Exons, Introns, and DNA Thermodynamics

Enrico Carlon,[1] Mehdi Lejard Malki,[1,2] and Ralf Blossey[1]

[1]*Interdisciplinary Research Institute c/o IEMN, Cité Scientifique, BP 60069, F-59652 Villeneuve d'Ascq, France*
[2]*Ecole Nationale Supérieure de Physique de Strasbourg (ENSPS), Parc d'innovation, BP 10413, F-67412 Illkirch, France*

(Received 5 October 2004; published 2 May 2005)

The genes of eukaryotes are characterized by protein coding fragments, the exons, interrupted by introns, i.e., stretches of DNA which do not carry useful information for protein synthesis. We have analyzed the melting behavior of randomly selected human cDNA sequences obtained from genomic DNA by removing all introns. A clear correspondence is observed between exons and melting domains. This finding may provide new insights into the physical mechanisms underlying the evolution of genes.

DOI: 10.1103/PhysRevLett.94.178101 PACS numbers: 87.15.−v, 05.70.Fh, 87.14.Gg

One of the most striking aspects of the human genome is the presence of long stretches of DNA with no apparent (or known) significance [1]. This is what biologists refer to as *junk* DNA, and it comprises the majority of our DNA. In the human genome (and that of other higher eukaryotes) not only are the genes very sparse, but most of them are interrupted by sequences, the introns, which are noncoding; i.e., they do not carry information for protein synthesis [1]. During transcription, introns are therefore removed from the messenger RNA (mRNA), which is assembled only from the expressed parts of the gene, the exons. In the human genome, introns are on average 10 times longer than exons and thus constitute the majority of the gene. Procaryotes (such as bacteria) instead have a very compact genome without introns [1].

The discovery of introns in 1977 triggered a debate around their significance and origin, which lead to the formulation of the ''introns-early'' [2–4] and the ''introns-late'' theories [5–7]. According to the introns-early viewpoint, the introns appeared at the origin of life and the exons were small ancient genes. The bacteria then lost the introns due to selective pressure in order to keep their genome short. The introns-late theory instead claims that introns must have appeared much later, i.e., during the early eukaryotic evolution. A consensus between these opposing views has meanwhile been reached in recent years. The analysis of an increasing number of genes showed that most of the introns have a ''recent'' origin, although a few are still believed to be very old [8]. The mechanism by which introns were included into the genome is, however, still poorly understood (for a recent discussion, see, e.g., Ref. [9]).

In this Letter, we present the results of a study of the physical properties of human DNA sequences which points to a possible pathway leading to intron insertion in genes. By means of a statistical mechanics approach, we analyze the thermodynamic stability (''melting'') of DNA sequences obtained by assembling the exons together. This is known as complementary DNA (cDNA) and can be obtained in the laboratory by reverse transcription of mRNA. As illustrated in Fig. 1, cDNA is characterized by exon-exon boundaries and the boundaries between the coding sequence (CDS) and the untranslated region (UTR). If introns were inserted ''recently'' into the genome, then the cDNA roughly resembles an ancient gene, apart from the mutations that have occurred since the insertion of the first introns (see below). We find that the exon-exon boundaries in cDNA sequences are strongly correlated with their melting domains.

DNA melting is the process by which the double-stranded molecule in solution dissociates into two separate strands by an increase of temperature [10]. Fragments which are longer than 1000 bp (base pairs) dissociate through a multistep process in which different parts of the chain melt at different temperatures. These ''melting domains'' are typically a few hundreds of nucleotides long. The thermodynamics of the DNA melting process has been investigated both experimentally [10] and by means of numerical calculations based on the statistical mechanics of the dissociation process [11,12]. The latter approach allows one to calculate θ_i, the probability that the ith

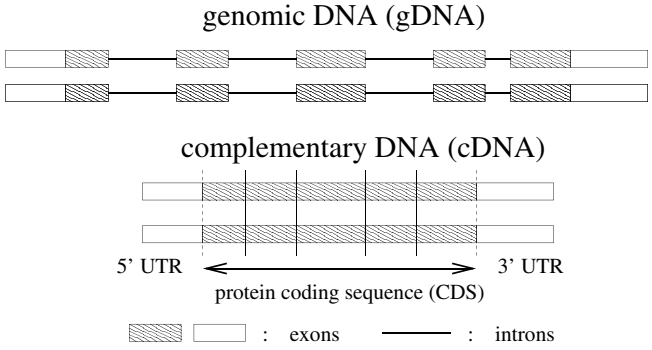

FIG. 1. Schematic view of genomic and cDNA. *Exons* are the segments of the gene transcribed into mRNA, while *introns* are spliced out. The cDNA, which is a reverse-transcription from the single-stranded mRNA, contains no introns. It is characterized by exon-exon boundaries (vertical solid lines) and boundaries between the protein CDS and the UTRs (vertical dashed lines). The part of the exons shown in black contains the protein coding sequence. The $3'$ and $5'$ ends refer to those of the single-stranded mRNA.

base pair is bound at a temperature T. The total fraction of bound base pairs is then $\theta = \sum_i \theta_i / N$, where N is the number of nucleotide pairs in the molecule. The multistep nature of the melting transition can be seen in a plot of θ vs T. This quantity vanishes while increasing the temperature through a series of jumps which correspond to sharp peaks in a plot of $-d\theta/dT$ vs T, the differential melting curve. The parameter θ can also be measured by UV absorption of DNA in a solution [10]. Typically, statistical mechanics programs reproduce the experimental results quite well [13].

In Fig. 2 we plot the melting curve $-Nd\theta/dT$, as obtained by a statistical mechanical calculation [13], for the human β-actin cDNA, where N is the total length of the sequence ($N = 1792$ in this case). We used the same stacking energies and loop entropic parameters as in Ref. [14]. The salt concentration was fixed at 0.05 M. The three main melting peaks of Fig. 2 indicate three sharp subtransitions which characterize the melting of the sequence. The evolution of the average configuration of the sequence as a function of T can be read off from the vertical bars, which denote, at the given temperatures, the regions which are more likely to be dissociated, i.e., where $\theta_i < 1/2$. For instance, the bar shown at $T \approx 85\,^{\circ}\mathrm{C}$ indicates that the region with $i \gtrsim 850$ is dissociated, while that with $i \lesssim 850$ is in a helical state. These melting domains are plotted for temperatures between melting peaks, so that, by comparing the configurations at temperatures below and above each peak, one can visualize the regions of the sequence involved in the multistep melting. We refer to the nucleotides separating two neighboring regions of the sequence with $\theta < 1/2$ and $\theta > 1/2$ as the *thermodynamic boundaries*. In Fig. 2 exon-exon boundaries are indicated as horizontal solid lines, while the boundaries between the CDS and the UTRs are shown as dashed lines. The numbers on the left vertical axis, located at the exon-exon boundaries, show the length on the introns in the genomic DNA.

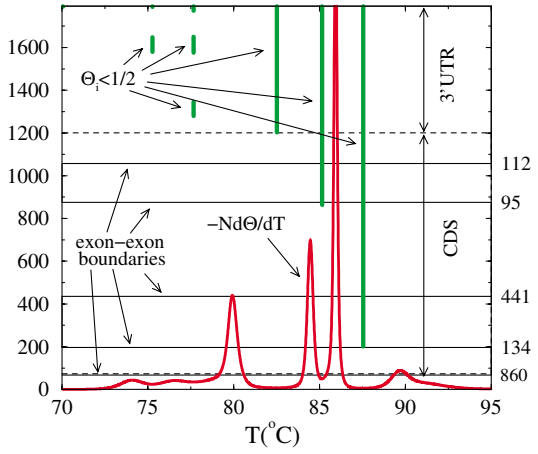

FIG. 2 (color online). Differential melting curve and melting domains for the human β-actin cDNA (NCBI entry code NM_001101). Horizontal axis, temperature; vertical axis, $-Nd\theta/dT$ and sequence length. Vertical bars in the graph indicate the regions along the chain for which $\theta_i < 1/2$. Horizontal solid lines are exon-exon boundaries and dashed lines are boundaries between the protein CDS and the UTRs. A remarkable overlap between the genomic and thermodynamic domains is observed.

In the example of Fig. 2, the melting process starts with the opening of small loops in the 3′UTR, while the first sharp peak at 80 °C is the dissociation of the whole 3′UTR. The next peak at about 84 °C is due to the melting of exons 5 and 6 (numbering them from the 5′ region), while the melting of exons 3 and 4 occurs at higher temperature ($\approx$86 °C). A remarkable overlap between the locations of the thermodynamic and genomic domains is observed.

An equally striking correspondence is found in most of the human cDNA sequences we investigated. Figure 3 shows the melting curve for the cDNA of the cyclin dependent kinase (CDK4). Occasionally, we found some discrepancies, as can be seen, e.g., in Fig. 4, which shows the cDNA for the human HAADH gene. The inset shows an enlargement of the region for the temperature interval of 80 °C–85 °C. Note that the thermodynamic boundary indicated by the arrows splits the third exon of the sequence into roughly two equal parts.

Long cDNA sequences ($\gtrsim$3000 bp) may have a very complex melting curve with many overlapping peaks. In order to have a better criterion for the definition of thermodynamic boundaries, we have performed a temperature scan from 60 °C to 100 °C and calculated the boundaries separating the $\theta_i < 1/2$ to the $\theta_i > 1/2$ regions at a fixed interval $\Delta T = 0.01\,^{\circ}\mathrm{C}$. We have then derived a histogram h_i over all base pairs i as follows: if i is found to be a thermodynamic boundary between two temperatures T_1 and $T_2 > T_1$, the contribution to the histogram is $h_i = (T_2 - T_1)/\Delta T$.

Figure 5 shows the histogram h_i as function of i for the human ILF2 cDNA (thick line). The solid and dashed thin vertical lines denote the exon-exon and CDS-UTR boundaries, respectively. The histogram is characterized by a few

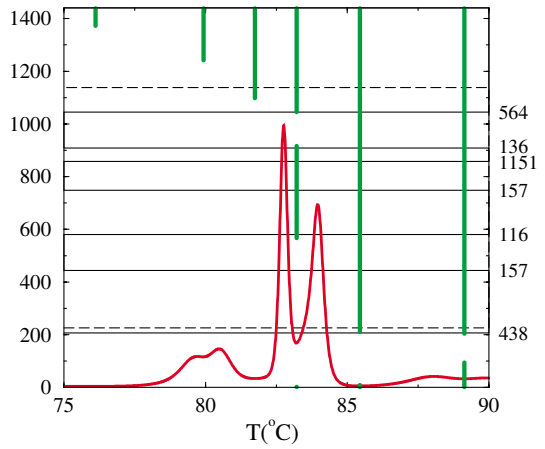

FIG. 3 (color online). As in Fig. 2, but for the human CDK4 cDNA (NCBI entry code NM_000075).

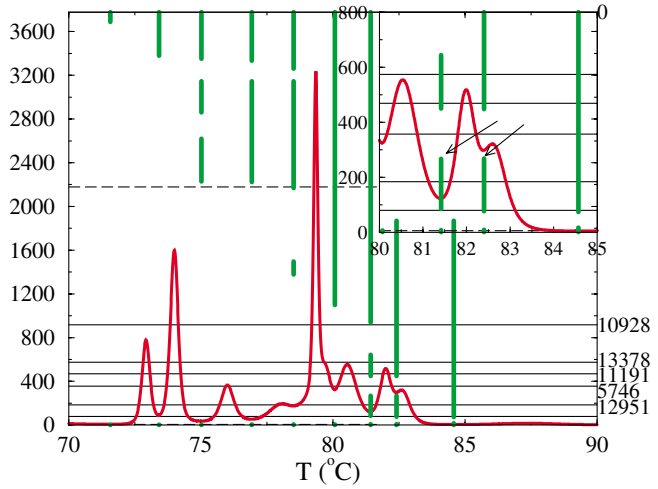

FIG. 4 (color online). As in Fig. 2, but for the EHHADH human gene (NCBI entry code NM_001966).

strong thermodynamic boundaries, well above the noise level (as typically observed in all cases). The advantage of h_i, with respect to a plot of the differential melting curves, is that boundaries appear more clearly also for long sequences, and their stabilities can be quantified from the height of h_i. The ILF2 cDNA melting of Fig. 5 is yet another example of the good correspondence between thermodynamics and genomic features. The exon-exon boundaries "detected" by thermodynamics are shown as vertical arrows in Fig. 5.

The correspondence between melting domains and genomic features has also been explored recently in studies of lower eukaryotes, such as *S. Cerevisiae* [15], *P. Falciparum* [16], or *D. Discoideum* [17]. These studies focused on *genomic* DNA and, in particular, on the correspondence between exon-intron and thermodynamic boundaries. This correspondence allowed Yeramian *et al.* to locate genes in the *P. Falciparum* [18] and *D. Melanogaster* [19] genome. Exon-intron boundaries may be difficult to detect in higher

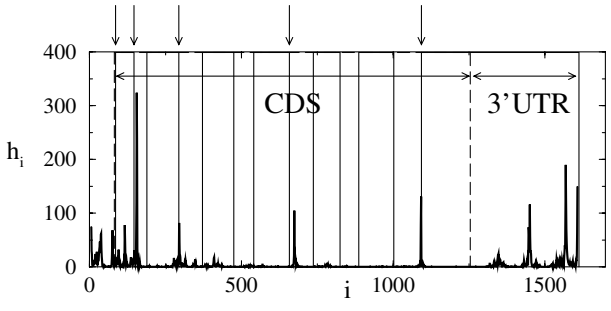

FIG. 5. Histogram of the thermodynamic boundaries (thick lines) of the human cDNA encoding for the interleukin enhancer binding factor 2 (ILF2) protein with gene bank entry code NM_004515. The thin vertical lines denote exon-exon (solid lines) and CDS-UTR (dashed lines) boundaries. The arrows indicate the exon-exon boundaries detected by the melting analysis.

eukaryotes by melting analysis as introns tend to be very long. We have illustrated this in Fig. 6, which shows the melting histogram for the human ribosomal protein L11 cDNA sequence (a) and for the cDNA in which the second intron has been inserted. In the cDNA the strongest peaks in the histogram are correlated with exon-exon boundaries; this holds for boundaries 1, 2, and 5 (notice the thermodynamic boundary 2 is shifted by 15 bp compared to the exon-exon boundary), but notice also some weaker signals close to boundaries 3 and 4 (such weak signals have not been taken into account in the histogram of Fig. 6). When the intron is inserted [Fig. 6(b)], many other "spurious" peaks appear; therefore exon-introns boundaries may be difficult to detect from thermodynamics, although they still persist.

We have analyzed 48 genes for which exons-introns boundaries have been experimentally confirmed, selected randomly from the GenBank Refseq set [20], and 35 housekeeping (HK) genes, taken from Ref. [21] (HK genes are virtually expressed in all tissues [1]). For each sequence, we have produced a histogram of thermodynamic boundaries as those of Figs. 5 and 6. We recorded the position of the major peaks and calculated the scaled distance from exon-exon boundaries. As an example, if a thermodynamic boundary is found at position $i_{\rm th}$ and it is contained in an exon beginning at i_1 and ending at i_2, the scaled distance is defined as $x = (i_{\rm th} - i_1)/(i_2 - i_1)$; thus $0 \le x \le 1$. Figure 7 shows a plot of the function $N_\alpha(x)$, defined as the number of observed thermodynamic boundaries in the interval $x - \alpha/2$, $x + \alpha/2$. For reference, this result is compared to that obtained from random distributions of uncorrelated boundaries, which is a constant (dashed lines in Fig. 7). The plot clearly demonstrates the significance of the observed correlations. As is clear from the Figs. 2–5, not all exon-exon boundaries are detected by thermody-

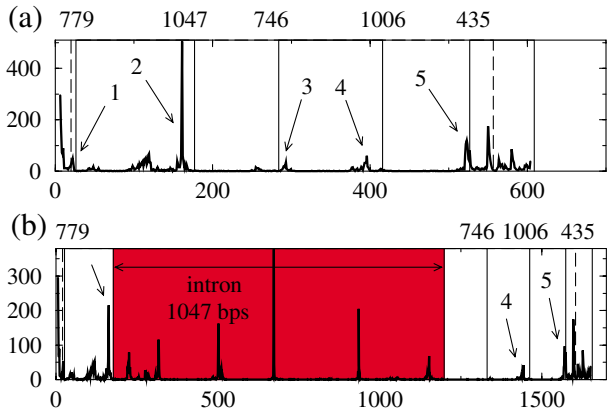

FIG. 6 (color online). (a) Melting histogram for the cDNA sequence encoding for the ribosomal protein L11 (RPL11) with NCBI entry code NM_000975. The numbers above the exon-exon boundaries denote the length of introns in the genomic DNA. (b) The same sequence as in (a), to which the intron 2 of 1047 bp is added (shaded area).

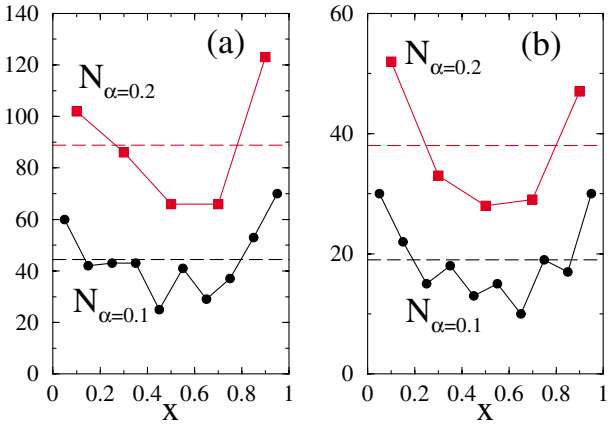

FIG. 7 (color online). Plots of $N_\alpha(x)$, the number of observed thermodynamic boundaries in the interval $x - \alpha/2$, $x + \alpha/2$. (a) Set of 48 human cDNAs selected at random from the Refseq Genbank set [20] and (b) set of 35 human cDNA from housekeeping genes taken from [21]. The dashed lines correspond to a random distribution of uncorrelated exon-exon and thermodynamic boundaries.

namics. Our statistical analysis indicates that the detection score is roughly 35%.

In our view, the correspondence between exon-exon and thermodynamic boundaries suggests a possible pathway of introns insertion in the genome during evolution. Introns seem to have targeted exposed bases at a "fork" between a double helix and an open DNA region as schematically shown in Fig. 8. The exposed bases could have provided sites where binding with "foreign" introns sequences was possible, probably because the two strands are still sufficiently close to each other to integrate efficiently the intron. To our knowledge, the idea that the thermodynamic stability of the double-stranded DNA played a role in the intron insertion in eukaryotic genomes has not been considered in other studies. The emphasis so far has been put on the specific sequence composition of a few base pairs around the insertion site; for instance, it has been shown [6] that the upstream exon tends to end with (C/A)AG and the downstream exon tends to start with GT, which were interpreted as sequence specific targeting of some still uncharacterized intron insertion machinery. As the precise biochemical mechanism for the introns insertion is not yet understood, the hypothesis of insertion both at specific sequence sites or at thermodynamic boundaries remains plausible. Since about $1/3$ of the exon-exon boundaries are detected by our thermodynamic analysis, it is also possible that other mechanisms of introns insertion may have been used during evolution as well. A systematic study of the relationship between thermal and exon-exon boundaries combined with a phylogenetic analysis for different species in the eukaryotic kingdom may provide further insights into these issues.

Finally, the boundaries separating regions of DNA with different stability properties, found in the melting analysis,

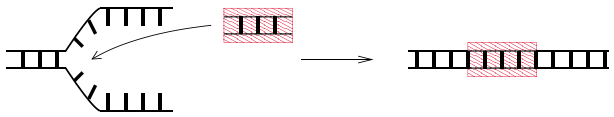

FIG. 8 (color online). Schematic view of a possible pathway of intron insertion in the genome: introns seem to have targeted preferentially thermodynamic boundaries.

should manifest themselves also under nonequilibrium conditions. Mechanical unzipping of DNA, for instance, is known to generate metastable forks [22]. Moreover, the thermodynamic boundaries obtained from statistical mechanics approaches are robust with respect to "realistic" changes in the stacking energies and entropic parameters [14,15]. These modifications do not affect the location of the thermodynamic boundaries noticeably. Likewise, as we explicitly verified, a small percentage of mutations does not have a strong influence on the thermodynamic boundaries. As the coding parts of the genes tend to be highly conserved across distant species [1], we expect that the melting features of cDNA sequences are as similar to that of the old "protogenes" as they were before the intron insertions.

[1] B. Alberts *et al.*, *Molecular Biology of the Cell* (Garland Science, New York, 2002).
[2] W. Gilbert, Nature (London) **271**, 501 (1978).
[3] C. C. F. Blake, Nature (London) **273**, 267 (1978).
[4] W. Gilbert, Cold Spring Harbor Symp. Quant. Biol. **52**, 907 (1987).
[5] T. Cavalier-Smith, Nature (London) **315**, 283 (1985).
[6] N. J. Dibb and A. J. Newman, EMBO J. **8**, 2015 (1989).
[7] A. Stoltzfus *et al.*, Science **265**, 202 (1994).
[8] S. W. Roy, A. Fedorov, and W. Gilbert, Proc. Natl. Acad. Sci. U.S.A. **99**, 15 513 (2002).
[9] A. Coghlan and K. H. Wolfe, Proc. Natl. Acad. Sci. U.S.A. **101**, 11 362 (2004).
[10] R. M. Wartell and A. S. Benight, Phys. Rep. **126**, 67 (1985).
[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] R. D. Blake *et al.*, Bioinformatics **15**, 370 (1999).
[14] R. Blossey and E. Carlon, Phys. Rev. E **68**, 061911 (2003).
[15] E. Yeramian, Gene **255**, 139 (2000).
[16] E. Yeramian, Gene **255**, 151 (2000).
[17] K. A. Marx *et al.*, J. Biomol. Struct. Dyn. **16**, 329 (1998).
[18] E. Yeramian *et al.*, Bioinformatics **18**, 190 (2002).
[19] E. Yeramian and L. Jones, Nucleic Acids Res. **31**, 3843 (2003).
[20] http://www.ncbi.nlm.nih.gov/.
[21] See http://www.cgen.com/supp_info/Housekeeping_genes.html.
[22] J. D. Weeks *et al.*, Biophys. J. **88**, 2752 (2005).