2756–2763 Nucleic Acids Research, 2008, Vol. 36, No. 8 Published online 26 March 2008
doi:10.1093/nar/gkn086

Spontaneous symmetry breaking in genome evolution

Yaroslav Ryabov[1,*] and Michael Gribskov[2]

[1]Department of Chemistry, Purdue University, 560 Oval drive, Box 202 and [2]Department of Biological Sciences, Lilly Hall of Life Sciences 915 W. State Street, Purdue University, West Lafayette, IN, 47907, USA

Received November 14, 2007; Revised February 7, 2008; Accepted February 11, 2008

ABSTRACT

The quest for evolutionary mechanisms providing separation between the coding (exons) and noncoding (introns) parts of genomic DNA remains an important focus of genetics. This work combines an analysis of the most recent achievements of genomics and fundamental concepts of random processes to provide a novel point of view on genome evolution. Exon sizes in sequenced genomes show a lognormal distribution typical of a random Kolmogoroff fractioning process. This implies that the process of intron incretion may be independent of exon size, and therefore could be dependent on intron–exon boundaries. All genomes examined have two distinctive classes of exons, each with different evolutionary histories. In the framework proposed in this article, these two classes of exons can be derived from a hypothetical ancestral genome by (spontaneous) symmetry breaking. We note that one of these exon classes comprises mostly alternatively spliced exons.

INTRODUCTION

A substantial fraction of the genomic DNA sequence does not directly encode the primary structure of any cellular protein, or any other cellular product (1). This is largely due, in eukaryotes, to the division of genes into introns (noncoding parts of DNA) and exons (coding parts of DNA), each of which have pronounced size distributions (2). The actual mechanism (or mechanisms) of intron insertion is the subject of intense discussion and a currently popular working hypothesis suggests that introns may be largely produced by insertion of transposons—DNA elements that can move around to different positions within the genome (3). Very frequently, this point of view implicitly assumes that there is a higher probability of splitting longer exons since they are larger targets for transposons. Here, we show that the distributions of exon sizes for different organisms have a general property: the presence of two distinguishable classes of exons with different size distributions. We present an idealized scheme that explains how the observed distribution of exon sizes can be derived from a common ancestral genome by a random (quasi)-evolutionary process. This formal model makes it possible to investigate the evolution of the genomes of particular organisms, and to estimate the number of evolutionary steps that separate them from the hypothetical ancestral genome. Conceptually, our results support the opinion that at the initial stages of evolution, simple genomes had a lower fraction of introns (introns late hypothesis). The model we propose explains the observed lognormal distribution of exon sizes, and suggests that introns may be inserted by a process that is independent of exon size. Our findings also can be rationalized in relation to the phenomenon of alternative splicing (4).

MATERIALS AND METHODS

In our study, we used genome data for 12 animal species provided by Ensembl (5) (http://www.ensembl.org/), the joint project of the European Molecular Biology Laboratory—European Bioinformatics Institute and the Sanger Institute, and a plant genome sequence from The Arabidopsis Information Resource (TAIR, http://www.arabidopsis.org/) (6). All the entries annotated as 'exon' were retrieved for each complete genome. All duplicates of exons (which originate primarily from multiple accessions) starting and ending at the same points were excluded from the analysis. The natural logarithms of exon sizes were divided into bins of width $\Delta \ln(\textit{exon size}) = 0.2$ to obtain the exon size distributions presented in Figure 1 and in Supplementary data. The distributions were fit using the unweighted χ^2-measure normalized over the number of degrees of freedom (7) as the criterion of fitting quality. The values of fitted

*To whom correspondence should be addressed. Tel: +301 435 9034; Fax: 301 480 0028; Email: yryabov@mail.nih.gov

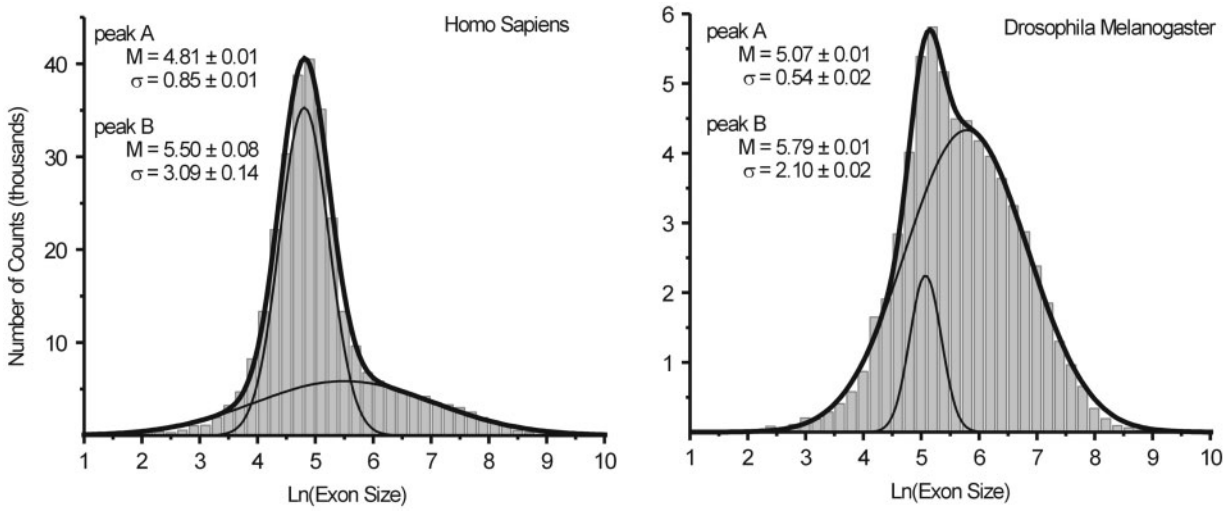

Figure 1. Distributions of the natural logarithms of exon sizes for the *Homo sapiens* and *Drosophila melanogaster* genomes. Both data sets can be approximated by a sum of two Gaussian peaks with very high correlation between the data and the best-fit distributions: Pearson correlation coefficients $r = 0.997$ and $r = 0.998$ for *Homo sapiens* and *Drosophila melanogaster* genomes, respectively. Data are shown as gray bars; thick lines represent the best fit approximation, while thin lines depict individual locations and shapes of two Gaussian peaks.

parameters are presented in Table 1. Table 2 shows the corresponding values of χ^2 and the Pearson correlation coefficient, r, which characterize the agreement between original data and the fitted model.

Three classes of fitting models were used:

(I) A model based on the assumption that an exon can be split at any position with equal and constant probability; this model leads to an exponential distribution of exon sizes:

$$dN = N\lambda e^{\ln E - \lambda e^{\ln E}} d\ln E,$$

where E is exon size, dN is the number of exons in a bin, N is the amplitude of the peak and λ is the probability of splitting an exon at a particular place.

(II) Two models that produce lognormal distributions of exon sizes. These models are based on a Kolmogoroff process, which does not assume any relationship between exon size and probability of splitting an exon. Particularly, to fit the data we used a model of single lognormal peak

$$dN = \frac{N}{\sigma\sqrt{\pi/2}} e^{-2((\ln E - M)/\sigma)^2} d\ln E,$$

and a mixture of two lognormal distributions,

$$dN = \frac{N_a}{\sigma_a\sqrt{\pi/2}} e^{-2((\ln E - M_a)/\sigma_a)^2} + \frac{N_b}{\sigma_b\sqrt{\pi/2}} e^{-2((\ln E - M_b)/\sigma_b)^2} d\ln E,$$

where, N, N_a and N_b are the amplitudes, M, M_a and M_b the mean positions, and σ, σ_a and σ_b the variances of lognormal peaks observed in the data.

(III) A combination of a Weibull distribution and exponential distribution

$$dN = \left(N\lambda e^{\ln E - \lambda e^{\ln E}} + N_w e^{c\ln E - \lambda_w e^{c\ln E}}\right) d\ln E,$$

where N_w, λ_w and c are the amplitude, frequency parameter and shape parameter of the Weibull distribution, respectively.

Sequences of pseudorandom numbers were obtained using the *Mersenne Twister* algorithm (8) implemented in the standard MATLAB 7.1 installation. The sequences of pseudorandom numbers were seeded with the values 14 (sequence **A**) and 16 (sequence **B**), and were used to generate the multiplicative processes presented in Figure 2. Each pseudorandom sequence consists of 10^7 random numbers, i.e., 10^7 exon splitting events. The ratio between the initial lengths of the two ancestral exons used to generate the exon size distributions presented in Figure 2 was 1:1000.

The confidence intervals shown for the fitted parameters in Figure 1, error bars in Figure 3, Tables 1 and 3, and in the Supplementary data are 95% confidence intervals estimated from the covariance matrix (7).

For *Homo sapiens* and *Mus musculus* genomes, we also analyzed distributions of alternatively spliced exons. These data were received from the third release of the Alternative Splicing Database (ASD), (http://www.ebi.ac.uk/asd/altsplice/). This resource provides manually curated data on alternative splicing with all exons confirmed by EST/mRNA alignments (9,10). These data were also divided into bins of width $\triangle \ln$ (*exon size*) = 0.2 and fitted using the unweighted χ^2-measure to models of single exponential peak, single Weibull peak, and single lognormal distribution (Figure 4 and Table 3).

In the case of fitting the model with two lognormal peaks, the fractions of exons contributing to each peak

Table 1. Parameters of fitting for different models of exon size distribution

Species names	*Caenorhabditis elegans*	Arabidopsis thaliana	*Anopheles gambiae*	*Tetraodon nigroviridis*	*Bos taurus*	*Drosophila melanogaster*	*Canis familiaris*	*Gallus gallus*	*Danio rerio*	*Macaca mulatta*	*Mus musculus*	*Homo sapiens*	*Pan troglodytes*
Exponential peak													
$N \times 0^{-3}$	28 ± 1.5	33 ± 1.7	12 ± 0.5	48 ± 3.5	52 ± 3.3	13 ± 0.5	52 ± 3.4	44 ± 2.9	64 ± 4.0	56 ± 3.6	52 ± 3.9	67 ± 4.8	55 ± 3.9
$\lambda \times 10^3$	5.3 ± 0.4	5.1 ± 0.4	3.6 ± 0.2	6.7 ± 0.7	7.7 ± 0.7	2.8 ± 0.2	7.5 ± 0.7	7.6 ± 0.7	7.7 ± 0.7	7.3 ± 0.7	7.0 ± 0.7	6.9 ± 0.7	7.2 ± 0.7
Lognormal peak													
$N \times 10^{-3}$	23 ± 1	30 ± 1	11 ± 1	36 ± 1	38 ± 1	13 ± 1	38 ± 1	33 ± 1	47 ± 1	42 ± 1	39 ± 1	50 ± 1	41 ± 1
σ	1.3 ± 0.1	1.8 ± 0.1	1.7 ± 0.1	1.0 ± 0.1	1.1 ± 0.1	2.0 ± 0.1	1.0 ± 0.1	1.0 ± 0.1	1.0±0.1	1.1 ± 0.1	1.0 ± 0.1	1.0 ± 0.1	1.0 ± 0.1
M	5.0 ± 0.1	5.0 ± 0.1	5.4 ± 0.1	4.9 ± 0.1	4.8 ± 0.1	5.6 ± 0.1	4.8 ± 0.1	4.8 ± 0.1	4.8 ± 0.1	4.8 ± 0.1	4.8 ± 0.1	4.8 ± 0.1	4.8 ± 0.1
Exponential peak plus Weibull peak													
$N \times 10^{-3}$	13 ± 2	25 ± 1	9.3 ± 0.3	18 ± 1	12 ± 1	11 ± 0.3	14 ± 1	11 ± 1	22 ± 2	14 ± 1	13 ± 1	18 ± 1	13 ± 1
$\lambda \times 10^3$	3.5 ± 0.5	3.7 ± 0.2	2.7 ± 0.1	4.3 ± 0.3	4.1 ± 0.3	2.1 ± 0.1	4.7 ± 0.3	4.2 ± 0.3	5.7 ± 0.4	3.0 ± 0.2	2.2 ± 0.2	2.1 ± 0.2	2.5 ± 0.2
$N_w \times 10^2$	32 ± 44	0.8 ± 2.2	0.1 ± 0.2	3.5 ± 2.7	62 ± 16	0.1 ± 0.1	33 ± 10	46 ± 11	23 ± 12	73 ± 19	31 ± 12	59 ± 24	53 ± 17
c	2.3 ± 0.3	3.2 ± 0.6	3.1 ± 0.3	2.9 ± 0.2	2.4 ± 0.1	3.1 ± 0.4	2.5 ± 0.1	2.4 ± 0.1	2.6 ± 0.1	2.4 ± 0.1	2.5 ± 0.1	2.5 ± 0.1	2.4 ± 0.1
$\lambda_w \times 10^6$	10 ± 20	0.3 ± 0.9	0.1 ± 0.2	0.5 ± 0.4	8.6 ± 2.3	0.1 ± 0.2	4.5 ± 1.3	7.3 ± 1.7	2.8 ± 1.5	9.2 ± 2.5	3.8 ± 1.5	5.7 ± 2.4	6.3 ± 2.1
Two lognormal peaks													
$N_a \times 10^{-3}$	5 ± 0.8	9 ± 0.9	1.3 ± 0.1	21 ± 1.3	21 ± 1.5	1.5 ± 0.1	20 ± 1.4	17 ± 1.2	26±1.2	28±1.1	29 ± 0.9	37 ± 0.9	30 ± 0.9
$N_b \times 10^{-3}$	18 ± 0.9	21 ± 1.0	10 ± 0.1	18 ± 1.3	22 ± 1.5	11 ± 0.1	23 ± 1.4	20 ± 1.2	29±1.1	21±1.2	16 ± 1.1	23 ± 1.1	19 ± 1.0
σ_a	0.7 ± 0.1	0.8 ± 0.1	0.6 ± 0.1	0.8 ± 0.1	2.1 ± 0.1	0.5 ± 0.1	2.1 ± 0.1	2.2 ± 0.1	2.5±0.1	0.8±0.1	0.8 ± 0.1	0.9 ± 0.1	0.8 ± 0.1
σ_b	1.5 ± 0.1	1.8 ± 0.1	1.9 ± 0.1	1.9 ± 0.1	0.8 ± 0.1	2.1 ± 0.1	0.8 ± 0.1	0.8 ± 0.1	0.8±0.1	2.7±0.2	3.1 ± 0.2	3.1 ± 0.1	3.1 ± 0.2
M_a	4.7 ± 0.1	4.5 ± 0.1	5.1 ± 0.1	4.9 ± 0.1	4.6 ± 0.1	5.1 ± 0.1	4.6 ± 0.1	4.6 ± 0.1	4.5±0.1	4.8±0.1	4.8 ± 0.1	4.8 ± 0.1	4.8 ± 0.1
M_b	5.2 ± 0.0	5.4 ± 0.0	5.5 ± 0.0	5.0 ± 0.0	4.8 ± 0.0	5.8 ± 0.0	4.8 ± 0.0	4.8 ± 0.0	4.8±0.0	4.9±0.1	5.4 ± 0.1	5.5 ± 0.1	5.1 ± 0.1

Table 2. Fitting quality, χ^2 and r, together with the number of data points and number of degrees of freedom, df

Species names	*Caenorhabditis elegans*	Arabidopsis thaliana	*Anopheles gambiae*	*Tetraodon nigroviridis*	*Bos taurus*	*Drosophila melanogaster*	*Canis familiaris*	*Gallus gallus*	*Danio rerio*	*Macaca mulatta*	*Mus musculus*	*Homo sapiens*	*Pan troglodytes*
Number of data points	50	45	51	50	51	51	51	50	51	51	51	51	51
Exponential peak													
$\chi^2 \times 10^{-3}$	3081	3577	259	14 947	13 991	316	14 876	10 874	19 826	16 196	18 559	29 373	18 882
r	0.831	0.847	0.910	0.748	0.777	0.902	0.771	0.770	0.786	0.773	0.725	0.731	0.743
df	48	43	49	48	49	49	49	48	49	49	49	49	49
Lognormal peak													
$\chi^2 \times 10^{-3}$	301	1324	69	1071	1579	113	1540	1400	3400	2014	1995	3749	2148
r	0.984	0.945	0.977	0.982	0.975	0.966	0.977	0.971	0.964	0.972	0.971	0.966	0.971
df	47	42	48	47	48	48	48	47	48	48	48	48	48
Exponential peak plus Weibull peak													
$\chi^2 \times 10^{-3}$	757	1043	42	612	161	53	145	93	459	280	548	1135	476
r	0.961	0.958	0.986	0.990	0.998	0.985	0.998	0.998	0.995	0.996	0.992	0.990	0.994
df	45	40	46	45	46	46	46	45	46	46	46	46	46
Two lognormal peaks													
$\chi^2 \times 10^{-3}$	53	48	8	162	308	6	256	244	341	277	322	342	264
r	0.997	0.998	0.998	0.997	0.996	0.998	0.996	0.995	0.997	0.996	0.996	0.997	0.997
df	44	39	45	44	45	45	45	44	45	45	45	45	45

(Table 4) were estimated by formal integration of the peak areas, which, for a standard Gaussian distribution, are proportional to the peak amplitudes.

F-test critical values, $F_{0.05}$ presented in Table S1 in Supplementary data, were calculated for significance level of 5% and number of degrees of freedom, *df*, given in Table S1 and Table 3. For each pair of models under consideration, we compared the ratio, $\chi^2_{\mathrm{I}}/\chi^2_{\mathrm{II}}$, to the calculated $F_{0.05}$. When $\chi^2_{\mathrm{I}}/\chi^2_{\mathrm{II}} > F_{0.05}$, the hypothesis that the variance σ_I associated with χ^2_{I} is greater than variance σ_{II} associated with χ^2_{II}, i.e. the assumption that Model II is a better description of the data than Model I (11), is verified at the $P < 0.05$ level.

RESULTS

The *Homo sapiens* (12) and *Drosophila melanogaster* (13) genomes show (Figure 1) a striking similarity in the distributions of exon sizes; in both, the distribution of the logarithm of exon size forms two distinctive peaks. Other genomes [see Supplementary data in which we analyze all complete genome data provided by Ensembl (5) and TAIR (6)] show similar patterns. A simplistic model of intron insertion assumes that the probability of inserting an intron is equal at all positions of an exon (making it more likely that a longer exon will be split). This type of process would lead to an exponential distribution of exon

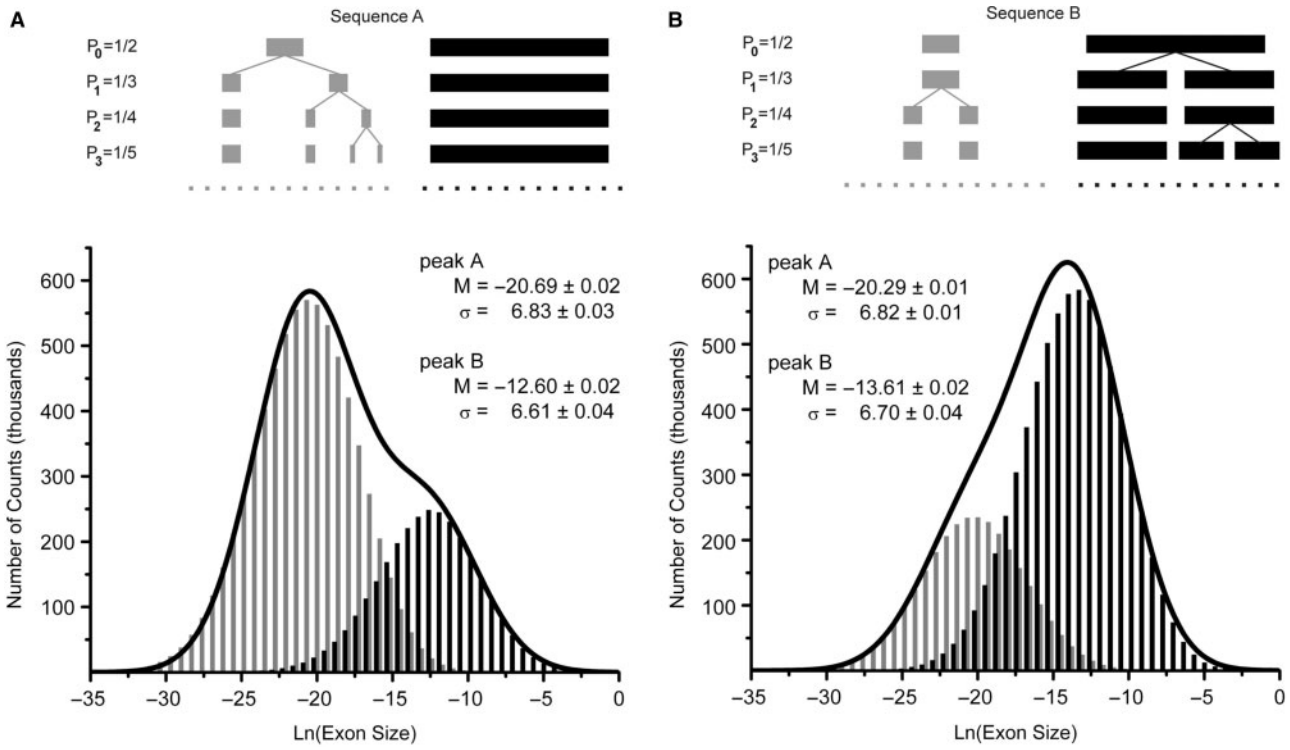

Figure 2. Distributions of the natural logarithm of exon sizes for two different realizations of a random multiplicative process starting from an ancestral genome comprising two exons. The diagrams at the top of the figure illustrate two different splitting patterns for sequences **A** and **B** (see Methods section). The solid line represents the sum of the two peaks. The peaks shown with gray bars in the plot originated from the short exon in ancestral genome (also gray in diagram); those shown in black bars originated from the long exon in ancestral genome (also black in diagram). The probability, P, of a splitting event for a single exon at each step of the process is also indicated. The figure illustrates symmetry breaking between the two parts of model genome. For the three first steps of the pseudorandom number sequence shown in (**A**), all splitting events happened to occur for the exons that originated from the short part of the model ancestral genome. Thus, by the third step of the process, four out of five exons originated from this part of model ancestral genome. This subset of exons has largest cumulative probability ($P = 4/5$) for the next splitting event. For the sequence in (**B**), at the third step of the process the largest cumulative probability for the next splitting event ($P = 3/5$) belongs to the subset of exons that originated from the long part of ancestral exon. This initial break of symmetry between the two subsets of exons in the model genome persists and produces the two different peak dispositions shown in panels (**A**) and (**B**).

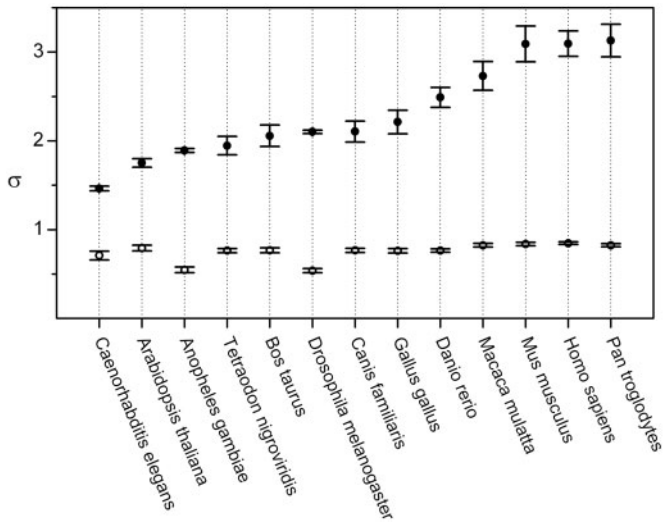

Figure 3. Widths, σ, of exon size distributions for several complete genomes (see Table 1 and in Supplementary data). Open symbols correspond to the narrow peaks; full symbols to the wide peaks. All the species were ordered according to the width of wider peak in the exon distributions. Error bars indicate 95% CI.

sizes showing a single peak on a logarithmic scale (14,15). Recently, the simplistic exponential model of exon size distribution was reconsidered by Gudlaugsdottir and co-authors (15) who suggested treating distributions of exon sizes as a combination of Weibull (16) and exponential distributions. However, they did not provide a model for an evolutionary process, which could lead to the Weibull distribution but rather consider this distribution to be only an empirical approximation of the observed exon size data.

Our analysis of thirteen genomes (Table 2), shows that the exponential distribution model always produces the largest χ^2-values, which signifies that it is the least adequate fit to the observed exon size distributions. The model of a single lognormal peak is a closer fit to the data. The mixture of Weibull and exponential distributions suggested by Gudlaugsdottir and co-authors is a better approximation, but for most genomes (except *Bos taurus*, *Canis familiaris* and *Gallus gallus*) a combination of two lognormal distributions produces the best agreement between the real data and a fitted model. The length distributions of alternatively spliced exons in *Homo sapiens* and *Mus musculus*, drawn from ASD (9,10), reveal a single

Table 3. Parameters of fitting for different models approximating size distributions of alternatively spliced exons (see Methods Sections)

Species names	*Mus musculus* 28 data points	*Homo sapiens* 35 data points
Exponential peak	$df = 26$	$df = 33$
$N \times 10^{-3}$	33.3 ± 3.7	39.5 ± 3.9
$\lambda \times 10^{3}$	8.1 ± 1.3	7.9 ± 1.1
$\chi^2 \times 10^{-3}$	16 880	18 313
r	0.648	0.677
Weibull peak	$df = 25$	$df = 32$
$N_w \times 10^{2}$	40 ± 22	51 ± 22
c	2.4 ± 0.1	2.4 ± 0.1
$\lambda_w \times 10^{6}$	6.9 ± 3.2	7.4 ± 3.5
$\chi^2 \times 10^{-3}$	522	756
r	0.990	0.987
Lognormal peak	$df = 25$	$df = 32$
$N \times 10^{-3}$	23.5 ± 0.4	28.2 ± 0.1
σ	0.9 ± 0.1	0.9 ± 0.1
M	4.8 ± 0.1	4.8 ± 0.1
$\chi^2 \times 10^{-3}$	378	329
r	0.992	0.994

peak pattern (Figure 4). In this case, analysis of χ^2-values also shows that an exponential distribution is poorest fit to the data, and the lognormal distribution is the best (Table 3). More sophisticated comparison between different models used in this work involves *F*-test criterion (11) and suggests similar hierarchy of the models (see Methods section and Table S1 in Supplementary data).

DISCUSSION

The lognormal distribution is a normal (Gaussian) distribution of the logarithm of some quantity. This kind of distribution commonly originates from a Kolmogoroff random multiplicative process (17), which was originally introduced to describe the distribution of ore particle sizes observed in geological samples (18,19), and later was found to be useful as a paradigm for a whole universe of different breakage and splitting processes. A modified version of this process, in the context of our problem, can be described as follows. Let us consider a single exon, which is split by a random mutational process into two parts (equal in size, for the sake of illustration). Then, in the next step of the process, let us assume that one randomly selected part of the ancestral exon undergoes the same splitting. Subsequently, this process repeats for a large number of splitting events. The key assumption for this process, which is the same as the assumption of Kolmogoroff, is independence of the probability of undergoing a splitting event from exon size. This version of a random multiplicative process is slightly different from that discussed by Kolmogoroff, which assumes a constant frequency of splitting events for all parts of the fractionating set (exons in a genome in our case) at every time, while the process considered here assumes a single exon splitting event at each step of the process. Thus, the probability of breaking a particular exon at the next step of the process is not constant, but constantly decreases with the increase in the number of exons. When started from a single ancestral exon, this model of the multiplicative process, similarly to Kolmogoroff's version, produces a lognormal distribution of exon sizes. The resulting distribution of exon sizes obtained in this manner is independent of the actual sequence of random splitting events. After a sufficiently large number of steps, the mean position of the peak in the distribution of ln (*exon size*), M, shows an asymptotic linear shift with the logarithm of the number of splitting events, N_{spl} (for the process introduced here) or on time, t (for the Kolmogoroff process), $\ln E_0 - M \sim \ln N_{\mathrm{spl}} \sim t$, where E_0 is the length of the ancestral exon. Similarly, the peak width, σ, is related to the same quantities, $\sigma \sim \sqrt{\ln N_{\mathrm{spl}}} \sim \sqrt{t}$. The particular values of the coefficients for these proportionalities depend on the details of the process and can be ignored for our current purposes. However, we assume that the details of the splitting processes, i.e. the coefficients of proportionality, are the same for all species.

The assumption that the splitting probability is independent of exon size is essential for a lognormal distribution. A similar hypothesis, that selection of exons from open reading frames is independent of exon size for large exons (larger than a certain threshold of 105–110 bp), was discussed earlier (20). Our model assumes that the probability of exon splitting is independent of exon size for all exons, irrespective of any threshold. We note, in this regard, that the currently most common point of view is that the evolutionary process splits exons by intron insertions. As mentioned in the Introduction section, this generally acknowledged hypothesis of exon splitting by transposon insertion implicitly assumes a larger probability of splitting longer exons because they are larger targets for transposons.

In contrast to this, the lognormal distribution of exon sizes suggests that the mechanism of intron insertion should be independent of exon size. Obvious candidates for such a process would be mechanisms involving exon–intron boundaries, of which are always two for each exon.

A Kolmogoroff process provides a conceptual background for understanding the lognormal nature of exon size distribution. However, the distribution of exon length in real genomes is somewhat more complicated and generally reveals two lognormal peaks (Figure 1 and Supplementary data). To demonstrate how this two-peak pattern can be obtained for a random exon splitting process, let us consider a random process, similar to the one described before, initiated for a model ancestral genome that comprises two exons of unequal lengths. In this case, the two parts of the ancestral genome generate two distinct peaks in the exon size distribution. However, unlike previously, the resulting distribution of exon sizes is heavily dependent on the actual splitting pathway. Figure 2 demonstrates this fact for two splitting patterns generated from two sequences of pseudorandom numbers (see Methods section). This figure reveals a (spontaneous) symmetry breaking between the two peaks in the exon size distribution. In a certain sense, this phenomenon is similar to bifurcation, but, in contrast to an instant bifurcation event, this type of behavior can be regarded as a kind of 'soft' breaking of symmetry. The initial steps of the

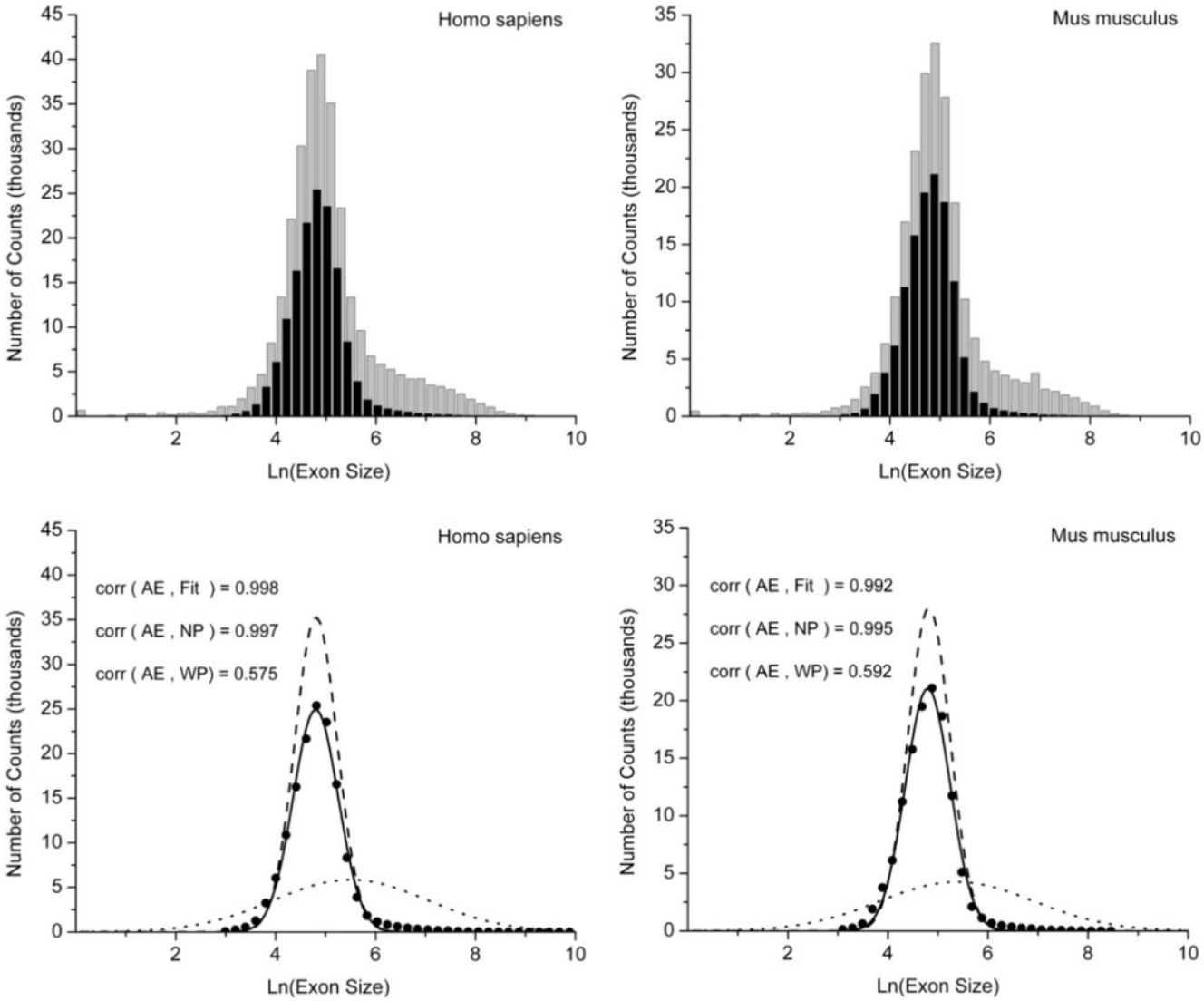

Figure 4. Top row presents comparison between statistical distributions of all exons (gray bars) with the distributions of alternatively spliced exons (black bars) from ASD. The data for *Homo sapiens* are in left panes; for *Drosophila melanogaster* in right panels. Panels in bottom row show the fit of the alternatively spliced exon length distribution to a single lognormal peak (full lines) as well as indicate locations of narrow (dashed lines) and wide (dotted lines) peaks obtained for the statistical distributions of all exons. Large dots indicate the observed distribution of alternatively spliced exon lengths. Inserts in bottom panels show values of correlations coefficients between the statistical distributions of alternatively spliced exons (AE) and fitted curve (Fit); narrow peak (NP) and wide peak (WP).

splitting process have a major impact on this phenomenon. Every subsequent splitting event has progressively less effect, and after ~5000 steps, the ratio between numbers of exons contributing to the two different peaks remains nearly constant. We believe that this phenomenon of symmetry breaking plays a major role in the diversity of patterns of exon size distributions.

The idealized model presented above can be reformulated in a different way where, instead of splitting exons on every step of the process, one randomly selected exon produces a descendant of larger size (a doublet, for example). From a formal point of view, these two variants of the process correspond to two possible directions for the time variable. From a biological prospective, they can be viewed as processes producing orthologous and paralogous gene families: the exons produced by duplications could be considered as paralogs, while exons derived from splitting as orthologs. Both variants of the model assume that the process starts from an ancestral genome with few exons, which are later split into smaller parts (or duplicated to produce longer parts). Thus, the proposed model, based on Kolmogoroff's ideas, supports the opinion that, at the initial stages of evolution, simple genomes had a lower fraction of introns (introns late hypothesis).

Figure 3 shows the parameters fit to the exon size distributions for 13 different genomes, using a two component lognormal model. All genomes show the presence of a narrow and a wide peak. Since peak width, σ, is a direct measure of the number of exon splitting events, our model suggests that exons in all of

Table 4. Distributions of exons between narrow and wide peaks together with peak positions, exp M, and peak widths, exp σ

Species	Narrow peak			Wide peak		
	exp M (bp)	exp σ (bp)	Fraction (%)	exp M (bp)	exp σ (bp)	Fraction (%)
Caenorhabditis elegans	115	2.03	22	185	4.32	78
Arabidopsis thaliana	90	2.21	30	233	5.76	70
Anopheles gambiae	159	1.73	12	252	6.63	88
Tetraodon nigroviridis	128	2.15	53	149	7.00	47
Bos taurus	125	2.16	52	99	7.82	48
Drosophila melanogaster	160	1.71	12	326	8.18	88
Canis familiaris	126	2.15	54	101	8.20	46
Gallus gallus	126	2.14	54	96	9.13	46
Danio rerio	126	2.15	53	90	12.05	47
Macaca mulatta	123	2.28	57	129	15.34	43
Mus musculus	124	2.31	64	217	21.98	36
Homo sapiens	122	2.33	63	245	22.06	37
Pan troglodytes	123	2.28	61	161	22.84	39

these genomes are distributed between two groups having different evolutionary histories. The width of the narrow peak is approximately equal for all species. The width of wider peak varies among species. For simple organisms, like *Caenorhabditis elegans*, its width is close to the width of narrow peak while for higher organisms, such as *Homo sapiens* and *Pan troglodytes*, it is about three times wider than the width of narrow peak. This may suggest both: that exons in this wider peak are older or are evolving more rapidly.

We argue that the biological origins of the exons contributing to the narrow and wide peaks are related to the appearance of new splicing mechanisms, such as alternative splicing (4). There recently has been great progress in the development of probabilistic methods for determination of exon boundaries (21–25). However, computational prediction of alternative splicing boundaries is still imperfect. Thus, to test the hypothesis that the observed distribution of exons between the two classes correlates with alternative splicing, we used the sets of alternatively spliced exons, which were manually curated and confirmed by EST/mRNA aligning (9,10). Unfortunately, curated datasets are available only for human and mouse, but there is a remarkable correlation between these data and the narrow peaks in *Homo sapiens* and *Mus musculus* (see data in Tables 1 and 3 and Figure 4). This strongly suggests that alternatively spliced exons are major contributors to the narrow peaks in all of the species examined here. This observation favors the hypothesis that the narrow peak is younger, since alternative splicing is a comparatively advanced evolutionarily mechanism. If one assumes that the evolutionary process modifies all exons in all the species with the same rate, the observation of nearly equal width narrow peaks in all species leads one to conclude that those peaks appeared at approximately the same and quite recent time. At least hypothetically, one may conclude that existence of the narrow peak in all discussed genomes is a manifestation of a spontaneous symmetry break in genome evolution, which is correlated with the evolution of alternative splicing.

The assumption that narrow peak is more recent does not completely eliminate the hypothesis that that this peak is more conserved. Indeed, one may assume is that exons contributing in narrow peaks are simultaneously more recent, and more conserved (with respect to the exon size distributions) than those contributing to the wide peak. In support of this point of view, one may note that the fraction of exons contributing to the narrow peak, in general, correlates with the width of wide peak (see data in Table 4). With two exceptions for genomes of *Anopheles gambiae* and *Drosophila melanogaster*, the increases in the fraction of exons in the narrow peak parallel the increases of width, σ, of the wider peak. In other words, its looks like the exons in the narrow peak are less frequently split by evolutionary processes (narrow peak width), but more frequently duplicated (larger fraction of exons) than are exons in the wide peak. This point of view is in agreement with the general pattern of alternative splicing in which a single copy of an exon is replaced by two (or several) isoforms of similar lengths.

CONCLUSIONS

The model presented here does not directly implicate a specific biological mechanism. However, we have shown that a very simple process, length independent splitting of exons, can produce the observed lognormal distribution of exon sizes. This suggests that it would be profitable to focus more attention on biological processes that are length independent, or on processes that could constrain the length of exons independently of exon length. Processes involved in mRNA splicing or associated with intron–exon boundaries are obvious candidates.

Furthermore, the observation that all eukaryotic organisms possess two exon length classes, one in common among all eukaryotes and one variable, suggests that exon splitting has played a key role in the evolution of eukaryotes. The difference in the peak widths of these two length classes suggests that the wider peak is older, or undergoes more rapid splitting, and mostly comprises nonalternatively spliced exons. The narrow peaks mostly comprise alternatively spliced exons, which are rather conserved in length but multiplied in number in more rapidly evolving genomes. The presence of a narrow peak

in all genomes examined could be explained as a manifestation of a 'spontaneous' symmetry break in genome evolution associated with the appearance of the mRNA splicing mechanism.

More detailed investigations, including characterization of conservation of exons in narrow and wide peaks between different species and between different subclasses within a particular genome, are extremely intriguing but beyond the scope of this article. The main goals of this work are to draw attention to statistical properties of exon size distribution and to highlight the utility of Kolmogoroff process model in understanding of genome evolution background.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Y.R. acknowledges stimulating discussions with Profs Yu. Feldman, D. Fushman, A. Sokolov and S. Mount, and expresses deep gratitude for constant support from Mrs. Natalia Grishina. Y.R. was supported by NSF award DBI-0515986 (M.G. PI) and, when working on the revision of this article by the National Research Council Associateship Program (Award # 0710430). Funding to pay the Open Access publication charges for this article was provided by NSF award DBI-0515986.

Conflict of interest statement. None declared.

REFERENCES

1. Gilbert,W. (1978) Why genes in pieces. *Nature*, **271**, 501–501.
2. Zhang,M.Q. (1998) Statistical features of human exons and their flanking regions. *Hum. Mol. Genet*., **7**, 919–932.
3. Roy,S.W. (2004) The origin of recent introns: transposons? *Genome Biol*., **5**, 251.
4. Brett,D., Pospisil,H., Valcarcel,J., Reich,J. and Bork,P. (2002) Alternative splicing and genome complexity. *Nat. Genet*., **30**, 29–30.
5. Curwen,V., Eyras,E., Andrews,T.D., Clarke,L., Mongin,E., Searle,S.M.J. and Clamp,M. (2004) The Ensembl automatic gene annotation system. *Genome Res*., **14**, 942–950.
6. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G.H., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al*. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res*., **31**, 224–228.
7. Press,W.H. (1992) *Numerical Recipes in C: the Art of Scientific Computing*, 2nd edn. Cambridge University Press, Cambridge, New york.
8. Matsumoto,M. and Nishimura,T. (1998) Mersenne Twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model Comp. Simul*., **8**, 3–30.
9. Thanaraj,T.A., Stamm,S., Clark,F., Riethoven,J.J., Le Texier,V. and Muilu,J. (2004) ASD: the Alternative Splicing Database. *Nucleic Acids Res*., **32**, D64–D69.
10. Stamm,S., Riethoven,J.J., Le Texier,V., Gopalakrishnan,C., Kumanduri,V., Tang,Y.S., Barbosa-Morais,N.L. and Thanaraj,T.A. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res*., **34**, D46–D55.
11. Snedecor,G.W. and Cochran,W.G. (1989) *Statistical Methods*, 8th edn. Iowa State University Press, Ames.
12. Collins,F.S., Lander,E.S., Rogers,J., Waterston,R.H. and Conso,I.H.G.S. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
13. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al*. (2000) The genome sequence of Drosophila melanogaster. *Science*, **287**, 2185–2195.
14. Balakrishnan,N. and Basu,A.P. (1995) *The Exponential Distribution: Theory, Methods, and Applications*. Gordon and Breach, Amsterdam, United States.
15. Gudlaugsdottir,S., Boswell,D.R., Wood,G.R. and Ma,J. (2007) Exon size distribution and the origin of introns. *Genetica*, **131**, 299–306.
16. Weibull,W. (1951) A statistical distribution function of wide applicability. *J. Appl. Mech-T. Asme*., **18**, 293–297.
17. Kolmogoroff,A.N. (1941) Concerning the logarithmic normal distribution principle of dimensions of particles during dispersal. *Cr. Acad. Sci. Urss*., **31**, 99–101.
18. Razumovsky,N.K. (1940) Distribution of metal values in ore deposits. *Cr. Acad. Sci. Urss*., **28**, 814–816.
19. Razumovsky,N.K. (1941) On the role of the logarithmically normal law of frequency distribution in petrology and geochemistry. *Cr. Acad. Sci. Urss*., **33**, 48–49.
20. Hoglund,M., Sall,T. and Rohme,D. (1990) On the origin of coding sequences from random open reading frames. *J. Mol. Evol*., **30**, 104–108.
21. Burge,C.B. and Karlin,S. (1998) Finding the genes in genomic DNA. *Curr. Opin. Struc. Biol*., **8**, 346–354.
22. Pertea,M., Lin,X.Y. and Salzberg,S.L. (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res*., **29**, 1185–1190.
23. Mathe,C., Sagot,M.F., Schiex,T. and Rouze,P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*., **30**, 4103–4117.
24. Zhang,X.H.F., Heller,K.A., Hefter,L., Leslie,C.S. and Chasin,L.A. (2003) Sequence information for the splicing of human Pre-mRNA identified by support vector machine classification. *Genome Res*., **13**, 2637–2650.
25. Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol*., **11**, 377–394.