Universal power law behaviors in genomic sequences and evolutionary models

Loredana Martignetti* and Michele Caselle†
Dipartimento di Fisica Teoric, Università di Torino and INFN, Via Pietro Giuria 1, I-10125 Torino, Italy
(Received 20 February 2007; published 2 August 2007)

We study the length distribution of a particular class of DNA sequences known as the 5′ untranslated regions exons. These exons belong to the messenger RNA of protein coding genes, but they are not coding (they are located upstream of the coding portion of the mRNA) and are thus less constrained from an evolutionary point of view. We show that in both mice and humans these exons show a very clean power law decay in their length distribution and suggest a simple evolutionary model, which may explain this finding. We conjecture that this power law behavior could indeed be a general feature of higher eukaryotes.

DOI: 10.1103/PhysRevE.76.021902 PACS number(s): 87.10.+e, 87.14.Gg

I. INTRODUCTION

Recently, there has been a lot of effort devoted to trying to find universal laws in nucleotide distributions in DNA sequences. A typical example was the identification more than 10 years ago of long-range correlations in the base composition of DNA (see, for instance, [1,2], and references therein). With the availability of complete sequenced genomes, the correlation property of length sequences has been studied separately for coding and noncoding segments of complete bacterial genomes, showing a rich variety of behavior for different kinds of sequences [3,4]. This line of research has been recently extended to the search for similar universal distributions of more complex features of eukaryotic DNA sequences, for instance, the 5′ untranslated regions (5′UTR) lengths [5], UTR introns [6], or strand asymmetries in nucleotide content [7,10]. The main reason of interest for this type of analysis is the search of general rules behind the observed universal behaviors. The hope is to obtain, in this way, new insight in the evolutionary mechanisms shaping higher eukaryotes genomes and to understand the functional role of the various portions of the genome. An intermediate important step of this process is the construction of simplified (and possibly exactly solvable) stochastic models to describe the observed behaviors. This is the case, for instance, of the model discussed in [8] for base pair correlations or the model proposed in [5] for the 5′UTR length. In this paper we describe a similar universal law for the exon length in the 5′UTR of the human and mouse genomes. Looking at the 5′UTR exons collected in the existing genome databases for the two organisms we shall first show that they follow with a high degree of confidence a power law distribution with a decay exponent of about 2.5 and then suggest a simple solvable model to describe this behavior.

We shall also compare the impressive stability of the power law decay of 5′UTR exons with the distributions in the case of the 3′UTR and coding exons which turn out to be completely different. This is most probably due to the different evolutionary pressures that are subject to the three types of sequences.

We think that the behavior that we observed should indeed be a general feature of higher eukaryotes, however its identification requires a very careful annotation of the 5′UTR which exist, for the moment, only for a human and a mouse (see Table I).

This paper is organized as follows. After a short introduction to the biological aspects of the problem (Sec. II) we discuss the exon length distribution in Sec. III. Section IV is then devoted to the discussion of a simple stochastic model which gives as equilibrium distribution the observed power-like behavior. Details on the model are collected in the Appendix.

II. BIOLOGICAL BACKGROUND

In eukaryotic organisms, DNA information stored in genes is translated into proteins through a series of complex processes, carefully controlled at each step by specific regulatory mechanisms activated by the cell. In particular, two crucial events in this process are the production of an intermediate molecule, the messenger RNA (mRNA) transcript, and the translation of the mRNA into proteins. The cell provides fine regulatory systems to regulate the gene expression both at transcriptional and post-transcriptional levels, using several *cis*-acting signals located in the DNA sequence. A common molecular basis for much of the control of the gene expression (whether it occurs at the level of initiation of transcription, mRNA processing, translation, or mRNA transport) is the binding of protein factors and specific RNA elements to regulatory nucleic acid sequences.

Once mRNA is transcribed, it usually contains not only the protein coding sequence, but also additional segments, which are transcribed but not translated, namely a flanking

TABLE I. Estimate of critical index α and length threshold $l_{\min}$ for the power law distribution of the 5′UTR exons in a human and a mouse.

Species	$\tilde{\chi}^2$	α index	$l_{\min}$ (bps)
H. sapiens	0.52	2.56(2)	150
M. musculus	0.74	2.61(2)	140

*martigne@to.infn.it
†caselle@to.infn.it

5′ untranslated region and a final 3′ untranslated region.+[11] Nucleotide patterns or motifs located in 5′UTR and 3′UTR are known to play crucial roles in the post-transcriptional regulation. Most of the primary transcripts of eukaryotic genes also contain sequences (named "introns") which are eliminated during a maturation process named "splicing." The sequences which survive this splicing process are named "exons," they are glued together by the splicing machinery and form the mature mRNA transcript. Both the UTR and the coding portions of the mRNA are usually composed by the union of several exons. It is thus possible to classify the exons as coding, 3′UTR, and 5′UTR, depending on the portion of the mRNA to which they belong.[22]

A cell can splice the "primary transcript" in different ways and thereby make different polypeptide chains from the same gene (a process called alternative RNA splicing) and a substantial proportion of higher eukaryotic genes (at least one-third of human genes, it is estimated) produce multiple proteins in this way (isoforms), thanks to special signals in primary mRNA transcripts.

Some hints about the 5′ and 3′ role in gene expressions can be derived from a quantitative analysis of the UTR length.

Recent large scale databases suggest that the mean 3′UTR length in human transcript is nearly 4 times longer than the mean human 5′UTR length [9] and that the evolutionary expansion of 3′UTR in higher vertebrates, not observed in 5′UTR, is associated to their peculiar regulatory role. Very recent works revealed the existence of an extremely important post-transcriptional regulatory mechanism, performed by an abundant class of small noncoding RNA, known as microRNA (miRNA), that recognize and bind to multiple copies of partially complementary sites in the 3′UTR of target transcripts, without involving the 5′UTR [11–13].

Differently, 5′UTR sequences are expected to be constrained mainly by the splicing process and translation efficiency. The exons in the 5′UTR are usually termed "noncoding exons," since they are not included in the protein coding portion of the transcript. However, their characteristics, as their length, secondary structure, and the presence of AUG triplets upstream of the true translation start in mRNA, known as upstream AUGs, have been shown to affect the efficiency of translation and to be preserved in the evolution of these sequences [5,14,15]. The 5′UTR exon length can vary between few tens until hundreds of nucleotides, without typical length scale around the favorite size, and the lower and upper bounds of this distribution are likely to be shaped by splicing and translation efficiency: exons that are too short (under 50 bps) leave no room for the spliceosomes (enzymes that perform the splicing) to operate [16], while exons that are too long can contain signals that affect translation efficiency. The 5′UTR "noncoding exons" are also free from selective pressure acting on coding exons, which strongly preserves the amino acid information written in triplets of nucleotides in the protein coding exons.

For these reasons, in our analysis we decided to construct strictly disjoint subsets of exons, according to their position in the transcript (5′UTR exons, protein coding exons, or 3′UTR exons).[33] Moreover, we created nonredudant genome-wide data sets of exons, considering only one isoform for each gene, the most extended one.

Curated information about DNA sequences and annotation of eukaryotic organisms are provided by the Ensembl project, based on a software system which produces and maintains automatic annotation on selected eukaryotic genomes [17].

III. ANALYSIS OF EXON DISTRIBUTION

We downloaded from the Ensembl database (release 40 [17]) all the available transcripts annotated as protein coding for different organisms, and we created a filtered data set of nonredundant exons, considering the most extended transcript for each gene. We eliminated all the exons with mixed annotations and grouped the remaining ones in three classes: 5′UTR, protein coding exons, and 3′UTR.

Plotting the length distribution of exons, separately for 5′UTR, coding exons and 3′UTR, we clearly observe different behaviors, which we think should reflect different evolutionary constraints acting on these classes of DNA sequences [Figs. 1(a)–1(c)]. In particular, the 5′UTR exon size distribution shows a remarkably smooth power decay for large enough values of the exon length. To assess this point and to evaluate the threshold above which the power law behavior starts, we fitted the observed distributions with a power law,

$$N(l) = l^{-\alpha}, \tag{1}$$

where $N(l)$ is the number of exons of length l.

In order to evaluate the goodness of the fits that we performed, we divided the set of all exons into 18 equivalent bins and then assumed the variance of these bins as an indication of the statistical uncertainty of our estimates (results are independent from the binning choice). This allowed us to perform a meaningful χ^2 test on the fits. This test is commonly used when an assumed distribution is evaluated against the observed data [18]. The quantity χ^2 may be thought of as a measure of the discrepancy between the observed values and the respective expected values. It is convenient to compute the reduced $\tilde{\chi}^2$ [i.e., the ratio $\chi^2/(N_p - N_f)$, where N_p is the number of points included in the fit and N_f is the number of parameters of the fit]. With this normalization one can immediately see if the fitting function correctly describes the data (which requires $\tilde{\chi}^2 \leq 1$). When instead $\tilde{\chi}^2 > 1$ the absolute value of $\tilde{\chi}^2$ gives a rough

[1] 5′ and 3′ refer to the position (5′ and 3′, respectively) of the carbon atoms of the mRNA backbone at the two extrema of the mRNA and are conventionally used to denote the "upstream" (5′) and "downstream" (3′) sides of the mRNA chain.

[2] Obviously in several cases one can have exons which are partially included in one of the two UTR and partially in the coding portion of the mRNA. These mixed exons were excluded from our analysis.

[3] Obviously in several cases one can have exons which are partially included in one of the two UTR and partially in the coding portion of the mRNA. These mixed exons were excluded from our analysis.

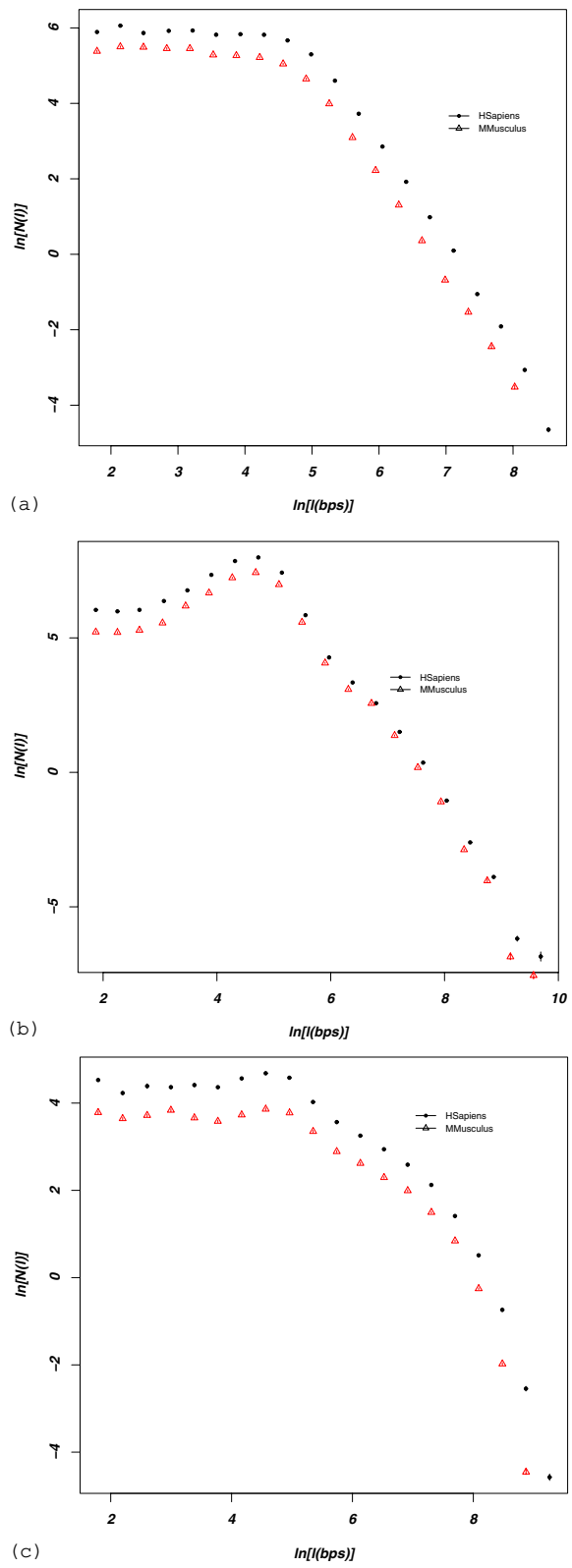

FIG. 1. (Color online) Exon length distribution in (a) 5′UTR, (b) protein coding exons, and (c) 3′UTR in human and mouse genome reported in ln-ln histograms (with bin size growing logarithmically). Plot errors are derived by dividing the complete data set into subsets of comparable dimension, avoiding biological biases, and averaging the length distribution of each subset.

TABLE II. $\tilde{\chi}^2$ values for the linear fit of the protein coding exons and the 3′UTR exons length distribution, in the same range where we are able to fit the power law decay of the 5′UTR exons length.

Species	Protein coding exons	3′UTR exons	$l_{\min}$ (bps)
H. sapiens	84.37	13.46	150
M. musculus	153.31	5.91	140

estimate of how inaccurate the tested distribution is to describe the data.

We fitted the data for the 5′UTR exons setting a minimum threshold on the exon length and then gradually increasing this threshold until a reduced $\tilde{\chi}^2$ value smaller than 1 was obtained. The rationale behind this choice is that (as we shall see below) the power law decay is likely to be an asymptotic behavior which is violated for short exon lengths. Starting from $l_{\min} \sim 150$, both in human and in mouse, good $\tilde{\chi}^2$ values were obtained and we could estimate the critical index to be $\alpha \sim 2.5$. Detailed results of the fits are reported in Table I. The $\tilde{\chi}^2$ values that we found support in a quantitative way the power law behavior of the data, which was already evident looking at Fig. 1(a).

On the contrary, the coding exons and the 3′UTR exon length histograms display (on a ln-ln scale) nonlinear distributions with peaks of population around the favorite sizes. In the range, where we are able to fit the power law decay of 5′UTR exon length, $\tilde{\chi}^2$ values for linear fits in the other classes of exons are completely unacceptable (Table II).

The same plots for other organisms show an exactly analogous trend, but they are affected by poor annotation of 5′UTR and 3′UTR, which are very difficult to identify entirely (see Table III). In Table III we reported the total number of annotated protein coding genes, annotated 5′UTR, and annotated 3′UTR for four different mammalian genomes, according to Ensembl database release 40. These data underline the current lack in the annotation of 5′UTR and 3′UTR for other mammals, besides *H. sapiens* and *M. musculus*. For this reason, the same analysis performed for *H. sapiens* and *M. musculus* exon length distribution is prevented for other organisms.

In order to understand this peculiar behavior of the 5′UTR exons we propose and discuss in the following section a simple model of exon evolution. Our goal is to understand if it is possible to associate the different behavior that we observe to the greater freedom from selective pressure of

TABLE III. Annotated protein coding genes, 5′UTR and 3′UTR, in Ensembl database release 40.

Species	Annotated protein coding genes	Annotated 5′UTR	Annotated 3′UTR
H. sapiens	23735	18333	18592
M. musculus	24438	15945	16429
C. familiaris	18214	5925	6298
G. gallus	18632	7463	7670

the 5′UTR exons with respect to the coding and 3′UTR exons.

IV. MODEL

Evolutionary models describe evolution of the DNA sequence as a series of stochastic mutations. There are three major classes of mutations: changes in the nucleotide type, insertions or deletions of one or more nucleotides. The various existing models differ with each other for the different assumptions they make on the parameter which control these changes (for a review see, for instance, [19–21]). From a biological point of view the two main assumptions of any evolutionary model are as follows:

(i) Evolution can be described as a Markov process, i.e., the modifications of a DNA sequence only depend on its current state and not on its previous history.

(ii) Evolution is "shaped" by functional constraints: DNA sequences with a negligible functional role evolve at a higher rate with respect to functionally important regions. This implies that regions with different functional roles must be described by different choices of the various mutational rates. The free evolution of sequences without functional constraints is usually called "neutral evolution."

Let us provide a few examples:

(i) Protein coding exons are usually strongly constrained since the proteins they code have an important role in the life of the cell, however due to the redundancy of the genetic code, the third basis of each codon in the coding exons is free to mutate. On the contrary insertions and deletions are suppressed because they can dramatically affect the shape and function of the protein.

(ii) Sequences devoted to transcriptional regulations (which very often lie outside exons) are usually so important for the life of the cell that they are kept almost unchanged over millions of years of evolution.

(iii) Regulatory sequences on the messenger RNA (mRNA) whose function often depends on the tridimensional shape of the RNA molecule and not on its exact sequence are in an intermediate situation between the above cases and the neutral evolution: they can tolerate mutations which do not modify their tridimensional shape (typically these are pairs of pointlike changes of bases and are usually called "compensatory mutations"). Most of the mRNA regulatory signals of this type are located in the 3′UTR exons.

(iv) Sometimes the 5′UTR contain regulatory sequences of the transcriptional type (which, as mentioned above, are strongly conserved under evolution), but their relative position does not seem to have a crucial functional role. They can thus tolerate insertion and deletions as far as they do not affect the regulatory regions.

Since in our model we are only interested in the exon length distribution we may neglect the nucleotide changes and concentrate only on insertions and deletions. From this point of view, according to the above discussion, both coding and 3′UTR should behave as highly constrained sequences, while the 5′UTR should be more similar to the neutrally evolving ones. With this picture in mind we decided to model the neutral evolution of a DNA sequence under the effect of insertions and deletions only, to see which general behavior one should expect for the length distribution and then compare it with the data discussed in the preceding section.

To this end, let us define n_j as the number of 5′UTR exons of length j in the genome and let N be the total number of such exons. Let $x_j \equiv n_j/N$ be the fraction of exons of length j.

If we assume that the exon length distribution evolves as a consequence of insertions and/or deletions of single nucleotides we find the following evolution equation for the $x_j(t)$ (where t labels the time step of this process):

$$x_j(t+1) = x_j(t) + (j-1)\alpha x_{j-1}(t) - j\alpha x_j(t) + (j+1)\beta x_{j+1}(t) - j\beta x_j(t), \tag{2}$$

where α and β denote the insertion and deletion probabilities, respectively, and we have taken into account the fact that for an exon of length j there are exactly j sites in which the new nucleotide can be inserted (i.e., that the insertion and deletion probabilities are linear functions of j, since the implied assumption is that all sites in our sequences are independent of one another).

At equilibrium the exon length distribution must satisfy the following equation (we omit the t dependence which is now irrelevant):

$$(j-1)\alpha x_{j-1} - j\alpha x_j + (j+1)\beta x_{j+1} - j\beta x_j = 0. \tag{3}$$

It is easy to see that the only solution compatible with this equation is a power law of this type: $x_j = cj^{\eta}$ with c a suitable normalization constant. Inserting this proposal in Eq. (3) one immediately finds $\eta = -1$.

This result is very robust, it does not depend on the values of α and β and, what is more important, it holds also if instead of assuming the insertion (or deletion) of a single nucleotide, we assume the insertion or deletion of oligos (i.e., small sequences of nucleotides) of length k, with any choice of the probability distribution for the oligos length as far as k is much smaller than the typical exon length. Moreover, one can also show that the power law decay still holds if we add to the process a fixed background probability of the creation of new exons of random length as far as this probability is smaller than $x_{j_{\max}}(\alpha - \beta)$, where $j_{\max}$ is the largest exonic length for which the power law is still observed. This is rather important since it is known that retrotransposed repeats (in particular of the Alu family) may in some cases (with very low probability) become new active exons and represent one of the major sources of evolutionary changes in the transcriptome.

On the contrary, this power law disappears if we assume that there is a finite probability that, as a consequence of the new insertion or deletion, the exon is eliminated. In this case the power law changes into a exponential distribution. This may explain why the power law decay is not observed in the coding and the 3′UTR portion of the genes which are under a much stronger selective pressure (the 3′UTR contain a lot of post-transcriptional regulatory signals).

Since the critical index that we observe in the actual exon distribution in a human and a mouse is much larger than 1, it is interesting to see which type of evolutionary mechanism could lead to a $\eta > 1$ behavior while keeping a power law decay. It is easy to see that this can be achieved assuming that the insertion (or deletion) probability is not linear with the length of the exon but behaves, say, as $p_{\text{insertion}} = \alpha j^{\lambda}$ with $\lambda > 1$. Then, following the same derivation discussed above, we find at equilibrium an exon length distribution $x_j = cj^{-\lambda}$.

A possible explanation for such nonlinear insertion rate comes from the observation that the transcribed portions of the genome (like the 5′UTR exons in which we are interested), besides the normal mutation processes typical of the intergenic regions, are subject to specific mutation events due to the transcriptional machinery itself (see, for instance, [7]).

It is clear from the above discussion that in this case the critical index of the exon distribution, strictly speaking, is not any more an universal quantity, but depends on the particular biological process leading to the $p_{\text{insertion}} = \alpha j^{\lambda}$ probability discussed above. However it is conceivable that similar mechanisms should be at work in related species. This, in our opinion, explains why the critical indices associated to the mouse and human distributions are so similar and lead us to conjecture that similar values should be found also in other mammalians as more and more 5′UTR sequences will be annotated.

Let us conclude by noticing that this whole derivation is based on the assumption that the system had reached its equilibrium distribution. This is by no means an obvious assumption and it is well possible that the fact that we observe a critical index larger than 1 simply denotes that the system is still slowly approaching the equilibrium distribution. There are three ways to address this issue. First, one should extend the analysis to other organisms (however, as we discussed above, this will require a better annotation of the UTR in these organisms). Second, one could reconstruct, by suitable aligning procedures, the UTR exons of the common ancestor between mouse and man and see if they also follow a power law distribution and, if this is the case, which is the critical index. Third, one could simulate the model discussed above and look to the behavior of the exon distribution as the equilibrium is approached. We plan to address these issues in a forthcoming presentation.

This work was partially supported by FIRB Grant No. RBNE03B8KK from the Italian Ministry for Education, University and Research. The authors would like to thank D. Corà, E. Curiotto, F. DiCunto, I. Molineris, P. Provero, A. Re, and G. Sales for useful discussions and suggestions.

APPENDIX: DERIVATION OF THE POWER LAW

Inserting the distribution $x_j = cj^{\eta}$ in Eq. (3) we find

$$\alpha(j-1)^{\eta+1} - \alpha(j)^{\eta+1} + \beta(j+1)^{\eta+1} - \beta(j)^{\eta+1} = 0, \quad \text{(A1)}$$

which can be expanded in the large j limit as

$$j^{\eta+1}\left[\alpha\left(1 - \frac{\eta+1}{j}\right) - \alpha + \beta\left(1 - \frac{\eta-1}{j}\right) - \beta\right] = 0, \quad \text{(A2)}$$

which implies

$$(\beta - \alpha)\frac{\eta+1}{j} = 0, \quad \text{(A3)}$$

which (assuming $\beta \neq \alpha$) implies, as anticipated, $\eta = -1$.

A few observations are in order:

(a) It is clear from the derivation that the result is independent from the specific values of α and β as far as they do not coincide. This independence from the details of the model holds also if we assume at each time step a finite, constant (i.e., not proportional to j) probability $\alpha'(\beta')$ of random insertion (deletion) of a nucleotide. In this case the evolution equation becomes

$$x_j(t+1) = x_j(t) + (j-1)\alpha x_{j-1}(t) - j\alpha x_j(t) + (j+1)\beta x_{j+1}(t) - j\beta x_j(t) + \alpha'[x_{j-1}(t) - x_j(t)] + \beta'[x_{j+1}(t) - x_j(t)], \quad \text{(A4)}$$

which still admits the same asymptotic distribution $x_j = cj^{-1}$.

(b) If we include a fixed exonization probability p_e to create a new exon from, say, duplicated or retrotransposed sequences the evolution equation changes trivially by simply adding such a constant contribution. The solution, in this case, becomes $x_j = cj^{-1} + d$, where the constant d is related to p_e as follows: $d = p_e/(\alpha - \beta)$, and it is negligible as far as it is smaller than $x_{j_{\max}}$.

(c) Remarkably enough, the above results are still valid even if the inserted (or deleted) sequence is composed by more than one nucleotide. Let us study as an example the situation in which we allow the insertion of oligos of length k with $0 < k < L$ and L smaller than the typical exon length. Let us assume for simplicity to neglect deletions and let us choose the same insertion probability α for all values of k. The evolution equation becomes

$$x_j(t+1) = x_j(t) + \alpha\left(\sum_{k=1}^{L} x_{j-k}(t)(j-k) - Ljx_j(t)\right), \quad \text{(A5)}$$

which implies

$$jx_j = \frac{1}{L}\sum_{k=1}^{L}(j-k)x_{j-k}. \quad \text{(A6)}$$

In the large j limit this equation admits again a power law solution $x_j = cj^{\eta}$. Inserting this solution in Eq. (A5) we find

$$j^{\eta+1}\alpha\left[\frac{1}{L}\sum_{k=1}^{L}\left(1 - \frac{k(\eta+1)}{j}\right) - 1\right] = 0, \quad \text{(A7)}$$

which is satisfied, as above, if we set $\eta = -1$.

(d) On the contrary, if we assume a finite probability $(1-\gamma)$ of elimination of an exon as a consequnce of the insertion (or deletion) event (as one would expect if the sequence is under strong selective pressure) we find the following evolution equation:

$$x_j(t+1) = x_j(t) + [(j-1)\alpha x_{j-1}(t)\gamma - j\alpha x_j(t)], \quad \text{(A8)}$$

where α is, as above, the insertion probability and we are assuming for simplicity single base insertions. This equation no longer admits a power law solution at equilibrium but requires an exponential distribution, $x_j = e^{-\lambda j} j^{\eta}$ with $\eta = -1$ and $\lambda = \ln(\gamma)$.

(e) It is instructive to reobtain the result discussed in (a) above by looking at the equilibrium equation as a recursive equation in j,

$$x_{j+1} = \frac{j}{j+1}\left(1 + \frac{\alpha}{\beta}\right)x_j - \frac{\alpha}{\beta}x_{j-1} \quad (j > j_{\min}) \quad \text{(A9)}$$

and

$$x_{j+1} = \frac{j}{j+1}\left(1 + \frac{\alpha}{\beta}\right)x_j \quad (j = j_{\min}), \quad \text{(A10)}$$

and construct recursively the solution for any j starting from $x_{j_{\min}} = c/j_{\min}$. The recursion can be solved exactly and gives

$$x_j = x_{j_{\min}} \frac{j_{\min}}{j} \frac{1 - \left(\frac{\alpha}{\beta}\right)^{j - j_{\min}+1}}{1 - \frac{\alpha}{\beta}}, \quad \text{(A11)}$$

which (assuming $\alpha < \beta$)[44] leads asymptotically to the solution $x_j = c/j$ with $c = x_{j_{\min}} \frac{j_{\min}}{1-\alpha/\beta}$. This result allows us to understand exactly the "finite size" corrections, with respect to this asymptotic solution, which turn out to be proportional to $\left(\frac{\alpha}{\beta}\right)^{j-j_{\min}+1}$ and vanish if only deletions (i.e., $\alpha = 0$) or only insertions (i.e., $\beta = 0$) are present. In these cases the asymptotic solution is actually the *exact* equilibrium solution of the stochastic model.

[4] If $\beta < \alpha$, one should study the inverse recursion relation starting from $x_{j_{\max}}$.

[1] W. Li, Comput. Chem. **21**, 257 (1997).
[2] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).
[3] Zu-Guo Yu, V. V. Anh, and Bin Wang, Phys. Rev. E **63**, 011903 (2000).
[4] Zu-Guo Yu, V. Anh, and Ka-Sing Lau, Physica A **301**, 351 (2001).
[5] M. L. Lynch, D. G. Scofield, and X. Hong, Mol. Biol. Evol. **22**, 1137 (2005).
[6] X. Hong, D. G. Scofield, and M. L. Lynch, Mol. Biol. Evol. **23**, 2392 (2006).
[7] M. Touchon, A. Arneodo, Y. d'Aubenton-Carafa, and C. Thermes, Nucleic Acids Res. **32**, 4969 (2004).
[8] P. W. Messer, P. F. Arndt, and M. Lassig, Phys. Rev. Lett. **94**, 138103 (2005).
[9] G. Pesole, S. Liuni, G. Grillo, F. Licciulli, F. Mignone, C. Gissi, and C. Saccone, Nucleic Acids Res. **30**, 335 (2002).
[10] M. Touchon, S. Nicolay, B. Audit, E. B. Brodie, Y. d'Aubenton-Carafa, A. Arneodo, and C. Thermes, Proc. Natl. Acad. Sci. U.S.A. **102**, 9836 (2005).
[11] D. P. Bartel, Cell **116**, 281 (2004).
[12] L. He and G. J. Hannon, Nat. Rev. Genet. **5**, 522 (2004).
[13] N. Rajewsky, Nat. Genet. **38**, S8 (2006).
[14] M. Iacono, F. Mignone, and G. Pesole, Gene **349**, 97 (2005).
[15] A. Churbanov, I. B. Rogozin, V. N. Babenko, H. Ali, and E. V. Koonin, Nucleic Acids Res. **33**, 5512 (2005).
[16] R. Sorek, R. Shamir, and G. Ast, Trends Genet. **20**, 68 (2004).
[17] Ensembl 2007, T. J. P. Hubbard *et al.*, Nucleic Acids Res. **35**, D610 (2007).
[18] *Statistical Methods in Bioinformatics*, 2nd ed., edited by W. J. Ewens and G. R. Grant (Springer, New York, 2005).
[19] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of DNA and Protein Sequences* (Cambridge University Press, Cambridge, 1998).
[20] G. Mitchison, J. Mol. Evol. **49**, 11 (1999).
[21] C. Kosiol, L. Bofkin, and S. Whelan, J. Biomed. Inf. **39**, 51 (2006).