Disorder and Power-law Tails of DNA Sequence Self-Alignment Concentrations in Molecular Evolution

Jian-Zhou Zhu¹, Kun Gao², HongGuang Sun¹,³,
¹Su-Cheng Centre for Fundamental and Interdisciplinary Sciences, Gaochun, Nanjing 211316 China and Li Xue Center, Gui-Lin Tang Lab., 47 Bayi Cun, Yong an, Fujian 366025 China
²Physics and Biology Unit, Okinawa Institute of Science and Technology (Graduate University), Okinawa
³College of Mechanics and Materials, Hohai University, Nanjing 210098, China

The self-alignment concentrations, \( c(x) \), as functions of the length, \( x \), of the identically matching maximal segments in the genomes of a variety of species, typically present power-law tails extending to the largest scales, i.e., \( c(x) \propto x^{-\alpha} \), with similar or apparently different negative \( \alpha \)s \( (< -2) \). The relevant fundamental processes of molecular evolution are segmental duplication and point mutation, and that recently the stick fragmentation phenomenology has been used to account the neutral evolution. However, disorder is intrinsic to the evolution system and, by freezing it in time (quenching) for the setup of a simple fragmentation model, we obtain decaying, steady-state and the general full time-dependent solutions, all \( \propto x^{-\alpha} \) for \( x \to \infty \), which is in contrast to the only power-law solution, \( x^{-3} \) for \( x \to 0 \) of the pure model (without disorder). We also present self-alignment results showing more than one scaling regimes, consistent with the theoretical results of the existence of more than one algebraic terms which dominate at different regimes.

PACS numbers: 87.10.Vg, 87.10.Ca, 87.18.Wd, 87.23.Kg

INTRODUCTION

The effects of duplication and mutation are crucial for the genome evolution dynamics and the generation of biodiversity (see, for example, Ref. ¹ for an overview of theory and mathematical models along with practical examples). The dynamics of duplication must be different, compared to other processes, such as recombination, thus it is helpful to isolate its fingerprints in the genome sequences, say, by masking simple repeats ², from the data for separate studies, say, the neutral evolution dynamics, among the various debatable considerations (cf. the recent dialog ³). One can also try to obtain information about life, concerning disease susceptibility and paralog v.s. ortholog issue etc., from studying the duplication and mutation (see, e.g., ⁴, ⁵). Fig. ¹ presents some examples of the concentration (histogram) \( c(x) \) of the maximal segments of exactly matching nucleotides as a function of the match length \( x \), from both eucaryotic and prokaryotic species. We see that the scaling exponents can be similar or apparently different for a variety of species; the power laws may not be the same for two distinct chromosomes of the same species (c.f. more details in the caption.) Massip and Arndt ⁷ computed and took the exponent of the repeat-masked whole genome sequence to be exactly \(-3\), a typical value for some specific chromosomes of various eukaryotic species ⁸, ⁹, and they also showed that the repetitive elements ² greatly deteriorate the scaling law. Li et al. ¹⁰, ¹¹ recently also discovered other relevant forms of power law distributions. It thus appears to us that a reasonable model, especially that from the null hypothesis of neutral molecular evolution, should present the power-law tails at large scales as well as the different possible scaling exponents with a common or somewhat “universal” mechanism among them.

Massip and Arndt ⁷ used an analytically tractable model with the pure fragmentation phenomenology (Koroteev and Miller ¹² had done simulations with descriptive procedures containing some of the essential features, and they found later similar results ¹⁴.) The scenario by these authors was the following: The maximal matching segments, defined by copies of nucleotide sequence that are the same but are different when extended oneither ends, come from duplications subject to mutations. Duplication (“Step 1” in Fig. ²) of a single sequence produce exactly matched segments of the same length, and random mutations “break” them into matching pieces of shorter lengths (Step 2 in Fig. ²). By suitably assigning the (constant) mutation rates and the (linear) balancing of gains and losses at each scale, they obtained the fragmentation model well studied in other disciplines: The solution to the model with a constant input at a fixed scale \( K \) (“monodispersion” contains a part scaling with the exponent \(-3\) as was also found earlier by Ben-Naim and Krapivsky ¹³ who also pointed out that the \(-3\) scaling was actually a small-scale asymptote, i.e., \( c(x) \propto x^{-3} \) as \( x \to 0 \) (compared to the system size, say), for the “head” instead of the “tail” \( x \to \infty \) compared to the number of alphabets, say,) of the concentration. Such ‘head’ versus ‘tail’ issue can be effectively distinct for the dynamics and thus calls for further studies. Although the notions of being ‘large’ compared to some small scale and being ‘small’ compared some large one do not directly conflict, we will show that physics of the scaling law are different.
FIG. 1: Concentrations, computed by MUMmer, of maximal matching segments in the self-alignments of the genome sequences of various species, shifted apart for better visualization, showing similar or apparently distinct scalings extending to the largest scales: All data, except those in the inset whose plot of human genome reproduces with Chromosome 1 the results of Massip and Arndt, are shown only for scales above 20bp below which there is no power law and the computation is also very expensive. Eucaryotic species, unlike the procaryotic ones, are in general seriously affected by simple repeats as shown by the data of homo sapiens (human) where the very small scale range is also presented to show that the masked data is basically exponential due to random matching from the finite number, now 4, of the alphabets and coagulation effects (simple equal probability assumption leads also to exponential distribution of coagulation effect, just as the random matches). Dashed lines are exact scaling laws for reference.

FIG. 2: The example of duplication and mutation, and, the fragmentation and coagulation effects.

Fragmentation and coagulation effects from duplication and mutation

Let’s call an $x$-length segment $x$-matching or $x$-unmatching, depending on whether it belongs to the set of matching segments of length $x$ or not. A mutation may have two-fold effects on the change of matching segments: One obvious effect is fragmentation, i.e., an $x$-matching segment is broken into shorter segments and become $x$-unmatching: In “Step 2” of Fig. 2 “GAGGCTATGT” fragments due to the mutations of “G” to “T” and “A” to “C” respectively; the other opposite effect, not discussed previously, is coagulation, i.e., a point mutation may unite its two sides to become a longer matching: In “Step 2” of
Fig. 2 “TGT” and “AAC” coagulates due to the mutation of “T” to “G” between them. Yet another “null” effect is that the mutation turns an $x$-matching segment into another $x$-matching one, without changing the number/concentration of the $x$-matching segments.

In general the coagulation-fragmentation effects of mutations should depend on the concentration itself and even on the detail structure of the sequences, which can be formally described by the general (stochastic) nonlinear integro-differential equation with stochastic (both in space and time) mutation rates, the coagulation-fragmentation model (see, e.g., [15] for deterministic but kinetic descriptions analyzed rigorously by mathematician already, and references therein.) A systematic derivation of the (random) coefficients in the reaction rates [15] has not been available, but we are not completely clueless. The molecular biology considerations and the accumulated vast amount wisdom about coagulation-fragmentation processes provide useful information for us to proceed tentatively. For example, given the mutation rate, the fragmentation effect may well be modeled by the conventional fragmentation model; as for the coagulation, due to the fact that the number of alphabet, 4, in genome sequences is very small compared to the total length, its effect should mostly concentrate at small scales (actually the coagulation effect and the finite-alphabet effect are not completely separatable). Thus, as an application and development, we start closely with the very classical one [7, 12, 13].

Disorder in genome sequences and dynamical processes

One possible origin of the disorder is that there are many different segments of the same matching length; or, in other words, an $x$-matching set contain segments of different local (arrangement of) nucleotides pairs and/or different “ribbon” writhe, torsion and twist. These $x$-matching but different segments characterize mutations and/or duplications differently. The (random) environments also add to the disorder in the duplications and mutations. Since we can not or need not know exactly all the details, specific statistical distribution is applied. For this, as an idealization, one can think of a big ensemble containing a distribution of sub-ensembles, or, in other words, all the $x$-matching segments do not correspond to independent-identical-distribution (i.i.d.) variables but contain (infinitely) many subsets of elements corresponding to i.i.d. random variables.

We have to do appropriate specification and simplification with “effective” parameters, starting with the model where concentration $c(x, t)$ of segments of length $x$ at time $t$ evolves according to the following rate equation [7, 12, 13]

$$\frac{\partial c(x, t)}{\partial t} = -\mu xc(x, t) + 2\mu \int_x^\infty dy c(y, t) + f(x),$$

where, compared to previous studies, the new element in the model lies in the disorders of the initial condition $c(x, 0)$, of $\mu$ and of the input $f(x)$. The input is used to model the gross contributions from the duplications and the coagulation effects of mutations. Instead of dealing directly with the stochastic equation, we freeze the disorder in time (quenching) for a simplification.

SOLUTIONS

It is not completely clear whether we should best treat our present genome data as the “decaying” state, given the duplications happened a long time ago and the on-going mutations, or as the statistical steady state, balanced among the effects of duplications and mutations, or most generally a state corresponding to the time-dependent solution with initial data and time dependent inputs. The answer probably depends on the time scale we want to put our current observation in. We thus check all possibly relevant solutions.

“Decaying” solution

Let’s start with the “decaying” case with $f(x) = 0$ which may correspond to the case dominated by the fragmentation process and can be accurate for large scales. Given the initial condition $c(x, 0) = \delta(x - K)$, Ziff and McGrady [7, 12] found the solution

$$c(x, t) = e^{-K\mu t} \delta(x - K) + [2\mu t + (\mu t)^2(K - x)] e^{-x\mu t}$$

for $0 < x \leq K$, otherwise null; and, in general

$$c(x, t) = e^{-x\mu t} \{c(x, 0) + \int_x^\infty c(y, 0) [2\mu t + (\mu t)^2(y - x)] dy\}.$$
Note that, due to “mass” conservation $M = \int_0^\infty x c(x, t) dx$, the long-time solution $c(x, \infty) = M \delta(x)/x$, which simply means that as time goes the stick will just breaks down into “powders” of infinitesimal length.

Time-averaging, by integration from $t = 0$ to $\infty$, of the exponentially “decaying” Eq. (2) does lead to a power-law distribution with exponent $-3$, as noticed in Ref. [7]. Such a result from collecting the snapshots for both the history and future, however, to our point of view is not appropriately or clearly related to the outputs in Fig. 1 (actually time accumulations of “history” ending at infinite time $+\infty$ and reaching the current moment $t$ are given precisely by the steady-state and general time-dependent full solution, respectively, to be presented below.) Note also that there is an annoying pulse at $K$ in such a treatment.

We first check an example with a realization with $\tilde{c}(x, 0) = e^{-\tilde{s}x}$, as also in Ziff and McGrady [12], but $\tilde{s}$ has quenched disorder: We always use a tilde to denote a realization of the disorder. Our solution can be obtained in two steps. First, fix $\mu = \tilde{\mu}$ to get

$$\tilde{c}(x, t) = \left(\tilde{\mu} t + \tilde{s}\right)^2 \exp\left\{-\left(\tilde{\mu} t + \tilde{s}\right) x\right\}. \tag{4}$$

Then we integrate over the distributions $P_{\tilde{\mu}}$ of $\tilde{\mu}$ and $P_{\tilde{s}}$ of $\tilde{s}$ to get the final averaged solution. The more general result in some appropriate conditions [18] is possible to be evaluated with Laplace’s method. We illustrate them with definite examples as follows. For instance, assuming

$$P_{\tilde{\mu}}(\mu) \propto \mu^ne^{-\lambda \mu} \quad \text{and} \quad P_{\tilde{s}}(s) \propto s^m e^{-\Lambda s},$$

with $n \geq 0$ and $m > 1$ (for convergence of the continuous integrations), gives for $x \to \infty$ and $t \to \infty$ asymptotically

$$c(x, t) \propto x^{\alpha_d} t^\beta : \alpha_d = -(n + m + 2) \quad \text{and} \quad \beta = -(n + 1). \tag{5}$$

The restriction $m > 1$ here is due to the requirement of the convergence of the continuous integral, and, in practice, with discrete and/or finite scales, it may be relaxed by easily tuning the ansatz at small argument or controlling the integration range. Fig. 3 shows, with $\lambda_2 = 2, \Lambda_2 = 5$ and $t = 2$, a typical plot of solutions for $n = 0, m = 1$ and $2$, compared to lines with exact slopes $-3$ and $-4$: As said, for $m = 1$ the integration diverge at $s = 0$, so we obtain the result approaching $x^{-3}$ in the figure by setting $P_{\tilde{s}}(s) = 0$ at very small $s$. We just remark that the above result turns out to be quite robust for a large class of reasonable ansatzes: One may still get algebraic tails with other ansatzes, such as

$$c(x, 0) = x^p e^{-\tilde{s} x^2}, \quad P_{\tilde{\mu}}(\mu) \propto \mu^n e^{-\lambda \mu^2}, \quad P_{\tilde{s}}(s) \propto s^m e^{-\Lambda s^2};$$

$$c(x, 0) = \frac{x^p}{(x + \tilde{s})^q}, \quad P_{\tilde{\mu}}(\mu) \propto \frac{\mu^n}{(1 + \mu)^r}, \quad P_{\tilde{s}}(s) \propto \frac{s^m}{(1 + s)^z}.$$ 

So, the algebraic tails appear to be the generic output of the combination of such distributions. Mathematically these results remind us but are different to the “weighted mixtures” of Willinger et al. [19] who requires the distribution (corresponding to
our concentration) be a scaling function. But, indeed now the power-law tail is also what they stated to be “more normal than normal [Gaussian]”.

In genome sequence, the question is then what exactly are the initial distribution, the disorder in it and in the duplication and mutation rates? We tend to believe that the above ansatzes, with possible quantitative modifications, used for the explicit calculations should be qualitatively ‘reasonable’ in describing what has been happening in nature, since they just simply represent the obvious facts of peaks at some (moderately) small values and the convergence properties. But all these follow the assumption of quenched disorder. Quenched disorder should be considered to be a working hypothesis or an effective modeling strategy (as widely applied for complex systems). Intuitively, being frozen in time of the disorder in the initial concentrations may sound more natural, but that of the mutation rates is just a working simplification.

Steady state solution

Since duplications also happen without stopping as time goes and since mutations also lead to coagulation, it may also be illuminating to check the other ideal extreme case of statistical steady state as follows.

The final steady-state solution with a realization of input reads [13]

$$\mu_{\infty}(x) = x^{-1} \tilde{f}(x) + 2 x^{-3} \int_x^\infty dy \tilde{f}(y). \quad (6)$$

So, whatever the steady input is, a tail of $-3$ for $x \to \infty$ is not consistent. One can just check situations with $\tilde{f}(x)$ decaying as, faster than or slower than $x^{-2}$. Such a model can produce genuine slope steeper than $-3$ only with an input of power law steeper than $-2$: From Eq. (6), if and only if $\epsilon > 0$, the input of slope $-2 - \epsilon$ produces a distribution of slope $-3 - \epsilon$. For instance, an exponential input gives an exponential tail with algebraic prefactor. Note in particular that $\epsilon$ cannot be 0. [Such observations were already partly made by Koroteev and Miller [16] semi-empirically. For the input $\tilde{f}(x) = \delta(x - K)$, a solution $\propto x^{-3}$ extends to all scales below $K$ \[^2\,\text{[13]}\], but we are afraid that such “monodisperse” be not the case of evolutionary genome sequences and that the result be not directly applicable (not to mention other issues such as a pulse at $x = K$).]

The restriction of power-law input for a genuine power-law tail is removed by disorder: For example, an input $\tilde{f}(x) = \tilde{\lambda} e^{-\tilde{\lambda} x}$ gives

$$\mu_{\infty}(x) = \frac{2 \tilde{\lambda} x + \tilde{\lambda}^2 x^2 + 2}{\tilde{\lambda} e^{\tilde{\lambda} x} x^3} \tilde{\lambda} e^{-\tilde{\lambda} x}$$

which, when $\tilde{\lambda}$ has disorder (quenched) of a distribution ansatz

$$P_\chi(\lambda) \propto \lambda^n e^{-\Lambda_\lambda} \text{ with } \Lambda > 0 \text{ and } n > 0,$$

produces an asymptotic tail ($x \to \infty$)

$$c_{\infty}(x) \propto x^{\alpha_s} : \quad \alpha_s = -(n + 3). \quad (7)$$

The exponent $\alpha_s$ is independent of $\mu$ disorder. Just as the decaying case, the condition $n > 0$ is for convergence of the continuous integration, and in practice with discrete and/or finite scales, this condition may be relaxed by tuning the ansatz at small arguments; and, also, some other ‘reasonable’ (again, in the sense of consistent with our understanding of the duplication and coagulation effects from mutations) ansatzes for the input also produce the power-law tails.

Full time-dependent general solution

Strong enough of the theoretical supports from the decaying and ultimate steady state solutions, it is still natural to consider the more general time-dependent solution with the memory of the initial supply (the first duplications) and the other inputs along the history, which is the \textit{full time-dependent general solution}, composed of two parts, the linear superposition of the solution “decaying” from some initial condition and that driven by some input. As already given with the Mellin transform and Charlesby method by Ben-Naim and Krapivsky [13] whose details we resist to reproduce here, the solution is simply the linear superposition of the previous decaying solution from some given initial condition without input and the final forced steady state solution forgetting the initial condition, a consequence of the linearity of the dynamics (as long as the forcing does not depend
on the solution itself). Thus, we can immediately conclude from Eq. (5,7) that the statistical average over the quenched disorder also presents asymptotic power-law tail:

\[ c(x, t) \propto C_1 t^{\beta x^{\alpha_d}} + C_2 x^{\alpha_s}. \]

(8)

We remark that at different scales separated far apart, the two components may dominate respectively. Actually we have already neglected subdominant algebraic terms for \( x \to \infty \) and/or \( t \to \infty \) while giving the final decaying or steady-state solution, but the subdominant algebraic term(s) could dominate at some intermediate regime(s), showing different power law(s): Whether or not this is happening in the data depends on the coefficient(s), \( C_i \), determined by the dynamics and on whether there are other contaminations. For example, the alignment concentrations of rice, as given by Fig. 4, appear to support two power-law regimes, especially for the unmasked data. Some other data also present such a similar feature [20], though do not always have as sharp results. On the other hand, as a ‘negative’ effect, such mixed algebraic components may also add to the ambiguity in detecting the power law where they don’t separate well or the data don’t have enough range of scale to separate them [8].

\[ \text{FIG. 4: The concentration of unmasked rice chromosome 1 shows an extra power-law regime, compared to the masked one. Unlike the human chromosome 1 (cf., inset of Fig. 1), in this case masking repeats does not seem to affect the scaling law, with exponent being roughly } -3.4 \text{ as the reference dashed lines shows, at the largest scales.} \]

**DISCUSSION**

We iterate that the classical result of the pure fragmentation model with constant input predicts a \( x^{-3} \) ‘head’ for \( x \to 0 \), though the ‘monodispersion’ [13] case with the input being a pulse/delta at a single largest scale extends the scaling to just below the forced scale; ‘monodispersion’ of duplicated segments of evolutionary genomes [7] however appears to be unrealistic for us. Coming back to nature’s data, we summarize that there are a variety of power laws which may be abstracted to be a problem of tails extending to the largest scales, ideally \( x \to \infty \). The physical intuition, the computed various scaling laws among different species or even the different chromosomes of the same species, and some of the poor convergence of power-law data (c.f., Fig. 1 and Refs. [8,11]), all call for the consideration of disorder which, as we have shown, turns out to help solve the problem. The exponential “head” of the concentration of course comes partly from the random matching due to the finite number, 4, of the alphabets as well as from the coagulation effect which has been taken into account in our input.

Introducing disorder into the pure fragmentation model for DNA evolution [7], even though a simplification of the supposed-to-be stochastic fragmentation-coagulation differential-integral equation, somehow opens a Pandora’s box. Various ansatzes of disorder were found to produce power-law tails in quite a generical way, which demonstrates some university in nature, but they have interesting differences in biophysics demonstrated through the ‘complexity in genomes’ (21 and references therein). So, a meaningful direction of study is to identify the disorders in the data. Further measurements/computations from the genomes to offer more information or to quantify the various ingredients such as the distributions of the duplication, the mutation rates, the coagulation effects will help to narrow down our somewhat general results to even more specific biophysics.

It is interesting to identify what is specific (especially concerning disorder) to the masked repeats and the possible relevance with selection: To our best knowledge, there is no established selective mechanisms that would be in favor of or against a power-law concentration. There had been some other models of molecular evolution emphasizing on different aspects of observations
(see, e.g., Refs. [21–25] for a partial list). And it is hoped that some of the considerations here may also be applicable to some of them.

This work benefited from the helpful discussions with J. Miller’s group, and we thank W. Li for kindly explaining his work to us.

[1] Evolution after Gene Duplication, Editor(s): Katharina Dittmar, David Liberles, 2010 Wiley-Blackwell.
[2] www.repeatmasker.org
[3] K. Laland and colleagues v.s. G. A. Wray, H. E. Hoekstra and colleagues, “Does evolutionary theory need a rethink?”, Nature 514, 161 (2014).
[4] J. A. Bailey and E. E. Eichler, Nat.Rev. Genet. 7, 552 (2006).
[5] K. Gao and J. Miller, Computational Biology and Chemistry 53(A), 59 (2014).
[6] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg, Genome Biol. 5, R12 (2004).
[7] F. Massip and P. Arndt, Phys. Rev. Lett. 110, 148101 (2013).
[8] K. Gao and J. Miller, PLoS ONE 6, e18464 (2011).
[9] E. Taillefer and J. Miller, Int. Conf. Natural Computation, Shanghai, China, pp. 1454 - 1460, 2011.
[10] W. Li, J. Freudenberg and P. Miramontes, BMC Bioinformatics 15, 2 (2014).
[11] W. Li and J. Freudenberg, Frontiers in Genetics 5, 1 (2014).
[12] R. M. Ziff and E. D. McGrady, J. Phys. A 18, 3027 (1985).
[13] E. Ben-Naim, P.L. Krapivsky, Phys. Lett. A 275, 48 (2000).
[14] M. Koroteev and J. Miller, “Fragmentation dynamics of DNA sequence duplications”, Preprint: http://arxiv.org/abs/1304.1409
[15] D. Broizat, Annales de l’Institut Henri Poincare (C) Non Linear Analysis 27, 809 (2010).
[16] M.V. Koroteev, J. Miller, Phys. Rev. E 84, 061919 (2011).
[17] R. D. Vigil and R. M. Ziff, Journal of Colloid and Interface Science 133, 257 (1989).
[18] C. Bender and S. Orszag, Advanced Mathematical Methods for Scientists and Engineers: Asymptotic Methods and Perturbation Theory (v. 1), Springer (1999).
[19] W. Willinger, D. Alderson, J. C. Doyle, L. Li, in Proceedings of the 2004 Winter Simulation Conference, R. G. Ingalls, M. D. Rossetti, J. S. Smith, B. A. Peters, Eds. (Institute for Operations Research and the Management Sciences, Hanover, MD, 2004), pp. 130 - 141.
[20] K. Gao, Private communication; see also Ref. 9.
[21] Y. Almirantis, P. Arndt, W. Li and A. Provata, Computational Biology and Chemistry 53, 1 (2014).
[22] W. Li and K. Kaneko, Nature 360, 635 (1992).
[23] P.W. Messer, P.F. Arndt, M. Lassig, Phys. Rev. Lett. 94, 138103 (2005).
[24] S. V. Buldyrev, A.L. Goldbergerin, S. Havlin, C-K. Peng, H.E. Stanley, M. Simons, Biophys. J. 65, 2673 (1993).
[25] A. Grosberg, Y. Rabin, S. Havlin, A. Neer, Europhys. Lett. 23, 373 (1993).