Scale-free duplication dynamics: a model for ultraduplication

M.V. Koroteev and J. Miller∗
Physics and Biology Unit, Okinawa Institute of Science and Technology
Suzaki 12-22, Uruma, Okinawa 904-2234, Japan

Empirical studies of the genome-wide length distribution of duplicated sequences have revealed an algebraic tail common to nearly all clades. The decay of the tail is often well-approximated by a single exponent that takes values within a limited range. We propose and study here scale-free duplication dynamics, a class of model for genome sequence evolution that generates the observed shapes of this distribution. A transition between self-similar and non-self similar regimes is exhibited. Our model accounts plausibly for the observed form of the algebraic tail, which is not produced by standard models for generating long-range sequence correlations.

PACS numbers: 05.50.+q,

I. INTRODUCTION

The field of comparative genomics - of pivotal importance to medicine, biotechnology and the basic biosciences - is in large part the game of inferring functionality from sequence conservation. Its premise is that selective adaptation acts on neutral (sequence) variation[1]. If for any given sequence, it can be established that its conservation among diverse species is improbable on neutral sequence variation alone, then negative selection on function of the given sequence is inferred de facto. This premise underlies the ‘conservation tracks’ at the UCSC genome browser, for example[2]. Consequently, the choice of model for neutral genome evolution can have a major impact on the computational inference of whether or not a sequence is functional.

The discovery in the early 1990’s of long-range algebraically decaying two-point base correlations (LRC) in natural genome sequences[3] received wide attention in the physics literature, wherein several models of neutral genome evolution were proposed to account for it. Over the years the Li expansion-modification model[4] seems to have achieved the greatest visibility, in part because it generates algebraic correlations (a non-local effect) via a local genome growth dynamics; values of the exponents for this model have been derived analytically[5].

By the term “local” is here meant “local with respect to the linear genome sequence.” Obviously, higher-order chromosomal structure could lead to effects that are local in space, but non-local on the genome sequence; indeed, at the time two further proposals for the origin of LRC, one by Grosberg and co-workers[6] and another by Stanley and co-workers[7], invoked such structure; the former as a collapsed polymer globule, the latter as a self-avoiding (non-Gaussian) polymer, whose interior loop lengths follow a slowly-decaying algebraic distribution[8], generating within the proposed models a source of DNA segments that can be excised from and reintegrated into a chromosome. Perhaps because no direct evidence of non-locality was forthcoming, these proposals appear to have been largely eclipsed by the expansion-modification mechanism.

Our recent observation of a phenomenon widespread in natural genomes that we called “ultra-duplication,” namely an algebraic tail in the distribution of duplicated sequence lengths[9–11], appears to directly implicate a specific non-local component to neutral genome evolution. In this paper we examine numerically scale-free duplication dynamics (SDD), a class of simple neutral models for sequence duplication that can account for this distribution; we exhibit simulations that appear to rule out other simple neutral models; and present evidence suggesting that the observed power-laws constrain the form of non-locality. As we will explore below, qualitatively, and in some respects quantitatively, this evidence seems to corroborate the idea behind Stanley’s proposal[7] even though ultra-duplication is an orthogonal effect: it is not generated by Stanley’s model, and constrains the dynamics more strongly. It is worth stressing that the algebraic distribution of duplication lengths is not a Zipf law[12]; there is no invertible transformation between the duplication length distribution and a rank-ordered (frequency) distribution.

The paper is structured as follows: We define the notion of ‘maximal’ subsequence that comprises the sets of sequence elements whose length distributions we study. Two forms of SDD are then introduced, ‘copy/substitute’ (or ‘copy/paste’), and ‘copy/delete/insert’ (or ‘copy/indel’). Both involve duplicating sequences with length m chosen randomly from a source distribution whose form we take to be algebraic \( p(m) \sim m^{-\gamma} \) or monoscale \( p(m) \sim \delta (m - m_0) \).

We demonstrate that for \( \gamma > 2 \) the dynamics achieves a steady-state duplication length distribution with \( \gamma_{\text{eff}} \sim \gamma + 1 \); but for \( \gamma < 2 \) no steady-state is achieved and the long-time distributions are not algebraic. For monoscale sources, we show that a power-law is obtained at best over a limited range, and that within that range, \( \gamma_{\text{eff}} \) does not exceed 2.

A comparison of our numerical steady-state to the duplication length distribution of the genome from the

∗Electronic address: jm@oist.jp
bacterium *Anabaena variabilis* is exhibited. We observe that expansion-modification models appear to yield exponential rather than algebraic duplication length distributions. The discussion addresses the relationship of our scale-free duplication models to earlier models of neutral evolution, to the Stanley and Grosberg explanations of the observed powers in terms of polymer physics, and to parallel developments in the literature.

II. MAXIMAL SEQUENCES

A. Maximal $m$-mers, maxmers

The primary object of study here is the length distribution of maximal duplications within a genome or chromosome sequence. If a sequence is duplicated, then so trivially are all its subsequences, and it would be redundant to count these subsequences independently. A maximal duplication (referred to here as a “maxmer”) satisfies the condition that it is not a subsequence of a longer duplication. Maxmers represent the simplest way of counting sequence occurrences without trivial redundancy. Although more complex objects than maxmers can be defined, these definitions tend to be ad hoc [13, 14]. Given a chromosomal sequence, we compute all maxmers, and evaluate $\rho(m)$, their number as a function of their length. There are certain subtleties with respect to whether each distinct sequence is counted exactly once, or whether the number of copies of each sequence is reported; however, we have confirmed that for natural genomes and the simulations described here, the impact of varying these different ways of counting on the length distribution turns out to be minor and can be safely disregarded for our purposes here.

In evaluating the ‘biological relevance’ of maximal subsequences, it should be kept in mind that quantifying neutral sequence evolution is in part the task of characterizing sequence elements that are not necessarily functional. In this sense, maxmers represent an elementary measure of the sequence variation of any symbol string. Whether maxmers exhibit properties that are similar to or different from maxmers in random strings (and if so, why?) is a fundamental question worth asking about any discrete time series.

As described in [15], the so-called ‘ultra-conserved’ sequences are precisely the maxmers identically conserved among human, rat, and mouse genomes (sequences contained in all three genomes that are not subsequences of any longer sequence also contained in all three genomes). Biologists have weighed in extensively on their biological relevance [16]; their contributions to fitness - if any - remain controversial. The notion applied here and in [10] is the natural generalization of the concept of maxmer from inter-genome comparison to intra-genome comparison. Furthermore, we’ve demonstrated that the set of maxmers is essentially identical to the set of perfectly conserved sequences obtained from mouse whole-chromosome self-alignment for lengths greater than around 30 bases [10], and that self-aligned Hox gene sequences alone satisfy approximately the same algebraic length distribution as the full genome [10].

B. Methods

Practical hash-table methods for rapid identification of maxmers shared between two whole genomes have been reported elsewhere [15]. They are readily modified to identify maxmers occurring more than once within a single genome and to generate the length distributions reported here.

III. THE MODEL

We study a one-dimensional lattice (or ‘string’) of length $L$ on which each site is occupied by a letter from a specified alphabet. When the alphabet consists solely of the letters (or “bases”) ‘A’, ‘G’, ‘C’, or ‘T’, we call such a lattice a “chromosome.” In practice, we have established for the models discussed in this paper that for the quantities we study it makes essentially no difference whether a chromosome is thought of as composed of all four bases, or just the two bases ‘A’ and ‘T’; in fact, an arbitrarily large alphabet yields distributions indistinguishable for long maxmers from those reported here for two bases (data not shown).

A chromosome evolves in time according to a rule that introduces certain classes of random change into the sequence. The transformations on the chromosome are introduced in loose analogy to processes occurring in the evolution of real genomes. One such process, duplication, has been described often in the literature [13, 14, 17]; Others include substitution, insertion, and deletion [18]. Our observation that an algebraic distribution of duplication lengths is generic among organisms whose genomes are subject to recombination places strong constraints on the implementation of duplication within a model for genome evolution.

In one version of the model we propose here, the chromosome evolves by substituting a sequence at a randomly-chosen location within the chromosome for a sequence at another randomly-chosen location within the same chromosome; ‘copy/substitute.’ Restricting the substitutions to solely within the immediate neighborhood of the original sequence (tandem duplication) does not alter the exponent of the steady-state duplication length distribution [21]. In this paper we study the dynamics of one chromosome only. In real genomes with multiple chromosomes, sequences effectively can and do jump among them, a process that could involve additional parameters and different dependences on length.

In order to account for the data observed in real genomes, we find empirically that we must choose the
FIG. 1: (Color online) Two models of neutral evolutionary genome sequence dynamics.

A length of the copied fragment according to a source distribution that takes a power-law form – but it turns out that in steady-state the observed power is not the same as the source power. This difference is not unexpected, because the duplicated sequences interact: a new duplication can partially cover (or be inserted within) an older duplication. We also study monoscale duplication, where the length of the duplication is fixed to a single time-independent value. The latter can be viewed as a drawing the length of a duplication from a \( \delta \)-function distribution. Monoscale source distributions can within our models yield at best powers that are inconsistent with those most often observed in nature. Provided the mean duplication length of the source is fixed, broadening the range of the source leaves the steady-state unaltered, as indicated in the discussion and illustrated in the supplementary figures[22].

Our dynamics is as follows (see Fig. 1):


(i) Choose two locations \( i \) and \( j \) in the sequence uniformly and at random.
(ii) Take the subsequence from \( x_i \) to \( x_{i+m} \) where \( m \) is generated according to a specified and fixed distribution.
(iii.I) Substitute the subsequence chosen in (ii) for the sequence starting at position \( j \) of the chromosome.
(iv) go to (i).


(iii.I) above is replaced by

(iii.II) Delete the sequence at position \( i \) with probability \( p \); insert the sequence at \( j \) with probability \( q \).

Thus, the copy/indel dynamics reduces the sequence length by \( m \) bases with probability \( p(1-q) \), increases the length by \( m \) bases with probability \( q(1-p) \), and leaves the length unaltered with probability \( pq + (1-p)(1-q) \). For the simulations described here, \( p = q \), and the sequence length is constrained to remain within 20% of its value at \( t = 0 \).

The distribution according to which \( m \) is generated is chosen as a \( \delta \)-function:

\[
p(m) = \delta(m - m_0)
\]

where \( m_0 \) is fixed in time, or

\[
p(m) \sim \frac{1}{m^\gamma}.
\]

It is not difficult to see that a sequence consisting of homopolymer (e.g. all T’s or all A’s) represents a stable fixed point of these dynamics. There are several ways of eliminating this artifact, among them:

(a) introduce point mutations in the sequence with a rate \( \mu \) per base per iteration.
(b) with some probability \( \sigma \), take the reverse complement of the subsequence before re-introducing it into the chromosome.

Under reverse complementation, A \( \leftrightarrow \) T and G \( \leftrightarrow \) C to preserve base-pairing within the DNA double helix. Both (a) and (b) are active in real genomes.

Except where indicated explicitly, we take \( \sigma = 1 \) and \( \mu = 0 \) in the simulations described here. Provided the ratio \( L/\mu \) is much greater than than the mean duplication length, the effect of non–vanishing \( \mu \) is quantitative but not qualitative. For sufficiently large \( \gamma \), the dynamics is dominated by the effect of \( \gamma \), and for large enough \( L \) and evolution time \( t(L, \gamma) \), it turns out the \( \gamma \) is all that matters for the shape of the tail in the duplication length distribution - but as mentioned, it turns out that the power \( \gamma_{eff} \) characterizing the tail is not equal to \( \gamma \), but is instead close to \( \gamma + 1 \).

\[\text{(a) } \text{introduce point mutations in the sequence with a rate } \mu \text{ per base per iteration.} \]
\[\text{(b) with some probability } \sigma, \text{ take the reverse complement of the subsequence before re-introducing it into the chromosome.}\]
IV. UNCORRELATED RANDOM SEQUENCE

The initial condition for our simulations ($t = 0$) consists of uncorrelated random binary sequence of length $L$. We count maxmers – duplicated sequences not contained within longer duplicated sequences – so that the distribution of duplication lengths falls off at $m$ less than around $\sim \log_2(L)$, yielding a peak at that value. The probability of duplication of a binary sequence of length $m$, where each base is chosen independently and its two values are equally likely, is proportional to $(1/2)^m = \exp\{-m \ln(2)\}$, generating an exponential tail that appears as a straight line of slope $- \ln(2)$ on a semi-log plot.

[19] demonstrates that the length $M_L$ of the longest match in a binary sequence is approximated by $M_L \sim 2 \log_2 L$ as $L \to \infty$. Combining these two observations yields a prefactor $L^2$ so that the distribution of duplication lengths is $\propto L^2 \exp\{-m \ln(2)\}$ for large $L$. Fig. 2 shows on a semi-log plot the duplication length distributions for random sequence and several values of $L$, confirming these estimates.

V. SIMULATIONS

A. Power-law source

The evolution in time of the length distribution of maximal $m$-mers under copy/substitute dynamics from a random initial condition at $t = 0$ for two different values of $\gamma$ is illustrated in Figs. 3 and 4.

The figures show the time evolution under copy/substitute dynamics of the length distribution of maximal $m$-mers, starting from a random sequence at $t = 0$. At finite times, the distribution can be for convenience divided into two parts, one composed of short lengths $m \lesssim \Lambda$, the other for $m \gtrsim \Lambda$, where $\Lambda \sim \log_2(L)$. The left part corresponds to the exponential distribution (Fig. 2) and the right part emerges under evolution of the system. For $\gamma = 2.4$ (Fig. 3), the distribution at finite times is at short lengths nearly indistinguishable from random sequence ($t = 0$). For larger $m$, the part of the distribution that we refer to as the ‘tail’ takes an algebraic form. Evidently for $\gamma = 2.4$, the curves evolve to a stationary state at long times.

On the other hand, for $\gamma = 1.8$ (Fig. 4), we have so far not been able to achieve a stationary state, and the curves are generally not very straight at large $m$. Furthermore, the height and location of the peak changes with time. If $\gamma_{\text{eff}}$ is the (effective) exponent describing the observed tail in steady-state of the dynamics (where the symbol $\gamma$ is reserved for the tail of the source distribution), then for $\gamma_{\text{eff}} \leq 2$, the total length of duplicated sequence no longer scales in direct proportion to $L$, namely

$$\frac{1}{L} \sum_{m=1}^{L} m (1/m^{\gamma_{\text{eff}}}) \sim \int_{1}^{L} (1/m^{\gamma_{\text{eff}}}) \, m \, dm$$
as the distribution may no longer be well-described by a single power over most of its range, or by the same power at different times.

The effective power $\gamma_{eff}$ derived from the the length-distributions for various $\gamma$ [20] is shown in Fig. 6 as a function of $\gamma$. It is remarkable that in real genomes, powers less than around $\gamma_{eff} = 3$ are exceptional [11].

\subsection*{B. Scale-invariance}

The general form of the tail of the distribution as $t \to \infty$ can be represented as $P = P(\gamma, m, L)$. At fixed $\gamma$, dividing by the length of the longest simulated sequence, $L_{max}$, rescales the $x$-axis to dimensionless form and collapses the curves for different $L$. We obtain the following invariant form of the length distribution for all $L$:

$$P(\gamma, m, L) \sim A(\gamma) \frac{L}{L_{max}} P_{L_{max}} \left( \frac{m}{L_{max}} \right).$$

Here $A(\gamma)$ is an a priori undetermined function that can be evaluated from the simulation. The outcome of computations based on the above formula is shown in Fig. 7. In the graphs we set $L_{max} = 10^8$. We find empirically that curves with differing values of $\gamma$ can’t be collapsed onto one another.

\subsection*{C. Tandem duplication}

A duplication may instead be restricted to occur near its source; e.g. a sequence of length $m$ beginning at base $i$ is substituted for another sequence beginning not at a randomly chosen location, but instead at location $i + m$, a form of tandem duplication. The exponents of the duplication length distributions obtained with tandem duplication only are unchanged from those obtained when...
the substitution occurs at locations uncorrelated with the source[21].

D. \(\delta\)-function source

If we assume that the only relevant length scales are the duplication length \(D = m_0\) and the chromosome length \(L\), then

\[
(D/L)\rho(m) \, dm = (D/L)\rho(Dm/D) \, dm
\]

\[
= (D^2/L)\rho(Dx) \, dx = \hat{\rho}(x) \, dx
\]

where \(0 \leq x \leq 1\), so that rescaling \(m \rightarrow x = m/D\) and \(\rho \rightarrow \hat{\rho}(x) = (D^2/L)\rho(Dx)\) should collapse steady-state distributions obtained for different \(D\) and \(L\). Figure 8 shows the steady-state maximal \(m\)-mer distribution for copy/substitute dynamics and a broad range of \(D\). Figure 9 shows data from 8 rescaled as indicated in the text, together with data for copy/indel dynamics where both \(D\) and \(L\) vary.

Broadening the variance of the source distributions at fixed mean length, for example drawing from a single Gaussian or a double-sided exponential source, yields steady-state distributions indistinguishable from the true monoscale source[22]. A Poisson source distribution \(\propto \exp(-m/\lambda)\) yields a tail that decays as \(\exp(-m/\lambda)/m\), i.e. faster than a pure exponential[23].

E. Comparison of steady state to natural genome

Figure 10 illustrates duplication length distributions for exact and reduced stringency self-intersections of the cyanobacterium *Anabaena variabilis* genome and the stationary state of a 4-base version of copy/indel dynamics on a single chromosome of the same length as the cyanobacterium genome, around \(6.5 \times 10^6\) bases. Given a single 4-base chromosome, in the 4 base intersections the four nucleotides \(A, G, C,\) and \(T\) are treated as distinct symbols whereas in the 2 base intersections, \(A\) and \(G\) (respectively, \(C\) and \(T\)) are taken as equivalent. Here \(\sigma = 0.5\), \(\mu = 2 \times 10^{-5}\), and a source distribution of the form \(1/(1 + (x/a)^{\gamma})\) with \(a = 16\) and \(\gamma = 2.4\) is applied. As in real genomes, the point mutations are such that transitions (\(A \leftrightarrow G\) and \(T \leftrightarrow C\)) are more likely...
than transversions; within our model, this difference in rates is essential to obtain the observed displacement of the 4-base versus 2-base curves in the algebraic regime. The correspondence is, we believe, suggestive.

F. Li expansion-modification model

Around twenty years ago, Li [4, 24] introduced an expansion-modification dynamics for genome sequence growth that incorporates short substitutions, insertions, deletions and local duplications. This local model yields algebraic decay of the spatial power spectrum corresponding to the (averaged) base-base correlation function, which may also be measured by the DFA exponent [25], \( \alpha \). In this class of model both duplications and insertions are at most a few bases long, and duplications always occur near their source, e.g. the copy is inserted directly next to the original. LRC was observed in natural genome sequences shortly after Li described his model, and has been intensively studied ever since [26, 27]. Within Li’s model, growth of the sequence is essential to produce these algebraic power spectra. More recently, Arndt and collaborators [5] have derived the exponents for this model analytically as a function of the rates of (small scale) substitution, insertion, deletion, and duplication.

We investigated for a variety of parameter values whether or not the expansion-modification dynamics could produce the algebraic length distributions observed in the data and in our simulations of SDD, with the package CORGEN [28]; the on-line [29] and stand-alone version (kindly provided to us by P.W. Messer) yielded identical outcomes.

The duplication length distribution generated by CORGEN was invariably exponential, as illustrated in Fig. 11; nevertheless, we obtained the theoretically expected exponent for the algebraic decay of the power spectrum for the LRC (Fig. 12b). In contrast, our model yields a white power spectrum (Fig. 12a).

In a further elaboration of his original idea, Li indicated how oligomeric repeats could contribute to LRC [30]; however, these duplications are explicitly short, whereas our model relies on a scale-free distribution of duplication lengths.

G. Summary

Under SDD, which yields exponentially decaying two-point base correlations, (i) the length of sequence does not grow; \( L \) is either fixed, or permitted to fluctuate within a small range of a fixed value; (ii) there is no intrinsic scale to the duplication lengths (this is the non-locality); (iii) duplications are not constrained to occur near their sources; the location of the copy does not depend on the location of the original.

Contraints on the location of the copy relative to the original can be applied; for example, requiring that all duplications be tandem; however, we find empirically that such constraints don’t alter the exponent of the length distribution from what is observed for random relative locations of the duplicates. We were at first surprised by this observation, but part of the value of studying the length distribution turns out to be its insensitivity to certain details. Therefore, condition (iii) appears essential to any conclusions we arrive at for our model here.

These differences, together with the discussion of the Stanley models below, indicate that our dynamics does not reduce to previously studied evolutionary dynamics.
FIG. 12: (a) Normalized power spectrum of the sequence after $10^8$ time steps of copy/substitute dynamics. $\gamma = 2.4$, $L = 10^8$. (b) Normalized power spectrum of the sequence obtained from CORGEN [28] for the parameters indicated in fig. 11. The straight line shows the theoretically expected slope corresponding to the algebraic decay exponent of the power spectrum[5] for the same parameter values.

One could say that the dynamics under comparison are orthogonal: Our models yield an algebraic tail of the distribution and white power spectrum for correlations; the Stanley and Li models yield exponential decay of the distribution and an algebraic power spectrum for correlations. Real genomes exhibit algebraic decay of length-distributions of maximal $m$-mers in addition to an algebraic power spectrum for correlations.

VI. DISCUSSION

A. Duplication is fundamental to genome evolution

Although the fundamental role of sequence duplication in genome evolution is long appreciated[17, 31–33], the first quantitative model-independent characterization of which we are aware is the recent work of Lee and co-workers on ‘genomic equivalence length’ [34] based on the study of $m$-mer entropy ($m \leq 9$) of modern natural genomes, which stresses the dominant role of duplication in genome growth. There exist other characterizations that may be more readily interpreted in terms of currently understood biological sequence type and function[13, 14], but they are model-dependent and may therefore not always be the most suitable tools for discovery of novel repeats and functional elements.

Lee et al. also investigated models of genome growth with monoscale ($\delta$-function) duplication lengths [34]; however, they did not study the distributions of longer $m$-mers, which constitute the focus of our work in general and this manuscript in particular; the calculations described above appear to eliminate monoscale duplication models as candidates for neutral evolution of most natural genomes. Lee et al. characterize nature as "the blind plagiarizer," duplicating sequences within a genome at random and letting selection determine what sticks. Our observations suggest that nature is in this sense blind also to scale. This seems in retrospect very reasonable, since the lengths of functional genomic elements span a very broad range, from short regulatory elements to operons comprised of many genes: it is unclear how a single characteristic scale could economically allow for duplications over this whole range. On the other hand, if duplication rate were wholly independent of duplication length, duplications would be overwhelmingly weighted toward lengths on order of the chromosome size.

B. Stanley generalized Levy walk (GLW) and related models for LRC

Stanley introduced several explicitly non-local models to account for the observation of long-ranged algebraically-decaying two-point base correlations in natural genomes.

One of them, the GLW model[35], involves successively appending random sequences of lengths drawn from an algebraic source distribution. This model is unphysical, but enables analytical computation of the correlation decay exponent, $\beta$, directly from the source distribution exponent, $\mu$. Stanley et al. propose that $\mu$ for natural genomes corresponds to the distribution $m^{-\mu}$ of interior looping segment lengths $m$ within a self-avoiding long polymer at equilibrium[8, 35]. For $d = 3$, $\mu \simeq 2.22$; it is claimed this value recapitulates the DFA exponent $\alpha$ observed in eukaryotic genome sequences, which can be computed trivially in terms of $\beta$. The analytical calculation is readily accomplished because the dynamics is comprised by successive concatenation of the sequences, which therefore effectively occur independently of one another; they don’t interact.

A more realistic model was proposed in [27], and involves the deletion and subsequent reinsertion at an uncorrelated location elsewhere in the chromosome, of segments chosen according to a source length distribution with exponent $\mu$. In addition, it is crucial that occasion-
ally, instead of reinserting the deleted segment, a random sequence of biased base composition of the same length be inserted in its place, generating a source of composition bias. Although so far not analytically calculated because of the complexity introduced by interaction (insertion or deletion), \( \beta \) may be computed from \( \mu \) via numerical simulation of the model.

There are two major problems in reconciling these models with natural sequence data: (i) Within the models, the DFA exponent achieves a steady-state value that depends only on \( \mu \), whereas natural genome sequences show DFA exponents that vary over a considerable range; (ii) There was at the time minimal, if any, direct evidence for scale-independent transfer of sequences within genomes[39] and therefore invoking the parameter \( \mu \) could not be independently validated.

C. LRC and the collapsed polymer globule

Grosberg’s 1993 prediction of the collapsed polymer globule structure of chromosomes[6] seems recently to have been confirmed by experiment[37], where what amounts to a value of \( \mu = 1 \) for the collapsed polymer globule is also derived. Based on a predicted typical structure, Grosberg derives a value of 2/3 for the DFA exponent \( \alpha \) that is broadly compatible with eukaryotic chromosome sequence data[38].

One difficulty in reconciling this proposal with Stanley’s is that the collapsed polymer globule structure and the conditions of the experiment[37] involve metaphase chromosomes and not meiotic or mitotic chromosomes; the latter, where deletion, hopping, and reinsertion of sequence elements are generally believed to occur[39], are widely believed to take on quite different structures than metaphase chromosomes.

D. Alternative mechanisms

Although both Stanley’s and Grosberg’s proposals could account for LRC, neither of them lead to a power-law distribution of duplication lengths[40]; however, it is believed that common pathways involving the bacterial protein recA and its eukaryotic homologs mediate processes involving duplication, insertion, and deletion of relatively long sequences that constitute ‘homologous recombination’[14], among them, gene conversion and unequal crossing-over. Thus, it is natural to conjecture that (a) a common power-law governs homologous recombination; and (b) this power-law would be reflected in LRC.

Observation of a duplicate sequence within a genome reflects several independent steps[41]:

s1. Duplication of the sequence or of a (longer) sequence containing it;

s2. Preservation of the duplicate, by selective or neutral processes, both of whose actions are incompletely understood;

s3. Experimental observation of the duplicate.

Our model is not intended to reflect particularities of either (s1) or (s2) alone, but the net singular behavior of the duplication distribution only. Thus, specific roles for transposable elements, homologous and illegitimate recombination, or RNA-mediated duplication are not readily inferred. Avenues for assessing specific contributions of (s1) and (s2) could be worth pursuing.

Bias in (s3) remains the major obstacle to understanding sequence duplication. Duplications are the most difficult-to-assemble parts of genome sequences; sequencing is itself expensive enough that our primary sources of information on laboratory time scales are indirect methods involving targeted search for specific duplications that introduce ascertainment bias (e.g. for or against gene sequences). This situation is likely to change dramatically over the next few years, as it becomes feasible to obtain sufficiently high quality whole genome sequences of sets of individuals from a single population.

As we study here duplicated sequences within a single sequenced genome only, the net impact of (s1-s3) is all that is observable. To lowest order, there appears to be universal agreement that (s1) is neutral; duplication occurs independent of functionality. Except perhaps for selection on total genome length, during (s2) selection presumably acts by definition solely on functional sequence. One might anticipate that studying matches independent of their function yields a more faithful reflection of neutral variation than than does studying duplicated genes alone. On the other hand, distinguishing the impact of selective versus neutral processes in (s2) remains one of the central challenges of comparative sequence analysis. There being no known or previously hypothesized selective mechanisms that would yield an algebraic distribution of duplicate sequence lengths, other sources seem worth exploring.

One candidate for a suitable neutral mechanism, the form of homologous recombination known as ‘gene conversion’ – a somewhat misleading term as it applies whether or not a sequence is part of a gene – is especially attractive as it is believed to operate similarly among homologous loci within a genome (inter-allelic), as it does between homologous loci of parental genomes (intra-allelic). Gene conversion refers to the non-reciprocal transfer of a tract of homologous sequence. The distribution of gene conversion tract lengths is conventionally taken to be exponential, but the range and resolution of the data in support of an exponential form are not sufficient to draw this conclusion[42]. We conjecture that an algebraic distribution of gene conversion tract lengths could contribute to the algebraic distributions observed for both duplicated sequence lengths and strongly-conserved sequence lengths.
E. Comparative sequence analysis

In practice, a “null model” for neutral drift forms the basis for determining whether conservation of a given sequence is stronger than would be expected on neutral drift alone. Stripped to their cores, existing methods of comparative genomics assess whether the probability of an observed sequence similarity, with a given length and fraction of identical bases, is sufficiently low within the null model. Standard null models are invariably up to inessential distinctions independent-site point mutation models, in which the likelihood of a change at one location in the sequence is independent of whether or not there is a change at a distant location. In other words, the null models are local.

In part for this reason, any non-locality in neutral drift could have an significant impact on inference that any given degree of conservation implies sequence functionality. Our research has identified two a priori independent indications of non-locality. The first was an algebraic length distribution of sequences perfectly or strongly conserved among multiple diverse genomes (correlations of conservation) - essentially linkage disequilibrium - which we initially interpreted as indicating selection[15]. More recently we described evidence that neutral effects are probably also major contributors[9]. The second was an algebraic length distribution of perfect or approximate duplication lengths within a single genome[10, 11]. The connection between these is the low probability that such long runs of identity could be generated by local neutral processes, such as point substitution or small insertions and deletions. Duplications are believed to be generated via recombative mechanisms that act on a contiguous chunk of sequence as a whole. Whether or not they persist after having first been generated, while in part a matter of natural selection and population genetics, can also be governed by the set neutral mechanisms referred to as 'concerted evolution.'

As always within the framework of comparative sequence analysis, the only observable is the net impact on sequence; disentangling neutral drift from selection can be accomplished only indirectly. For example, neutral substitution rates are inferred from existing natural genome sequences by identifying sequence elements that are believed (sometimes incorrectly) no longer to be under selection, such as pseudogenes or obsolescent transposable elements; computing the number of substitutions between aligned homologous elements; and coupling the latter to some estimate of how long ago the genomes branched from a common ancestor. An inappropriate model of neutral drift could bias several of these steps unpredictably. A universal contributor to neutral drift having recently been identified[9, 10, 15], our contribution in this manuscript is a phenomenological dynamical model for its origin.

VII. CONCLUSIONS

Long-range algebraic sequence correlations (LRC) have been the focus of intense research over the last two decades, in part because of their potential relationship with isochores[24, 26]. Several distinct models for their origin have been proposed in the physics literature. We recently described a complementary and more widespread property of genomes, namely an algebraic distribution of duplicated sequence lengths, which in contrast to LRC has the attractive feature that each of the duplicated sequences contributing to the distribution can be separately extracted from a genome and studied on its own merits, independently of the others. We have proposed and investigated numerically here a scale-free duplication dynamics (SDD) model that we show can account for a number of distinctive properties of this algebraic distribution that we have observed in natural genomes. The model we investigated is complementary to earlier proposals for LRC by Li[4, 5], Grosberg[6], and Stanley[7] that don’t generate an algebraic distribution of duplicated sequence lengths.

Li and Kaneko assert that their expansion-modification model accounts for the LRC observed in natural genome sequences[43]. It is certain that short indels contribute significantly to genome evolution, although they may be biased toward deletion - e.g. toward shrinkage rather than growth[44]. One can also view the expansion-modification primarily as a toy model that generates long-range algebraic correlation without including it as part of the model, e.g. within a purely local dynamics. It is plausible that if expansion-modification models were generalized so that the duplicated block lengths were not limited to the scale of a few bases, and were instead chosen from a long-tailed distribution, then the maxmer length distribution at steady-state would also exhibit a long tail - but then their dynamics would no longer be purely local.

Together with our calculations, the observation of algebraic duplication-length distributions suggests to us that proposed expansion-modification models may be of limited value in interpreting LRC in natural genome sequences. On the other hand, the phenomenology of expansion-modification models remains to be fully developed[45]; were it possible to generate algebraic distributions of duplication lengths similar to those observed in natural genomes within a (purely local, by our criterion) expansion-modification model, it could represent a significant theoretical advance.

Although we don’t know how the (local) Li expansion-modification dynamics could be modified to produce such a distribution without incorporating non-locality into its dynamics, the Stanley model is explicitly predicated on non-locality introduced by assuming an algebraic distribution of deleted and inserted sequence lengths, with exponent $\mu$. SDD assumes an algebraic source distribution for duplicated sequence lengths, with exponent $\gamma$, and simulations suggest that in the absence of this alge-
braic source it is not readily possible to recapitulate observed sequence data. Thus, together with the data, SDD provides independent evidence of an essential non-local contribution, identifying for the first time an observable algebraic length-distribution for various gamma.

VIII. ACKNOWLEDGEMENTS

Eddy Taillefer computed the maximum likelihood fits to a power–law in the figures, and wrote software (SEQ-ANALYSIS) that we used for some of the maximal m-mer computations. We gratefully acknowledge Dr. Philipp W. Messer for sharing with us his CORGEN source code. We thank Sydney Brenner for stressing the fundamental role of linkage in our earlier observations[15]; Holger Jenke-Kodama for suggesting cyanobacterial genomes; Alexander Mikheev, Sydney Brenner, Holger Jenke-Kodama, Ting Wu, UlI Skoglund and Sathish Venkatesan for valuable conversations; the three anonymous referees for their improvements to the manuscript. While this manuscript was under review, we analytically reproduced many of the simulated outcomes here reported.

[21] See Supplemental Material(figures 2, 3) at
[22] See Supplemental Material(figures 4, 6, 7) at
[23] See Supplemental Material(figures 5) at
[29] http://corgen.molgen.mpg.de/
quence; however, has in fact been demonstrated that once constraints on synonymous codon variation are properly accounted for, a DFA exponent close to the value for introns emerges instead[46], suggesting that the non-random DFA values for both coding and non-coding sequences have the same origin. Similarly, both coding and non-coding duplications follow the same length distribution[10].


[40] More recently, Stanley and co-workers introduced an unequal crossing-over model to explain the length distribution of simple repeats; see e.g. S.V. Buldyrev, N.V. Dokholyan, S. Havlin, H.E. Stanley, R.H.R. Stanley, Physica A 273:19 (1999) and references contained therein.


