 (actin γ_2, ACTG2). The sequences have high similarity to those of all other vertebrate actins, for which almost identical melting patterns are found. Hence only human actins are shown as representatives of those of all the vertebrates. In the plots the temperature is on the x axis and the sequence position on the y axis. The thick solid lines separate the low-temperature helix domain from the high-temperature coiled state. Horizontal dashed lines indicate the boundaries of the CDS and the solid lines the intron positions for the given sequence. Arrows point to the intron positions found in homologous sequences.

and are the site of interactions with many other proteins, as for instance motor proteins or actin-bundling proteins [1]. A mutation in a specific actin protein site may result in a change in its interactions with several proteins that bind near the mutated site. While the mutation can favor the interaction with one specific protein, it is likely to disrupt interactions with many other proteins. Hence, in order to maintain the multiple interactions with all its partners, actin proteins have been highly conserved during evolution. Obviously, there is lower conservation at the gene level compared with the conservation of the amino acid sequence for the corresponding protein, as the genetic code is degenerate and multiple

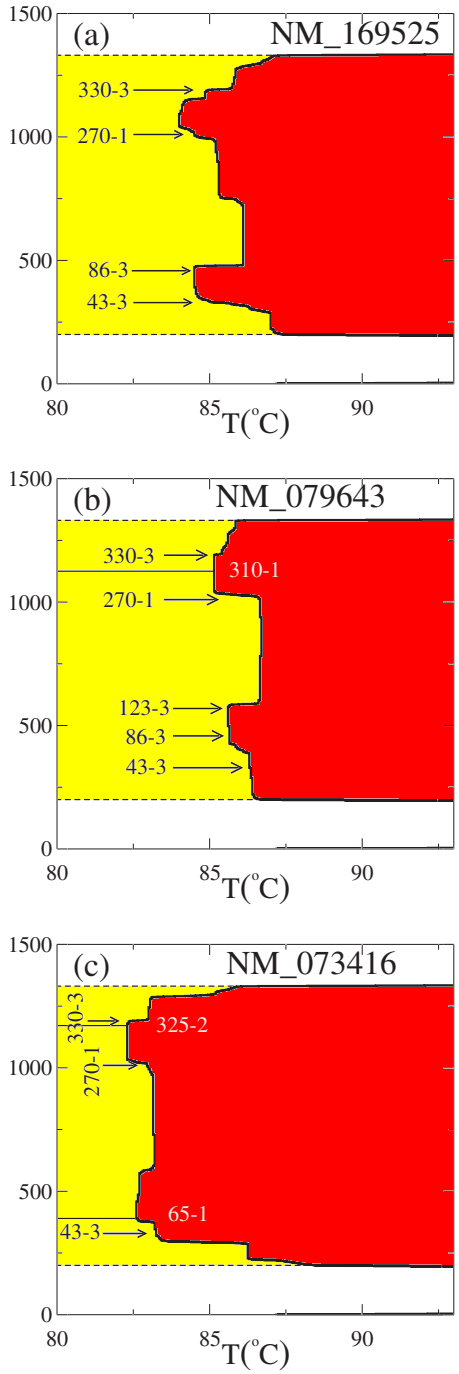

FIG. 3. (Color online) Melting domains for actin genes of *D. melanogaster*. (a),(b) and *C. elegans* (c). GenBank entries: (a) NM_169525 (Act87E), (b) NM_079643 (Act88F), and (c) NM_073416. The area descriptions are the same as in Fig. 2.

codons encode for the same amino acid. For actin, there is roughly 80% of sequence conservation between human and yeast (*S. cerevisiae*) genes, but 95% conservation of amino acids in the proteins [1].

Genomic evolution is believed to have occurred mainly through gene duplication and mutations [1]: At a given time, an error in the replication of DNA produces two copies of a gene, which are inherited by a daughter cell. These genes further evolve separately, accumulating different point mutations and thus diverging in time. As the process is repeated, one obtains from a single ancestor gene a family of closely related genes. In vertebrates, there are three classes of actins [1] known as α, β, and γ actins. The α actins are found in muscle cells, while the β and γ are found in nonmuscle cells. Plant actins also form a large family with even more genes than in vertebrates. For instance, more than ten different actin genes have been identified in the genome of the plant *Arabidopsis thaliana*.

A. Intron positions

Table I compares the intron positions of vertebrates and land plants. A more complete table which contains 56 different intron positions for actins of different organisms can be found in Ref. [32]. In total there are seven intron positions for vertebrate actins. These positions are labeled, following the notation of Ref. [32], by two numbers. The first number refers to the codon in the sequence and the second one (between 1 and 3) indicates where the intron is inserted in the codon. A 3 signifies that the intron is inserted after the third nucleotide of the codon; hence the intron does not break the codon. The codon numbers are given with respect to a reference sequence, which is the α actin of vertebrates. Although plants have more actin isoforms than vertebrates, somewhat surprisingly they have only three intron positions, one of which (152-1) is in common with the vertebrate lineage.

B. Melting domains for animal actins

We start with the description of melting domains in animal actins. Although Fig. 2 shows exclusively human actins, we found very similar melting profiles also in α, β, and γ actins of other vertebrates: *Canis familiaris* (dog), *Bos taurus* (cow), *Danio rerio* (zebrafish), *Gallus gallus* (chicken) etc. Hence, the conclusions drawn from the analysis of Fig. 2 are probably valid for all vertebrates.

Figure 2(a) shows the melting behavior of the human actin β (GenBank entry NM_001101). This sequence has four introns at positions 43-3, 123-3, 270-1, and 330-3 (see Table I). These positions are indicated by horizontal lines in Fig. 2(a). In this sequence melting is a three-state process. First, at around 85 °C, the exon bounded between introns 270-1 and 330-3 melts. Next, the whole CDS sequence melts except for a short fragment bounded by the intron at 43-3, which then melts only beyond 90 °C. There is a remarkable correspondence between the 43-3 position and a sharp stability boundary. Also positions 270-1 and 330-3 show a similar, although weaker, correspondence. This β actin sequence has already been analyzed in Ref. [5] (see Fig. 2 in [5]). In that analysis the correspondence with the intron 330-3 was missed because different boundary conditions were used. In this work the CDS is embedded between two poly(G) stretches, so that melting inside the sequence is always through the formation of loops bounded between two helical regions. In Ref. [5], some parts of the untranslated regions bounding the CDS were included in the analysis. As no stable boundary helical regions were included, part of the melting in Ref. [5] occurred through fork openings from the boundaries. The inclusion of untranslated regions, i.e., of the

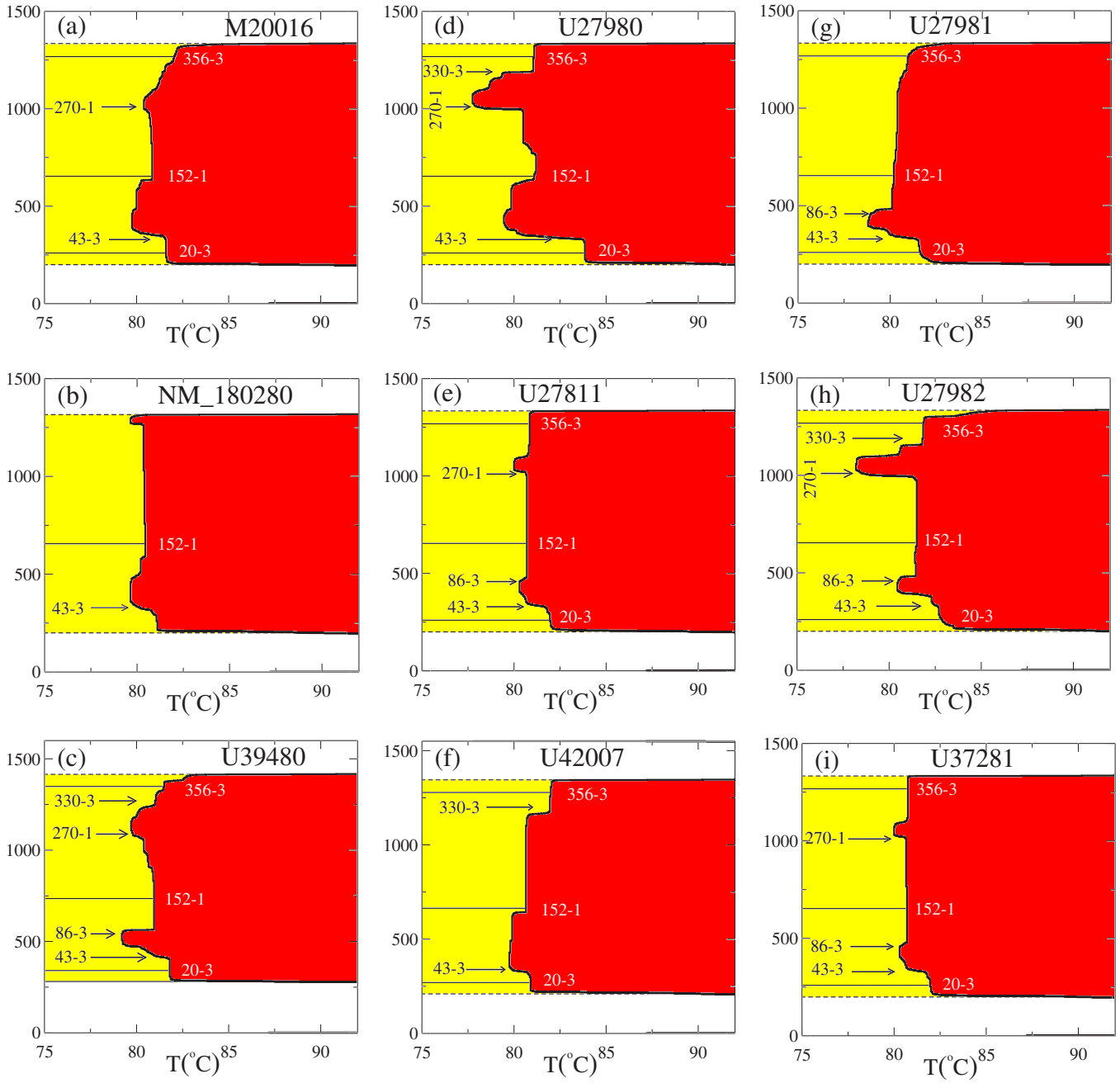

FIG. 4. (Color online) Melting domains for actin sequences of *A. thaliana*. The area descriptions are the same as in Fig. 2. GenBank entries: (a) M20016 (AAc1), (b) NM_180280 (ACT2), (c) U39480 (ACT3), (d) U27980 (ACT4), (e) U27811 (ACT7), (f) U42007 (ACT8), (g) U27981 (ACT11), (h) U27982 (ACT12), and (i) U37281 (actin 2).

original genomic neighborhood in the analysis, is probably not an optimal choice as these regions are poorly conserved during evolution. As the aim is to look for signals from ancient genomes, it is better to embed the CDS between two poly(G) stretches. In this way all the sequences analyzed are treated on equal footing.

Figures 2(b)–2(f) show other melting domains for homologous vertebrate actin genes. These sequences have four [Fig. 2(e)], five [Figs. 2(b) and 2(c)], and seven [Figs. 2(d) and 2(f)] intron positions. There are three intron positions which are in common in the vertebrate actins: 43-3, 270-1, and 330-3. These are also those for which a correspondence with the thermal boundary was found in the β gene of Fig. 2(a). The correspondence between thermodynamic boundaries and the intron positions 43-3 and 270-1 is also observed in the three other sequences of Figs. 2(b), 2(c), and 2(f). Note the very sharp signal from the 43-3 intron in Fig. 2(c). The intron at position 330-3 shows a good correspondence with stability boundaries for the sequences in Figs. 2(b) and 2(c). A much weaker, but noticeable, correspondence is with the intron at position 206-1 in the sequences in Figs. 2(c) and 2(d). The stability boundary is slightly shifted from the 206-1 in the sequences in Figs. 2(d) and 2(f). The correspondence

TABLE I. Table of intron positions in actin genes for vertebrates and green plants. The label refers to the codon position following the table of Ref. [32]. The three positions in bold are those for which a stability boundary was found.

	20-3	**43-3**	86-3	123-3	152-1	206-1	**270-1**	**330-3**	356-3
Vertebrates									
α_1		x			x	x	x	x	
β		x		x			x	x	
γ_1		x		x			x	x	
α_2		x	x	x	x	x	x	x	
γ_2		x	x	x	x	x	x	x	
Green plants	x				x				x

between intron positions and thermodynamic boundaries is absent in the sequence of Fig. 2(d).

Another interesting feature of vertebrate actins can be seen in Fig. 2(b). This sequence is the actin α_1, which hosts five introns in its coding region. These are marked by horizontal lines. The two remaining of the total seven intron positions of vertebrate actins, the 86-3 and 123-3, are indicated by horizontal lines. As can be seen from Fig. 2(b), these two positions correspond to stability boundaries. The correspondence of a stability boundary with the 123-3 is also visible in Figs. 2(c) and 2(e). In the latter example the 123-3 is a nucleation site for a small loop. The sequences of Figs. 2(d)–2(f) show a much weaker correspondence between intron positions and stability boundaries.

Figure 3 shows the melting curves for *Drosophila melanogaster* [Figs. 3(a) and 3(b)], the fruit fly, and *Caenorhabditis elegans* [Fig. 3(c)], a worm. The *Drosophila* actins have at most one intron in the coding region in either 15-1 or 310-1. These positions differ from the vertebrate positions discussed so far. The sequence shown in Fig. 3(a) has no introns. The melting analysis, however, reveals a few stability boundaries close to the positions 43-3, 86-3, 270-1, and 330-3, which are the intron positions of vertebrate actins. The 43-3 and 86-3 are particularly sharp. The next *Drosophila* sequence [Fig. 3(b)] with one intron at position 310-1 shows a stability boundary close to 270-1 and a weaker one close to 330-3. Compared to the case in Fig. 3(a), in this sequence the signals from 43-3 and 86-3 have been lost. However, a sharp boundary has appeared close to the vertebrate intron 123-3. The *C. elegans* sequence of Fig. 3(c) has two introns at "new" positions 65-1 and 325-2. As in the previous examples, one observes boundaries close to the 43-3, 123-3, 270-1, and 330-3 positions.

C. Melting domains for plant actins

Figure 4 shows the melting domains for actin genes of the green plant *Arabidopsis thaliana*. The intron positions of actin sequences of higher plants are highly conserved (see Table I), which indicates that these introns date back to the early evolution of land plants. In three out of the nine *Arabidopsis* sequences shown Figs. 4(a), 4(d), and 4(f) we find a correspondence of a thermal boundary and an intron at 152-1. This is the intron which is in common with vertebrates (see Table I). As in the *Drosophila* and *C. elegans* sequences (Fig. 3) in general stability boundaries tend to be found at vertebrate positions 43-3, 86-3, 270-1, and 330-3. In a few cases the correspondence is very striking, as in Fig. 4(d).

In order to corroborate these findings we extended the analysis to other plants. We considered 12 additional actins from *Nicotiana tabacum* (tobacco, GenBank X63603), *Oryza sativa* (rice, GenBank X15862, X15863, X15864, X15865), *Glycine max* (soyabean, GenBank J01298, V00450), *Solanum tuberosum* (potato, GenBank X55749, X55750, X55751, X55752), and *Striga asiatica* (GenBank U68461, U68462). With the nine sequences from *A. thaliana* we have in total 21 plant actin genes. For each sequence the melting curves were calculated and then averaged. The result is shown in Fig. 5. In the graph the x and y axes are reversed compared to Figs. 4. The x axis is the codon position, and the temperature, now in the y axis, is ordered as increasing from top to bottom. As reference, four of the most commonly found intron positions for actin genes are indicated as vertical lines. The averaging introduces some smoothing, but it confirms the existence of a sharp stability boundary close to the 43-3 position. Two weaker boundaries are found close to the positions 86-3 and 270-1.

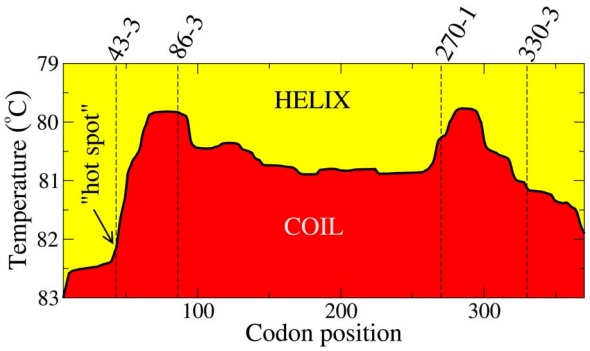

FIG. 5. (Color online) Average melting curves from 21 green plant actin genes. The axes of the diagram are swapped compared to those in Figs. 2–4. The horizontal axis is given in codon position; only the coding sequence is shown. The four vertical lines denote the four major intron positions found to correlate with stability boundaries in vertebrates.

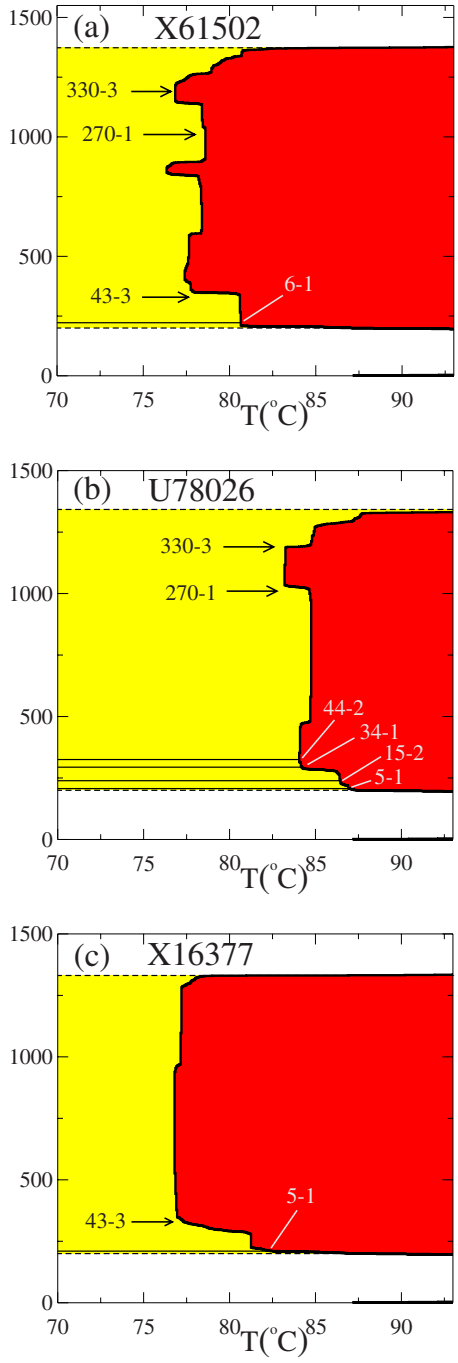

FIG. 6. (Color online) Melting domains for actin sequences of fungi. The area descriptions are the same as in Fig. 2. (a) *S. cerevisiae* (yeast), (b) *N. crassa,* and (c) *C. albicans*. GenBank entries: (a) X61502 (Act2), (b) U78026, and (c) X16377 (act1).

D. Melting domains for fungi actins

To conclude the analysis of the stability behavior of actin genes, we consider now fungi. Figure 6 shows the melting curves for the budding yeast *Saccharomyces cerevisiae* [Fig. 6(a)], for *Neurospora crassa* [Fig. 6(b)], and for *Candida albicans* [Fig. 6(c)]. In general the number of introns and their positions are highly variable in fungi actin genes: their number varies from zero to seven and the positions are most likely concentrated in the region before the 50th codon. This can be seen also in the sequences of Fig. 6, two of which have one intron and one has four. All introns are found before the codon 45. The melting behavior shown in Fig. 6 resembles that of the previous cases: A sharp stability boundary close to the 43-3 position and some weaker ones appearing close to positions 270-1 and 330-3 in the case of Fig. 6(b). This correlation is absent in the case of Fig. 6(a).

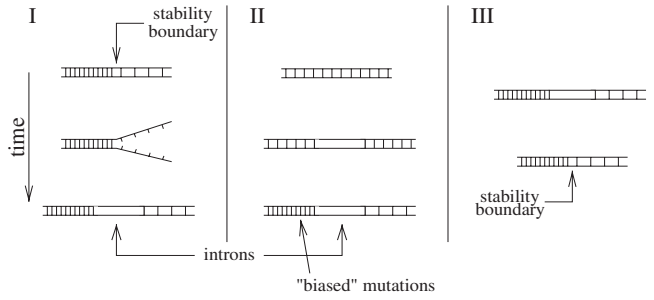

FIG. 7. Possible scenarios of gene evolution. (I),(II) introns late and (III) introns early perspectives. The scheme (I) is the insertion driven by thermodynamics, supported by the analysis of stability regions reported in this paper. The scheme (II) suggests the appearance of a thermal boundary after intron insertion, a scenario not supported by the results presented in this paper. Finally, the scheme (III) takes into account the introns early perspective, in which the ancient exons had already different stability properties.

E. A hot spot for intron insertion

In Ref. [32] the full table of intron positions in actin genes gives 56 different positions. There is an interesting remark concerning the position 43-3, which is shared by all vertebrate actins. Position 43-3 is relatively common in animals, but it is also the sole intron in the actin gene of the red alga *Chondrus crispus* [32]. Shifted by a single nucleotide at position 44-1, there is an intron in the alga *Cyanophora paradoxa* [32]. An intron at position 44-2 is present in the single-copy actin genes of the fungi *Thermomyces lanuginosa*, *Aspergillus niger*, *N. crassa* [see Fig. 6(b)], and *Trichoderma reesei* [32]. The only other positions with introns separated by a single nucleotide are at 34-1 and 34-2, which, however, are only found in some fungi. Hence the 43-3 is a unique site in the actin genes, which we can refer to as a "hot spot" for intron insertion.

IV. DISCUSSION

DNA sequences, which are hundreds of base pairs long, tend to melt through a series of separate temperature steps. Each step consists of the melting of a region of a few hundred base pairs. By following the melting process over a wide temperature interval, one can thus identify separate melting domains, i.e., parts of the sequence which dissociate at different temperatures. The domain boundaries are points in which the sequence tends to form in a relatively wide temperature interval a stable Y conformation separating a double helix from a coiled region (see Fig. 7).

A previous study [5] of about 80 human genes from which introns are removed and exons linked together revealed that stability domain boundaries tend to be localized at the ends of exons. This correspondence was found for about 35% of the exons analyzed. The correlation was found to be stronger for a class of so-called housekeeping genes, i.e., those genes involved in the basic cellular processes. These genes are expressed in all tissues and have been more conserved during evolution. Actin is in fact an example of a housekeeping gene. If one accepts an "introns late" viewpoint, the correlation between intron positions and stability boundaries suggests that some introns were inserted into genes at the ends of the melting domains in a process driven by thermodynamics. Such a process is illustrated in Fig. 7 (scheme I): an intronless fragment of a gene has naturally parts that are richer and poorer in CG nucleotides. When the two strands partially separate they may form a Y configuration, the end of a less stable domain and the beginning of a more stable one. Introns may have targeted these fork locations.

Another possibility that may have explained the correlation between thermodynamic boundaries and intron positions observed in Ref. [5] is schematically shown in Fig. 7 (scheme II). Originally the insertion site does not possess a thermodynamic boundary, so the intron is inserted through a process that does not depend on thermodynamics. Once the insertion has taken place and the two exons are separated by an intron stretch, mutations may have biased the CG content on the two exons so that their thermodynamic boundary originated after the intron insertion. However, this scheme is at odds with the results presented in this paper. We have indeed shown that boundaries in conserved positions are found in actin family genes where no introns are present close to those positions. For instance, in many actins of plants, fungi, and animals, there is a sharp stability boundary at the position 43-3 in sequences that have no intron at that position. Hence, being found in plants, animals, and fungi sequences, the stability boundary at 43-3 is rather a property of an intronless ancestor actin gene. The same is true for stability boundaries found in other intronless positions like, for instance, 86-3 and 270-1.

We further speculate on the "introns early" perspective, i.e., the possibility that introns were already present in early genomes and were selectively lost by some species. Our findings then imply that these introns would have separated early exons (or minigenes as they are also referred to) with different stability properties, as shown in the scheme of Fig. 7 (scheme III). Although possible, this scenario seems to be in contradiction with most of the recent phylogenetics-based studies, which favor an introns late theory.

In conclusion, our work supports the mechanism given in Fig. 7 (scheme I), i.e., a thermodynamically driven intron insertion. This does not necessarily mean that the actual insertion process took place through an equilibrium transition with a temperature rise to 80 °C. First of all the melting temperature depends also on other salt concentrations and the *p*H of the environment. Moreover, the boundaries found in the melting analysis should manifest themselves also under nonequilibrium conditions. Y configurations like those shown in Fig. 7 can also be generated by mechanical unzipping of DNA [33]. Quite remarkable is the fact that the sharpest boundary in actin genes (43-3) is also the locus in which intron insertion has been the most active in evolutionarily distant organisms. As we have pointed out, introns have also been found at positions 44-1 and 44-2. This fact is in agreement with the idea of an insertion driven by thermodynamics. As the boundary is particularly sharp, the mechanism of Fig. 7 (scheme I) could have occurred independently in three different families of actin genes.

The correlation between stability boundaries and intron positions is particularly sharp in several sequences analyzed, but in a few cases is absent. We believe that this is due to mutations having erased the correlation from the ancestral gene sequence. Although actin is highly conserved as a protein, there is no selective pressure against synonymous mutations which do not modify the amino acid sequence. Such mutations are known to have occurred at a roughly constant rate in all genes of a given organism [1]. Two genes of the same family in the same organism have evolved separately, and mutations may have accumulated at higher or lower rates in different parts of the sequence: In some genes the mutations may have erased the ancient stability boundaries.

The problem of intron evolution has been widely debated in the biological literature (for a recent review of the state of the art, see Ref. [34]). Even within the introns late perspective, there is no general consensus on the mechanism of insertion and several possibilites have been analyzed. For instance, Ref. [34] reports five different models of intron insertion. Most of these models in general discuss the mechanism of insertion without suggesting in which position of the sequence the insertion would have occurred. One exception is the protosplice site model [35], which suggests a bias toward a specific insertion sequence (C/A)AG(G/A) (here C/A denotes a site that can posses either a nucleotide C or A), referred to as the protosplice sites. This insertion would have led to a structure (C/A)AG-intron-(G/A). The protosplice model and other models for intron insertions are only partially supported by the analysis of genomic data [34].

Unfortunately, the genomes nowadays investigated have been heavily reshaped by hundreds of millions of years of evolution and are quite different from genomes of early eukaryotes. Hence the answer to the question of intron origin is not an easy one. Indeed, although introns were discovered 30 years ago, there is still an open debate on this issue. Moreover, evolution may have taken place through complex and diversified pathways so it is not unlikely that different mechanisms of insertion have coexisted. Certainly, the possibility that also the physical and thermodynamical stability of the double helix has played a role offers different insights and stimulates further research in this field.

[1] B. Alberts, A. Johnson, J. Lewis, K. Roberts, and P. Walter, *Molecular Biology of the Cell* (Garland Science, New York, 2002).
[2] W. Gilbert, Nature (London) **271**, 501 (1978).
[3] T. Cavalier-Smith, Nature (London) **315**, 283 (1985).
[4] S. W. Roy, A. Fedorov, and W. Gilbert, Proc. Natl. Acad. Sci. U.S.A. **99**, 15513 (2002).
[5] E. Carlon, M. L. Malki, and R. Blossey, Phys. Rev. Lett. **94**, 178101 (2005).
[6] J. SantaLucia, Jr., Proc. Natl. Acad. Sci. U.S.A. **95**, 1460 (1998).
[7] R. M. Wartell and A. S. Benight, Phys. Rep. **126**, 67 (1985).
[8] M. Peyrard and A. R. Bishop, Phys. Rev. Lett. **62**, 2755 (1989).
[9] K. A. Marx, I. Q. Assil, J. W. Bizzaro, and R. D. Blake, J. Biomol. Struct. Dyn. **16**, 329 (1998).
[10] D. Poland and H. A. Scheraga, J. Chem. Phys. **45**, 1456 (1966).
[11] D. Poland and H. A. Scheraga, *Theory of Helix-Coil Transitions in Biopolymers* (Academic Press, New York, 1970).
[12] D. Poland, Biopolymers **13**, 1859 (1974).
[13] M. Fixman and J. J. Freire, Biopolymers **16**, 2693 (1977).
[14] E. Yeramian, Gene **255**, 139 (2000).
[15] E. Yeramian, S. Bonnefoy, and G. Langsley, Bioinformatics **18**, 190 (2002).
[16] T. Garel and H. Orland, e-print arXiv:cond-mat/0304080.
[17] E. Tøstesen, Phys. Rev. E **71**, 061922 (2005).
[18] R. Metzler and T. Ambjörnsson, J. Biol. Phys. **31**, 339 (2005).
[19] B. Coluzzi and E. Yeramian, Philos. Mag. **87**, 517 (2007).
[20] R. Everaers, S. Kumar, and C. Simm, Phys. Rev. E **75**, 041918 (2007).
[21] F. Liu *et al.*, PLos. Computational Biol. **3**, e93 (2007).
[22] M. Peyrard, Nat. Phys. **2**, 13 (2006).
[23] A. Campa and A. Giansanti, Phys. Rev. E **58**, 3585 (1998).
[24] S. Cocco and R. Monasson, Phys. Rev. Lett. **83**, 5178 (1999).
[25] M. Barbi, S. Lepri, M. Peyrard, and N. Theodorakopoulos, Phys. Rev. E **68**, 061909 (2003).
[26] M. Joyeux and S. Buyukdagli, Phys. Rev. E **72**, 051902 (2005).
[27] T. Michoel and Y. Van de Peer, Phys. Rev. E **73**, 011908 (2006).
[28] G. Weber, N. Haslam, N. Whiteford, A. Pragel-Bennett, J. W. Essex, and C. Neylon, Nat. Phys. **2**, 55 (2006).
[29] J. W. Bizzaro, K. H. Marx, and R. D. Blake, in *Materials Science of the Cell*, edited by B. Mulder, V. Vogel, and C. Schmidt, MRS Symposia Proceedings No. 489 (Materials Research Society, Pittsburgh, 1998), p. 73.
[30] R. Blossey and E. Carlon, Phys. Rev. E **68**, 061911 (2003).
[31] R. D. Blake, Biopolymers **26**, 1063 (1987).
[32] D. Bhattacharya and K. Weber, Curr. Genet. **31**, 439 (1997).
[33] J. D. Weeks, J. B. Lucks, Y. Kafri, C. Danilowicz, D. R. Nelson, and M. Prentiss, Biophys. J. **88**, 2752 (2005).
[34] S. W. Roy and W. Gilbert, Nat. Rev. Genet. **7**, 211 (2006).
[35] N. J. Dibb and A. J. Newman, EMBO J. **8**, 2015 (1989).