Neutral Evolution of Duplicated DNA: An Evolutionary Stick-Breaking Process Causes Scale-Invariant Behavior

Florian Massip and Peter F. Arndt

Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany

(Received 20 November 2012; published 2 April 2013)

Recently, an enrichment of identical matching sequences has been found in many eukaryotic genomes. Their length distribution exhibits a power law tail raising the question of what evolutionary mechanism or functional constraints would be able to shape this distribution. Here we introduce a simple and evolutionarily neutral model, which involves only point mutations and segmental duplications, and produces the same statistical features as observed for genomic data. Further, we extend a mathematical model for random stick breaking to analytically show that the exponent of the power law tail is −3 and universal as it does not depend on the microscopic details of the model.

DOI: 10.1103/PhysRevLett.110.148101 PACS numbers: 87.10.Vg, 87.10.Ca, 87.18.Wd, 87.23.Kg

Ever since Susumu Ohno wrote his influential book to highlight the role of gene duplication in evolution [1], it has been well recognized that duplication and subsequent change of genetic material allow the exploration of evolutionary trajectories not accessible by point mutations only. Having completed the sequencing of the human genome, we know today that about 5% of primate genomes are composed of so-called segmental duplications often spanning tens of kbp [2,3]. The majority of those duplications are thought to have no direct function. In contrast to the very rich discussion about "the evolutionary fate and consequences of duplicated genes" [4], the destiny of duplicated nonfunctional DNA segments is in the majority of cases clear: they will dissolve into the genomic background by random mutations. However, as we show in this Letter, this dispersion process generates interesting statistical properties of the length distribution of identical sequence segments in genomes, which exhibits scale invariance with an integer exponent. We argue that this distribution is the characteristic mark of processes that are continuous and perpetual on evolutionary time scales and generate segmental duplications of genetic material and disperse them by random mutations into the genomic background.

Just after its duplication, a duplicated sequence segment will start out 100% identical to its original; subsequently random nucleotide substitutions and small scale insertions or deletions will break it into two and then more pieces, each being still identical to the corresponding segment in the original. This dispersion process can easily be observed in sequenced genomes when considering maximal segments of exactly matching nucleotides, i.e., copies of sequence segments that are equal over their entire length but differ on both ends. Such identical matches can easily be found using, for example, a gapless local alignment algorithm with infinite mismatch costs [5]. More advanced techniques employing suffix trees [6] or word counts are considerably faster for counting long and short segments, respectively. Independent of the algorithm, a self-alignment will include the global match along the diagonal of the alignment grid but will also show smaller (off-diagonal) matches representing duplicated segments along the sequence (see inset in Fig. 1 and the Supplemental Material [7] for more details on the computational procedures). For our purposes, we are not interested in the positions of those sequence segments but focus solely on their length distribution. As an example, we present the match length distribution (MLD) of the human genome in Fig. 1 [8,9]. We show two genomic distributions, before and after filtering for repetitive elements. Such elements cover about 45% of the human genome [9–11]. They have been copied into our genome multiple times in short bursts during evolution; their duplication dynamics is therefore remarkably different from the dynamics of (often) unique segmental duplications [3].

For small lengths of matching segments $r < 10$, the distribution is dominated by our neutral expectation for matching segments in random iid sequences (blue curve in Fig. 1). It is given by an exponential function in r

$$m_{\mathrm{iid}}(r) = \frac{1}{2} L^2 (1-p)^2 p^r, \tag{1}$$

where L is the total sequence length and p the probability of matching nucleotides, which is equal to $1/4$ for an iid sequence with equal proportions of nucleotides. Note that this exponential MLD leads to the well-known Gumbel distribution for *best* matches in an alignment of iid sequences, which is commonly used to assess the significance of local alignments [12,13].

Excluding duplications due to insertions of repetitive elements, the observed MLD carries an interesting

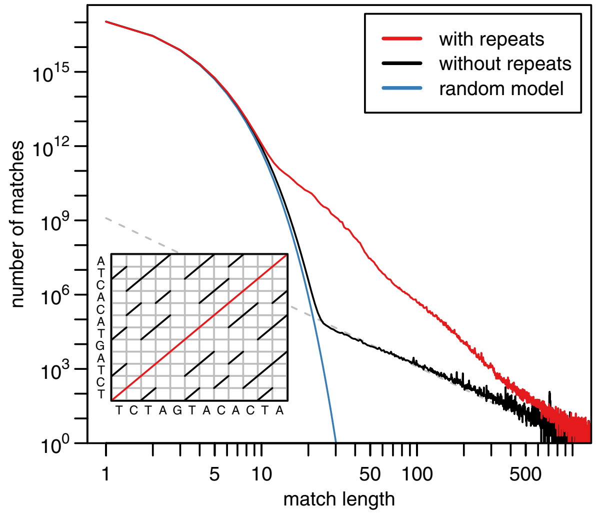

FIG. 1 (color online). The match length distribution for a self-alignment of the human genome. The MLD for the complete genome excluding repetitive sequences (in total $L = 1.23$ Gbp) is shown in black and shows the described power law tail. The MLD for a human sequence of the same length but including repetitive elements is shown in red. For small lengths both distributions coincide and are dominated by random sequence matches, which occur in randomly shuffled sequences (blue curve). The dashed line represents the function L/r^3, where r is the match length. The inset gives an example for an alignment grid of a self-alignment of a sequence of length 12. Matching nucleotides are marked by diagonal lines. The global alignment is shown in red along the main diagonal. Off-diagonal matches are depicted in black. The grid is symmetric and only matches above the main diagonal are counted. In this example there are six matches of length one, three matches of length two, and one match of length three.

statistical signature; i.e., it is well described by a power law with exponent -3 (black curve in Fig. 1), which can also be found in other species and was first reported by Gao and Miller [14]. Given this empirical result and considering that it only holds over one order of magnitude, it is reasonable to question if a meaningful model can be developed that also explains this finding mathematically; see Ref. [15] for a discussion on power law relationships in empirical data. However, although the genomic data might not be conclusive with respect to the exact functional form of the tail of the MLD, we will show in the following that a power law with exponent -3 can be understood through a simple sequence evolution model. The MLD of our model can be analytically computed using an integrated or time-averaged version of the stick-breaking model, a model which was first developed to describe the fragmentation process of polymers [16].

A sequence evolution model.—We introduce a sequence evolution model that includes two basic dynamic processes: point mutations and duplications of sequence segments. Both processes act on a sequence of nucleotides $A = (a_1, \ldots, a_L)$ of length L with $a_i \in \{\mathrm{A, C, G, T}\}$ representing the four possible states. The dynamics is Markovian and the mutation process changes the sequence $A \rightarrow A'$ at one random position k:

$$a'_i = \begin{cases} a' & \text{with } a' \neq a_i \text{ for } i = k, \\ a_i & \text{otherwise.} \end{cases} \tag{2}$$

This process happens with rate μ per site; i.e., in an infinitesimal small time interval dt it occurs with probability $\mu L dt$.

The second process in our model generates segmental duplications. A random segment of consecutive nucleotides of fixed length $K \ll L$ starting at a random position c in A, $(a_c, \ldots, a_{c+K-1})$, is copied and pasted to a random position v. The rest of the sequence stays unchanged; the new sequence A' is given by

$$a'_i = \begin{cases} a_{i-v+c} & \text{for } i \text{ with } v \leq i < v + K, \\ a_i & \text{otherwise.} \end{cases} \tag{3}$$

This process overwrites the K preexisting nucleotides a_{v+k} for $0 \leq k < K$ at the target site, and the total sequence length L stays constant. For simplicity we assume periodic boundary conditions and identify a_1 with a_{L+1}. Segmental duplications occur with rate γ per site, which is assumed to be smaller than the mutation rate.

Using only these two basic processes—mutations and segmental duplications—we will be able to generate sequences that exhibit power law match length distributions as they are observed in the human genome.

Match length distribution of the simulated sequences.—Given the above dynamics it is easy to simulate sequences and perform a self-alignment to find identical matching segments. We start each simulation with a random iid sequence with equal nucleotide frequencies. This sequence is then subjected to the above dynamics for a time long enough for each nucleotide to be mutated at least once on average and for a stationary state to be reached. The resulting sequence is then aligned to itself to find matching sequence segments. See the Supplemental Material [7] for computational details.

The MLDs for several simulations using generic parameters are shown in Fig. 2. The sequence length in all these simulations was $L = 10^6$. The distributions share the same behavior for small lengths, $r \lesssim 15$, which is dominated by the presence of small random matches whose distribution is an exponential as described in Eq. (1) with $p = 1/4$ and is plotted in the same figure.

Because of the continuous generation of segmental duplications, we also observe many more exact matches of length $r > 20$ than would be expected for random sequences. Interestingly, the length distributions follow a power law with exponent -3, as observed in genomic data from the human genome. Varying the parameters describing mutations and segmental duplications, μ and γ, does not change the exponent of the power law, i.e., the slope in

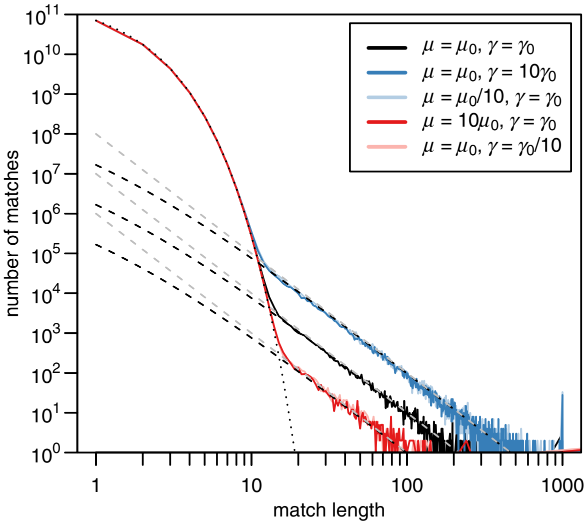

FIG. 2 (color online). The MLD for various values of μ and γ. For this plot we choose $\mu_0 = 0.1$, $\gamma_0 = 0.001$, $L = 10^6$, $K = 1000$. The two blue and red distributions are from sequences with the same ratio γ/μ. Using dashed lines we show theoretical predictions of the continuous (gray) and discrete (black) stick-breaking model. The dotted line represents the predicted MLDs for two random iid sequences of the same length.

the double logarithmic plot in Fig. 2. Observing the shifts of different distributions for varying μ and γ along the vertical axis, the prefactor of all distributions seems to be proportional to γ/μ. The other two parameters, the sequence length L and the length of segmental duplications K, have an influence on the prefactor of the power law tail but not on its exponent; see the Supplemental Material [7]. These observations can be explained by the following analytic considerations.

The stick-breaking process for the dispersion of a single duplication.—Let us first focus on the evolution of a single segmental duplication. Just after its generation we will find one full length 100% identical match of length K. This match will then be disrupted by mutations, which occur randomly in one of the two copies of the duplicated sequence. This fragmentation process of matching sequence segments can be mathematically described by the so-called stick-breaking process, a process that was once introduced to understand the fragmentation of long polymer chains [16]. In this framework, each segmental duplication is considered to be a full length stick, which is then broken up into smaller sticks. Following Ziff and McGrady [17], the dynamics of the length distribution of fragmented sticks in time can be solved analytically. For simplicity we assume that the stick length r is a continuous parameter and that mutations only break a stick without shortening it. If we denote the length distribution of sticks at time t after the segmental duplication event by $m(r, t)$, then it fulfills the following differential equation

$$\frac{\partial m(r, t)}{\partial t} = -2\mu r m(r, t) + 4\mu \int_r^\infty m(s, t) ds, \tag{4}$$

where the first term on the rhs represents the loss of matches due to mutations with rate μ in one of the two copies with total length $2r$. The second term describes the gain of matches due to a mutation in a longer match, which has to occur in one out of four possible locations, each being in distance r from one end of the two matches. At time $t = 0$ we start with one stick of length K, formalized by the initial condition $m(r, 0) = \delta(r - K)$, where $\delta(x)$ denotes the Kronecker delta function. The time dependent solution of this differential equation with this initial condition is known to be

$$m(r, t) = [4\mu t + 4\mu^2 t^2 (K - r)] \exp(-2\mu r t) \tag{5}$$

for $0 < r < K$, $m(r, t) = \exp(-2\mu K t)$ for $r = K$, and $m(r, t) = 0$ otherwise [17]. For large t and small r this is basically an exponential function in r.

The dispersion of a multiple duplication.—To finally understand the occurrence of power law tails in the match length distributions in simulated and genomic sequences, note that in these contexts we are likely to observe the remaining pieces of multiple ancient segmental duplications of different ages. Depending on their age, these segmental duplication will be broken into a different number of pieces. Assuming that duplications occur continuously we have to consider the time-averaged match distribution. Up to a normalization factor this distribution is equal to the integral

$$\int_0^\infty m(r, t) dt = \frac{K}{\mu r^3} \tag{6}$$

for $0 < r < K$ and equal to $1/(2\mu K)$ for $r = K$. This expression already shows the desired power law with the observed exponent. The appearance of a scale-invariant distribution in a process that is observed at different time points is not unexpected [18]. Surprisingly, in our integrated stick-breaking model the exponent is universal in the sense that it does not depend on the microscopic details of the model: the mutation and duplication rates.

The stationary state of a stick-breaking process with continuous duplications.—To deduce the correct normalization factor for the distribution [Eq. (6)], we consider a stick-breaking process, in which according to our evolutionary model segmental duplications of length K are continuously generated with rate γ per site. The dynamics in Eq. (4) for the distribution $m(r, t)$ then gains a third term on the rhs that describes the influx of new matches of length K in a system of size L,

$$\frac{\partial m(r, t)}{\partial t} = -2\mu r m(r, t) + 4\mu \int_r^\infty m(s, t) ds + \gamma L \delta(r - K). \tag{7}$$

In this setting we are interested in the stationary state distribution $m_\infty(r)$ and solve the differential equation (7) for $\partial m/\partial t = 0$. The solution is

$$m_\infty(r) = \frac{\gamma K}{\mu} \frac{L}{r^3} \tag{8}$$

for $r < K$ and $m_\infty(r) = \gamma L/(2\mu K)$ for $r = K$. Corresponding lines are shown in Fig. 2 and match our simulated data very well. We can match our observations even better by considering a discrete version of the stick-breaking model; see the Supplemental Material [7]. In essence these considerations yield a finite size correction to the asymptotic power law behavior in Eq. (8) for small r; see Fig. 2 for examples.

Considering the MLD per site, m_∞/L, the prefactor $A := \gamma K/\mu$ can be interpreted as the length of newly duplicated genomic sequence relative to the amount of mutated nucleotides. For the human genome this factor is close to one; see Fig. 1. This indicates that the amount of information that is ‘‘backed up’’ by segmental duplications is on average equal to the amount lost due to mutations. Note, however, that the spacial distribution of segmental duplications in the human genome is very complex and not specific to coding sequences. Therefore, this process might not save coding sequences from deterioration per se. For present day biological evolution, natural selection is probably a more powerful force to maintain and evolve genomic information over long periods of time.

Discussion.—We introduced a simple evolutionary model of segmental duplications and mutations that is able to give us insights into the occurrence of a power law tail in the length distribution of exact matches in self-alignments of genomic sequences. Using an extended version of the stick-breaking process for fragmentation we can also correctly deduce the empirically observed exponent -3 of this power law dependency. For the human genome, this tail comprises exactly matching sequences from length 25 to about 1000 bp; see Fig. 1. From our analysis we estimate that a total of about 50 Mbp (approximately 1.6% of the human genome) is part of at least one such match. The longest matching sequence segments are about 1000 bp in length, which suggests that the majority of segmental duplications spawn probably a few kbp, consistent with previous studies [2,3]. Furthermore, the prefactor of the power law tail in the MLD of the human genome is $A \approx 1$. From this observation, and assuming that mutations occur with a rate of about 1.5% per 10 million years [3], we can easily derive that a total number of about 4.5 Mbp per million years have been duplicated in the human lineage. This estimate agrees well with the one given in Ref. [3]. Assuming further that a typical duplication is 10 kbp long, the duplication rate γ would be of the order of 1.5×10^{-13} per bp and year. When restricting our analysis to coding sequences of the human genome, we find a similar power law tail with the same exponent -3 in the MLD; see the Supplemental Material [7]. Surprisingly, the prefactor A is about five, which is most likely due to a lower nucleotide substitution rate in these regions of the human genome.

Our model is very simple and uses fewer assumptions in comparison to a recently introduced model [19] that requires segmental duplications, whose length distribution needs to follow a power law from the outset. In our model the power law in the MLD can be derived without restrictive distributional assumptions and is solely generated through the interplay of the continuous duplication and mutation processes.

Further, we remark that, although similar in their definition of the underlying basic processes, our model is different from so-called expansion-modification models [20,21], which have been introduced previously to understand the observations of long-range correlations of the GC content along genomes [22]. The important difference is that in these models sequences as short as one nucleotide are inserted right next to its origin extending the total sequence length. This way no long matches are seeded. It can easily be tested that sequences that have been generated by an expansion-modification process [23] do not show a power law tail in the length distribution of identical matches. In this respect, these two phenomena, i.e., power law length distributions of matches and long-range correlations of the GC content, need not necessarily appear together.

Our results have also consequences for the assessment of the statistical significance of local sequence alignments of related species. The appearance of a power law tail in the length distribution of exact matches requires corrections to the Gumbel distribution of optimal scores in local sequence alignments. A quantification of the effects of segmental duplications on the score statistics will provide a better null model for local gapped alignment. However, because most segmental duplications occur in series and are interstitial [3], such corrections will probably only be relevant for regions where segmental duplications accumulated on evolutionary time scales.

In conclusion, we remark that in contrast to three-dimensional objects, which also show scale-invariant behavior in their fragment size distribution when broken [24], in our one-dimensional system, objects, i.e., segmental duplications, need to be continuously generated and broken up to give rise to the observed power law tail as a superposition of exponential distributions for different degrees of fragmentations. This condition of continuity seems to be sufficiently met for segmental duplications in the human lineage. This is not true for repetitive elements, which have been copied into our genome in irregular bursts. Therefore the match length distribution of the non-repeat-masked genome, which is clearly dominated by inter-repeat matches, does not have a power law tail with exponent -3; see Fig. 1.

The good fit of our model to empirical data is indicative of a constant accumulation of segmental duplications in the human genome, which are subsequently fragmented by random mutations during evolution. This process can mathematically be described by an extended version of the stick-breaking model, which explains the existence of the power law tail in the size distribution of fragments and the universality of its exponent with fascinating simplicity.

We thank Yves Clement, Navodit Misra, Dmitri Petrov, Martin Vingron, and especially Irina Czogiel for helpful discussions.

[1] S. Ohno, *Evolution by Gene Duplication* (Springer, New York, 1970).

[2] R. V. Samonte and E. E. Eichler, Nat. Rev. Genet. **3**, 65 (2002).

[3] J. A. Bailey and E. E. Eichler, Nat. Rev. Genet. **7**, 552 (2006).

[4] M. Lynch and J. S. Conery, Science **290**, 1151 (2000).

[5] T. F. Smith and M. S. Waterman, J. Mol. Biol. **147**, 195 (1981).

[6] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg, Genome Biol. **5**, R12 (2004).

[7] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.110.148101 for more details on the computational procedures, further simulations, and a derivation of the MLD in a discrete system.

[8] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh *et al.*, Nature (London) **409**, 860 (2001).

[9] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald *et al.*, Nucleic Acids Res. **40**, D84 (2011).

[10] J. Brosius, Science **251**, 753 (1991).

[11] J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz, Cytogenet. Genome Res. **110**, 462 (2005).

[12] S. Karlin and S. F. Altschul, Proc. Natl. Acad. Sci. U.S.A. **87**, 2264 (1990).

[13] S. Karlin and S. F. Altschul, Proc. Natl. Acad. Sci. U.S.A. **90**, 5873 (1993).

[14] K. Gao and J. Miller, PLoS ONE **6**, e18464 (2011).

[15] M. P. H. Stumpf and M. A. Porter, Science **335**, 665 (2012).

[16] W. Kuhn, Ber. Dtsch. Chem. Ges. **63**, 1503 (1930).

[17] R. M. Ziff and E. D. McGrady, J. Phys. A **18**, 3027 (1985).

[18] W. J. Reed and B. D. Hughes, Phys. Rev. E **66**, 067103 (2002).

[19] M. V. Koroteev and J. Miller, Phys. Rev. E **84**, 061919 (2011).

[20] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).

[21] P. W. Messer, R. Bundschuh, M. Vingron, and P. F. Arndt, Lect. Notes Comput. Sci. **3909**, 426 (2006).

[22] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).

[23] P. W. Messer and P. F. Arndt, Nucleic Acids Res. **34**, W692 (2006).

[24] L. Oddershede, P. Dimon, and J. Bohr, Phys. Rev. Lett. **71**, 3107 (1993).