Sequence Evolution and Long-range Correlations in Genomic Base Composition

Philipp W. Messer[1,2], Peter F. Arndt[2], and Michael Lässig[1]
[1] *Institute for Theoretical Physics, University of Cologne, Zülpicher Str. 77, 50937 Köln, Germany and*
[2] *Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany*
(Dated: April 14, 2005)

We study a minimal model for genome evolution whose elementary processes are single site mutation, duplication and deletion of sequence regions and insertion of random segments. These processes are found to generate long-range correlations in the composition of letters as long as the sequence length is growing, i.e., the combined rates of duplications and insertions are higher than the deletion rate. For constant sequence length, on the other hand, all initial correlations decay exponentially. These results are obtained analytically and by simulations. They are compared with the long-range correlations observed in genomic DNA, and the implications for genome evolution are discussed.

PACS numbers: 87.23.Kg, 87.15.Cc, 05.40.-a

I. INTRODUCTION

Over a decade ago, long-range correlations in the sequence composition of DNA have been discovered [1–3]. With the rapidly growing availability of whole-genome sequence data, the composition of genomic DNA can now be studied systematically over a wide range of scales and organisms. The statistical analysis is quite intricate since genomic DNA is a rather "patchy" statistical environment [4]: it consists of genes, noncoding regions, repetitive elements etc., and all of these substructures have a systematic influence on the local sequence composition. Variations in composition along the genome have been studied extensively by a number of different methods [5–12], and it is now well established that long-range correlations in base composition appear in the genomes of many species. These can be measured, for example, by the autocorrelation function $C(r)$ of the GC-content, which measures the likelihood of finding G-C Watson-Crick pairs at a distance of r bases along the backbone of the DNA molecule. However, the form of these correlations is much more complex than simple power laws. Within one chromosome, there is often a variety of different scaling regimes and effective exponents, and sometimes no clear scaling at all. Moreover, the effective exponents of comparable scaling regimes vary considerably between different species, and even between different chromosomes of the same species [10, 11, 13].

Despite the ubiquity of genomic correlations, little is known about their evolutionary origin. In this article, we address the question whether the observed correlations can be explained quantitatively by a biologically realistic "minimal" model of sequence evolution. We take into account four well known elementary evolutionary modes: single site mutations, duplications and deletions of existing segments of the sequence, and insertions of random segments. The duplication processes are believed to be a crucial mechanism of genome growth [14–16]; the length of the duplicated segments ranges from single letters to thousands of letters as in the case of gene duplications. The model is minimal in the sense that all four elementary modes are *local* stochastic processes compatible with *neutral evolution*, i.e., they do not require any assumption of natural selection. An alternative possible reason for the observed correlations may be *long-range interactions* likely to be caused by natural selection for a specific local GC-content. An example of such a selective process is the clustering of genes in some regions of a chromosome [17], but no plausible mechanism producing long-range interactions has been proposed so far.

Li's original work has shown that already a simple stochastic process consisting of duplications and mutations of single letters leads to generic power law correlations in the sequence composition [18]. Here we analyze in detail the generalized sequence evolution model introduced above. In particular, we calculate the stationary two-point correlation function $C(r)$. It is of power law form, $C(r) \propto r^{-\alpha}$, with a decay exponent α depending on only two effective parameters, which are simple functions of the rates of the elementary processes. These long-range correlations are generic as long as the rates of the processes result in a growing sequence. At constant sequence length, however, the stationary correlations in sequence composition vanish, and initial correlations from a previous growth phase decay. Our analytic results (which differ from Li's approximate expressions [18] and the results of [19]) are in excellent agreement with our numerical simulations. We use these results to infer from measured values of α a lower bound on the growth rate of the genome, which can be compared with independent estimates. The implications of our findings on the evolution of mammalian genomes are discussed at the end of this article.

II. SEQUENCE EVOLUTION MODEL

The stochastic evolution model generates sequences $(s_1, \dots, s_N)$ of variable length N. For simplicity, their letters are taken from a binary alphabet; $s_k = \pm 1$. (In the application to genomic systems, $s_k = +1$ denotes a GC-pair and $s_k = -1$ an AT-pair at backbone position k.) The elementary evolutionary steps are mutations, duplications, insertions, and deletions of single letters (the generalization to segments will be discussed below). They are Markov processes with rates μ, δ, γ^+, and γ^- acting on the sequences as

$$(\cdots, s, s', \cdots) \to \begin{cases} (\cdots, -s, s', \cdots) & : \text{ rate } \mu \\ (\cdots, s, s, s', \cdots) & : \text{ rate } \delta \\ (\cdots, s, x, s', \cdots) & : \text{ rate } \gamma^+ \\ (\cdots, s', \cdots) & : \text{ rate } \gamma^-, \end{cases} \tag{1}$$

where $x = \pm 1$ denotes an uniformly distributed random letter. Duplication and insertion events introduce a new letter next to an exiting one and shift all subsequent letters one position to the right, thereby increasing the sequence length by 1. Conversely, deletions shorten the length by 1. This type of Markov evolution model is widely used in computational biology, forming the statistical basis of sequence alignment algorithms [20]. Running all four processes over a time t produces a statistical ensemble of sequences; the corresponding averages are denoted by $\langle \dots \rangle(t)$. This ensemble is characterized by the rates δ, μ, γ^+, γ^-, and by the initial sequence. Here we use sequences of length 1 with a fixed letter, $(s_1) = 1$, or a random letter, $(s_1) = x$.

After a time t, the sequences have an average length $\langle N \rangle(t) = \exp(\lambda t)$ with the effective growth rate

$$\lambda = \delta + \gamma^+ - \gamma^-. \tag{2}$$

We are interested in two dynamical regimes, sequence growth from a single-letter initial state (i.e., $\lambda > 0$) and the evolution of sequences at stationary length $\langle N \rangle \gg 1$ (i.e., $\lambda = 0$), to which we now turn in order.

III. GROWTH DYNAMICS AND STATIONARY CORRELATIONS

The composition bias of the sequences at position k is measured by the expectation value $\langle s_k \rangle(t)$. It is easy to show that any initial composition bias decays due to mutations and random insertions. We note that each insertion can be regarded as a duplication with a subsequent mutation in half of the cases, resulting in an effective mutation rate

$$\mu_{\text{eff}} = \mu + \gamma^+/2. \tag{3}$$

We obtain $\langle s_k \rangle(t) \propto \exp(-2\mu_{\text{eff}} t)$ for fixed initial condition, while $\langle s_k \rangle(t) = 0$ for random initial conditions. The composition correlation $C(r) \equiv \langle s_k s_{k+r} \rangle(t)$ between two sequence positions at distance r is affected by all four processes and is independent of the initial condition. Its evolution equation can be derived by writing it as $C(r,t) = P_{\text{eq}}(r,t) - P_{\text{op}}(r,t)$, where $P_{\text{eq}}(r,t)$ and $P_{\text{op}}(r,t)$ denote the joint probabilities of finding two symbols of equal and opposite signs, respectively, at a distance r. The Master equation for $P_{\text{eq}}(r,t)$ takes the form

$$\begin{aligned} \frac{\partial}{\partial t} P_{\text{eq}}(r,t) &= 2\mu_{\text{eff}} \left[-P_{\text{eq}}(r,t) + P_{\text{op}}(r,t) \right] \\ &\quad + [r\delta + (r-1)\gamma^+] \left[P_{\text{eq}}(r-1,t) - P_{\text{eq}}(r,t) \right] \\ &\quad + r\gamma^- \left[P_{\text{eq}}(r+1,t) - P_{\text{eq}}(r,t) \right]. \end{aligned} \tag{4}$$

The first term on the r.h.s. describes the change in $P_{\text{eq}}(r,t)$ due to mutations and random insertions, while the second term specifies the probability current due to duplication of a site in the interval $(k, k+r-1)$ or insertion of a new site in the interval $(k, k+r-2)$. The third term gives the corresponding current due to deletions. By exchanging P_{eq} and P_{op}, we obtain a similar equation for $P_{\text{op}}(r,t)$. Hence we have

$$\begin{aligned} \frac{\partial}{\partial t} C(r,t) &= -4\mu_{\text{eff}}\, C(r) \\ &\quad + [r\delta + (r-1)\gamma^+] \left[C(r-1) - C(r) \right] \\ &\quad + r\gamma^- \left[C(r+1) - C(r) \right]. \end{aligned} \tag{5}$$

For the special case with only single-letter duplications and mutations ($\delta, \mu > 0$, $\gamma^+ = \gamma^- = 0$), which is equivalent to Li's original model [18], we find a simple analytical form for the stationary $C(r)$ by solving the recursion

$$C(r) = \frac{r}{\alpha + r}\, C(r-1) \quad \text{with} \quad \alpha = \frac{4\mu}{\delta} \tag{6}$$

and the initial value $C(0) = 1$. This gives

$$C(r) = \frac{\Gamma(r+1)\Gamma(1+\alpha)}{\Gamma(r+1+\alpha)} = \frac{\alpha}{1+\alpha}\, B(r,\alpha), \tag{7}$$

where $\Gamma(x)$ is the gamma function and $B(x,y)$ the beta function. Evaluating its asymptotic behavior for $x \gg 1$,

$$B(x,y) \propto \Gamma(y)\, x^{-y} \left[1 - \frac{y(y-1)}{2x} \left(1 + \mathrm{O}\left(\frac{1}{x} \right) \right) \right],$$

then produces the algebraic decay $C(r) \propto r^{-\alpha}$. For the general case including insertions and deletions, the asymptotic decay can still be obtained exactly in the continuum limit. For $r \gg 1$ and $\delta > 0$, the difference equation (5) becomes the differential equation

$$\frac{\partial}{\partial t} C(r,t) = -4\mu_{\text{eff}} C(r,t) - r\lambda \frac{\partial}{\partial r} C(r,t) \tag{8}$$

with the effective rates μ_{eff} and λ defined by (2) and (3). This has the stationary solution

$$C(r) \propto r^{-\alpha} \quad \text{with} \quad \alpha = \frac{4\mu_{\text{eff}}}{\lambda}. \tag{9}$$

Eq. (8) clearly shows the mechanism generating long-range correlations in this type of sequence evolution model. Correlations are continuously produced at small scales by duplications and transported to larger distances by the net exponential expansion of the sequence (resulting from duplications and insertions/deletions). On the other hand, correlations decay exponentially due to processes randomizing the sequence (i.e., mutations and random insertions). The competition between expansion and randomization produces the algebraic decay $C(r) \propto r^{-\alpha}$, which is highly universal. Microscopic details of the evolution processes are irrelevant, the exponent α is determined by a simple balance between the growth rate λ and the effective mutation rate μ_{eff}. Hence, an extended model containing duplications, deletions and random insertions of sequence *segments* of finite length $\ell = 1, 2, ..., \ell_{\max}$ with respective rates δ_l, γ_ℓ^-, and γ_ℓ^+ still has the same asymptotics (9) for $N(t) \gg \ell_{\max}$ and $r \gg \ell_{\max}$. The effective rates (2), (3) are now given by

$$\lambda = \sum_\ell \ell \left[\delta_\ell + \gamma_\ell^+ - \gamma_\ell^-\right], \quad \mu_{\text{eff}} = \mu + \frac{1}{2}\sum_\ell \ell\gamma_\ell^+. \quad (10)$$

This asymptotics can again be proved from an exact Master equation similar to (5) [13]. The extended model is important for genomic evolution since strong long-range correlations (i.e., small values of α) can be the combined result of segment duplications with different values of ℓ. Their individual rates δ_ℓ might be small and difficult to assess but the cumulative rate λ can still be estimated.

IV. STATIONARY-LENGTH DYNAMICS

It is obvious from Eq. (8) that stationary long-range correlations only exist as long as the sequence grows, i.e. for $\lambda > 0$. Consider now the following evolutionary scenario: sequence growth with rate $\lambda_1 > 0$ up to a length $N_0 = N(t_0)$, followed by a second phase with $\lambda_2 = 0$ and $\langle N\rangle(t) = N_0$ for $t > t_0$. The time-dependent solution of Eq. (8) for the asymptotics of $C(r,t)$ is then

$$C(r,t) = C(r,t_0)\, e^{-4\mu_{\text{eff}}\Delta t} \propto r^{-4\mu_{\text{eff}}/\lambda_1} e^{-4\mu_{\text{eff}}\Delta t} \quad (11)$$

with $\Delta t = t - t_0 > 0$. In the second phase, the long-range tails of $C(r,t)$ are preserved but their amplitude decays with a characteristic time scale $\tau = (4\mu_{\text{eff}})^{-1}$.

Numerical results.—We have performed extensive Monte Carlo simulations of our model. During each time step $\Delta t = [(\mu + \sum_\ell[\delta_\ell + \gamma_\ell^+ + \gamma_\ell^-])N(t)]^{-1}$ we choose a random site and apply one of the elementary processes with its relative weight. For a single realization of this dynamics, the correlation function $C(r)$ is well approximated by the sequence average $(N-r)^{-1}\sum_{k=1}^{N-r} s_k s_{k+r}$. Further averaging over 100 realizations produces very accurate measurements of $C(r)$.

Fig. 1(a) shows the numerical $C(r)$ for the single-letter duplication-mutation dynamics with various rates, which is in excellent agreement with the analytic expression (7). The same is shown in Fig. 1(b) for the general case with all types of processes present, verifying the asymptotic behavior (9) and (10). For completeness, we have also obtained power spectra and the mutual information function, as defined in [11], which have the expected decay exponents $1-\alpha$ and 2α, respectively.

The dynamical build-up of these correlations for growing sequences is seen in Fig. 2(a), which shows $C(r,t)$ at various intermediate times of the growth process. The correlation rapidly converges to the stationary form for all distances $r \lesssim N(t)$. This should be compared with the time-dependence of $C(r,t)$ at constant length in Fig. 2(b), which shows an algebraic tail with an exponentially decreasing amplitude as predicted by Eq. (11).

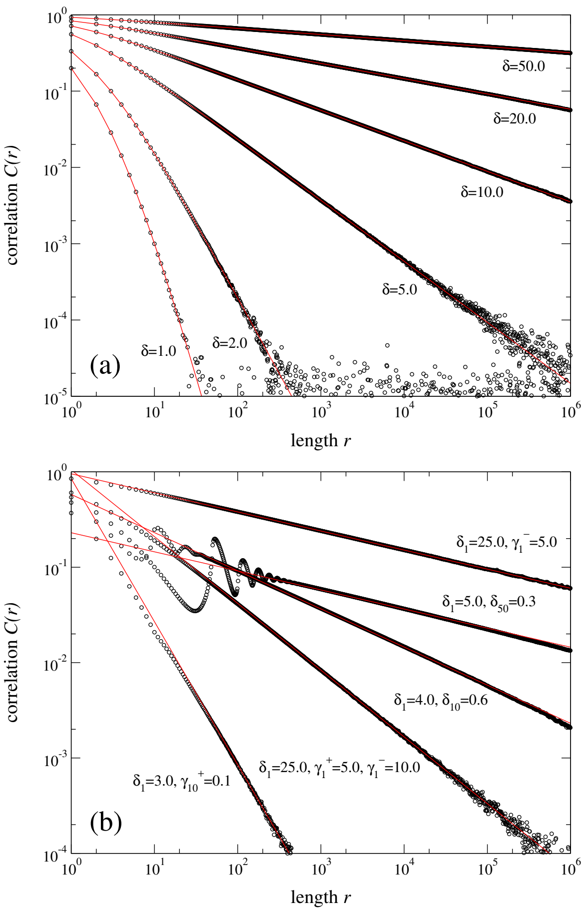

FIG. 1: Stationary $C(r)$ at different rates of the elementary processes. (a) Single-letter duplication-mutation model: Numerical results (circles) and the analytical form (7) (lines) for $\mu = 1$, δ varying. (b) Full model: Numerical results (circles) with the analytic asymptotics (9) and (10) (lines) for $\mu = 1$ and varying rates of the other processes (rates not specified in the plot are zero). The dynamics of the sequences was simulated until they reached a length of $N = 2^{27} \approx 10^8$; $C(r)$ was averaged over the sequence and over 100 runs.

V. GENOMIC EVOLUTION

As pointed out above, the processes discussed here build a minimal model for dynamically generated long-range correlations along a sequence. But can this model explain the observed correlations in genomic DNA? The correlation function $C(r)$ along human chromosomes shows a rather slow algebraic decay on distance scales $10^3 < r < 10^6$ with typical effective exponents $\alpha \approx 0.1$ [10, 11]. We have confirmed these measurements and found them to be consistent with sequence data from other mammals [13]. A lower bound of the effective mutation rate in mammals is $\mu_{\text{eff}} \approx 2\cdot 10^{-9}\text{a}^{-1}$ per site [21]. Assuming stationary growth, we can use these values of α and μ_{eff} to derive a lower bound on the genomic growth rate λ, resulting in a minimum value $\lambda \approx 10^{-7}\text{a}^{-1}$ per site according to Eq. (9). However, this rate is much

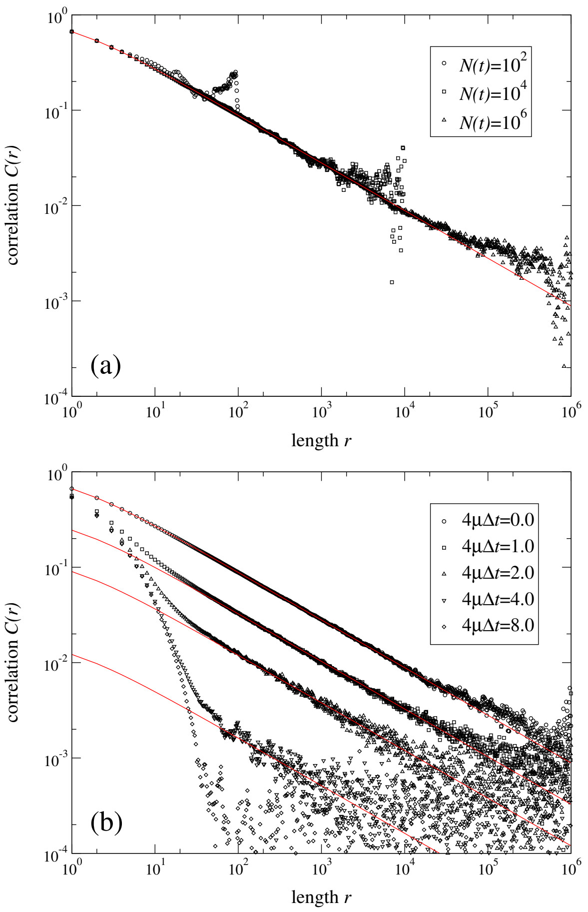

FIG. 2: Time-dependent correlations $C(r,t)$. (a) Build-up of long-range correlations by stationary growth. Measured $C(r,t)$ at various intermediate lengths $N(t) = 10^2, 10^4, 10^6$ (symbols) together with the stationary form (7) (line) for $\mu = 1$, $\delta_1 = \delta = 8$, all other parameters are zero. (b) Decay of correlations during sequence evolution at stationary length $N_0 = 10^6$. Measured $C(r,t)$ at various times Δt (symbols) together with the analytic decay of the long-range tail given by Eq. (11) (lines).

too high. Our genome would have expanded much faster than it is observed since the current human genome contains $N \approx 3 \cdot 10^9$ base pairs and, assuming the above rate of genome expansion, would have contained only about $4 \cdot 10^5$ base pairs at the time of mammalian radiation about 90 million years ago. This can clearly be rejected since approximately 40% of the human genome can be aligned to the mouse genome, representing most of the orthologous sequences that remain in both lineages from the common ancestor [22].

Over longer evolutionary periods, genomic expansion phases with rates $\lambda \approx 10^{-7}\text{a}^{-1}$ cannot be ruled out if we assume the history of the genome has been a *punctuated* process, with such expansion phases followed by periods of approximately constant length. In the human genome, there is by now ample evidence for growth by segmental duplications with various segment lengths [23, 24]. In a punctuated growth process, correlations are produced and transported during the expansion phases. During the stationary phases, the previously established correlations decay as given by Eq. (11). In mammals, the last likely period of rapid expansion has been the mammalian radiation, and the characteristic time scale of the decay is $\tau \approx 100$ Myr. Correlations present or generated at the time of the mammalian radiation would hence still persist. The succession of several distinct growth phases with different values of λ and μ_{eff} could even explain correlations $C(r)$ with several scaling regimes as found in human chromosomes [10]. Thus, a punctuated expansion-randomization process may be compatible with the correlations observed in mammals. Clearly, this scenario is a hypothesis at present, and other causes for the correlations are not ruled out. Indeed, the rather diverse functional forms found in different species may point towards more than one generating mechanism. If genomic expansion proves to be a significant contribution, composition correlations could be the "background radiation" of genomics, allowing us to trace the history of genomes far back in evolutionary time.

[1] W. Li and K. Kaneko, *Europhys. Lett.* **17**, 655 (1992).
[2] C.-K. Peng *et al.*, *Nature* (London) **356**, 168 (1992).
[3] R. F. Voss, *Phys. Rev. Lett.* **68**, 3805 (1992).
[4] S. Karlin and V. Brendel, *Science* **259**, 677 (1993).
[5] C.-K Peng *et al.*, *Phys. Rev. E* **49**, 1685 (1994).
[6] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, *Phys. Rev. Lett.* **74**, 3293 (1995).
[7] W. Li, *Comput. Chem.* **21**, 257 (1997).
[8] M. de Sousa Vieira, *Phys. Rev. E* **60**, 5932 (1999).
[9] H. E. Stanley *et al.*, *Physica A* **273**, 1 (1999).
[10] P. Bernaola-Galvan, P. Carpena, R. Roman-Roldan, and J. L. Oliver, *Gene* **300**, 105 (2002).
[11] D. Holste *et al.*, *Phys. Rev. E* **67**, 061913 (2003).
[12] Z. Ouyang, C. Wang, and Z. S. She, *Phys. Rev. Lett.* **93**, 078103 (2004).
[13] P. Messer, M. Lässig, and P. Arndt, to be published.
[14] R. V. Samonte and E. E. Eichler, *Nat. Rev. Genet.* **3**, 65 (2002).
[15] L.-C. Hsieh, L. Luo, F. Ji, and H. C. Lee, *Phys. Rev. Lett.* **90**, 018101 (2003).
[16] A. Goffeau, *Nature* (London) **430**, 25 (2004).
[17] M. J. Lercher, A. O. Urrutia, A. Pavlicek, and L. D. Hurst, *Human Mol. Genetics* **12**, 2411 (2003).
[18] W. Li, *Phys. Rev. A* **43**, 5240 (1991).
[19] R. Mansilla and G. Cocho, *Complex Syst.* **12**, 207 (2000).
[20] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis* (Cambridge University Press, Cambridge, England, 1998).
[21] P. F. Arndt and T. Hwa, *Bioinformatics* **20**, 1482 (2004).
[22] R. H. Waterston *et al.*, *Nature* (London) **420**, 520 (2002)
[23] J. A. Bailey *et al.*, *Science* **297**, 1003 (2002)
[24] E. E. Thomas *et al.*, *PNAS* **101**, 10349 (2004)