Chapter 3

Tandem duplications and genomic correlations

In this chapter, we study the statistical features of sequences evolving by mutational processes. Models that include mechanisms of local sequence randomization and segmental tandem duplication are found to constitute a universality class of non-equilibrium 1D expansion-randomization systems with generic stationary long-range correlations in a regime of growing sequence length. We analytically calculate the two-point correlation function of the sequence composition and the distribution function of the composition bias in sequences of finite length. The characteristic exponent of these quantities is determined by the ratio of two effective rates. It is calculated explicitly for several specific sequence evolution dynamics of the universality class. We also discuss the non-stationary build-up and decay of correlations, as well as more complex evolutionary scenarios with varying rates of the processes in time and space. At the end of this chapter, we address the question whether the observed correlations in eukaryotic genomes can indeed result from the mutational processes of our evolutionary model.

Our comparative genomics analysis of DNA insertions in the human lineage revealed that tandem duplication is the predominant mechanisms to generate insertions of short DNA segments in the genome during recent evolution. More than 90% of all identified insertions of single nucleotides, for example, were found to be consistent with tandem duplication events. This would only have been expected in approximately half of the cases under the assumption that inserted nucleotides were drawn randomly from the four different bases. Moreover, a prevalence of tandem duplications among insertions of DNA segments was observed on all length scales investigated in our analysis (1-100 bp), and the odds of observing a tandem duplication by chance vanish rapidly with increasing segment length. Although the duplication process does not seem to be perfect in a sense that newly inserted segments resemble exact copies of juxtapositional sequence (especially for longer segments), almost all insertions still show high sequence similarity to their direct vicinity.

The observed predominance of tandem duplication insertions substantially disagrees with the characteristics of indels in conventional models of sequence evolution, where DNA insertions are typically modeled as segments of independently drawn random

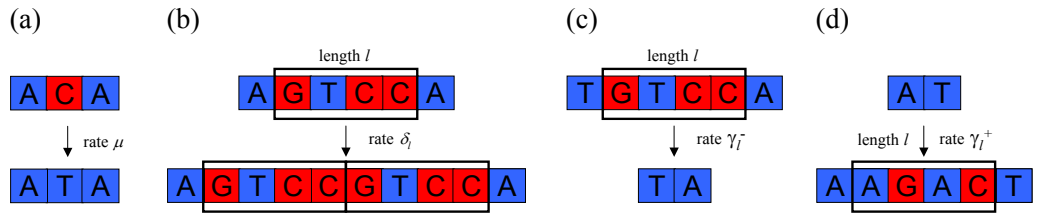

Figure 3.1: The four elementary processes of local sequence evolution incorporated in our dynamical model: (a) Single nucleotide mutations lead to an exchange between a GC and an AT base pair and occur at rate μ. Notice that in our binary model we do not need to consider A↔T and G↔C mutations. (b) Segmental duplications insert a new copy of a randomly chosen existing sequence segment of length ℓ next to it at rate δ_ℓ. (c) Deletions of existing sequence segments occur at rate γ_ℓ^-, respectively. (d) New segments composed of randomly chosen letters are inserted at rate γ_ℓ^+.

nucleotides. In the light of our findings, it remains questionable whether this simplification is still applicable when aiming for a realistic description of genome evolution and the resulting statistical properties of genomic sequences. In this chapter, we want to address this question by an in-depth analysis of the effects of tandem duplication processes on elementary statistical features of genome sequences. Our analysis will thereby shed light on the contributions of tandem duplications to a particularly interesting class of statistical sequence characteristics in genomes related to spatial long-range correlations and large fluctuations in genomic base composition. From a more conceptual perspective, our findings will also serve us to establish a powerful concept of statistical physics – universality classes – for the first time in the context of evolutionary biology.

3.1 Dynamical model of sequence evolution

We want to study a "minimal" stochastic model of genome evolution that incorporates the major local stochastic processes assumed to be acting on genomic DNA sequences during evolution [58], including tandem duplications of sequence segments. The focus of our analysis lies in the analytic calculation of general statistical properties of the sequences evolving under the evolutionary model. In particular, we are interested in magnitude and spatial structure of fluctuations in genomic base composition. Aiming primarily at large-scale characteristics, we can effectively simplify our analysis by making use of complementary strand symmetry, which has been shown to hold in eukaryotic genomes for first and also higher-order symmetries if investigated length scales are large enough [16]. Genome sequences $\vec{s} = (s_1, \ldots, s_N)$ of variable length $N(t)$ can therefore be modeled as binary sequences with letters $s_k = \pm 1$, where $s_k = +1$ denotes a GC Watson-Crick pair and $s_k = -1$ represents an AT pair.

Model definition The elementary evolutionary steps of our dynamical model are single site mutations, deletions and tandem duplications of existing sequence segments

of arbitrary lengths, and insertion of random segments (see Fig. 3.1). Formally, the dynamics of the processes can be defined by

$$\begin{array}{lcll} (\cdots, s, \cdots) & \to & (\cdots, -s, \cdots) & \text{mutation rate } \mu \\ (\cdots, (s)_\ell, \cdots) & \to & (\cdots, (s)_\ell, (s)_\ell, \cdots) & \text{duplication rates } \delta_\ell \\ (\cdots, s, \cdots) & \to & (\cdots, s, (x)_\ell, \cdots) & \text{insertion rates } \gamma_\ell^+ \\ (\cdots, (s)_\ell, \cdots) & \to & (\cdots, \cdots) & \text{deletion rates } \gamma_\ell^-, \end{array} \tag{3.1}$$

where $(s)_\ell$ denotes an existing sequence segment of length $\ell \geq 1$, and $(x)_\ell$ is a segment of length ℓ with uniformly distributed random letters $x_i = \pm 1$. Note that by convention we do not allow insertion of random segments prior to the first sequence element. Duplication and insertion events introduce a new sequence segment next to an existing one and shift all subsequent letters ℓ positions to the right, thereby increasing the sequence length by ℓ. Conversely, deletions shorten the length by ℓ. We will restrict all processes to a maximum range $\ell_{\max}$, i.e., all rates δ_ℓ, γ_ℓ^+, and γ_ℓ^- are zero for $\ell > \ell_{\max}$. Repeatedly running the processes over a time t produces a statistical ensemble of sequences; the corresponding averages are denoted by $\langle\cdot\rangle(t)$. This ensemble is characterized by the rates of the processes and by the initial sequence.

The statistical properties of sequences generated by our dynamical model have already been investigated for the special case of $\ell_{\max} = 1$, i. e. for a restricted set of evolutionary processes only incorporating single site duplications, single site duplications and deletions, and insertions of single random nucleotides [111]. In this analysis, it has analytically been shown that the restricted model generates long-range correlations in the composition of letters along the sequences. We will recall some of the major results of [111] later in this chapter, as they form an obvious special case of our more general dynamics defined in (3.1).

At first glance, the vast number of parameters in our general segmental dynamics might seem daunting in comparison to the simple single letter model of [111]. However, besides an obvious biological motivation arising from the fact that in molecular evolution insertions and deletions are not restricted to single base pairs, a careful treatment of the general model will also allow us to elucidate the emergence of universality in a broad class of one-dimensional dynamical systems, so-called expansion-randomization systems [114]. The concept of universality generally refers to the observation that "macroscopic" properties of a large class of systems are essentially independent of the "microscopic" dynamical details [150]. In our case, individual rates and characteristics of the local evolutionary processes constitute the microscopic details of the system. Macroscopic properties correspond to the statistical features of the generated sequences on length scales much larger than $\ell_{\max}$. It will turn out that these macroscopic properties are determined by just two effective parameters, the asymptotic growth rate λ and the effective mutation rate μ_{eff}, defined by

$$\lambda = \delta_{\text{eff}} + \gamma_{\text{eff}}^+ - \gamma_{\text{eff}}^- \tag{3.2}$$

$$\mu_{\text{eff}} = \mu + \frac{1}{2}\gamma_{\text{eff}}^+. \tag{3.3}$$

Both are simple functions of the cumulative rates of the microscopic processes,

$$\delta_{\text{eff}} = \sum_{\ell=1}^{\ell_{\max}} \ell\delta_\ell, \quad \gamma_{\text{eff}}^{+} = \sum_{\ell=1}^{\ell_{\max}} \ell\gamma_\ell^{+}, \quad \text{and} \quad \gamma_{\text{eff}}^{-} = \sum_{\ell=1}^{\ell_{\max}} \ell\gamma_\ell^{-}. \tag{3.4}$$

A numerical implementation of this dynamics is described at the end of Section 3.5. We use the simulations to verify analytic results of the following sections.

Statistical properties of interest For a systematic analysis of the statistical features of sequences generated by our dynamical model we start from the most simple statistical properties and then gradually advance the complexity of the measures under investigation. In a field-theoretic framework, the simplest non-trivial quantity of a one-dimensional system is its sequence length $\langle N \rangle$, also called the system's zero-point function. Ascending one step on the ladder of complexity, the one-point function of the system measures the composition bias $\langle s_k \rangle$ at a single position k of the sequence. Consequently, the two-point function measures the correlations in the composition of two different sequence positions, often specified as a function of the distance between the two sites along the sequence. In a natural manner, n-point functions with $n > 2$ are then defined as higher-order correlations between n different points along the sequence. All of these quantities are in fact specified by distributions of measured values in single sequence realizations over the full dynamical ensemble. Thus, for each n-point function, there are also different levels of complexity depending on whether we are interested in the distribution, or only in some of its elementary features, for example its mean. For many cases, the latter is obviously far easier to obtain than the full distribution.

We first shortly recall known results of the average sequence length in the next section and will then present a detailed analysis of the average sequence composition bias. In Section 3.3, stationary solutions for the two-point function will be derived. The full distribution function of the average composition bias in sequences of finite length will be calculated in Section 3.4. In Section 3.5, we will investigate extensions of the model to biased insertion rates and asymmetric mutation rates and discuss the universality of our model. Technical details and numerical implementation of the measurement of correlation functions and finite-size composition bias distribution are described at the end of Section 3.5. A recapitulation of the non-stationary build-up and decay of correlations in more complicated scenarios of dynamical process rates will be presented in Section 3.6. The general four-letter model and the web service CorGen are described in Section 3.7. We will conclude this chapter with the discussion of a possible causal connection between the observed correlations in eukaryotic genomes and the mutational processes of our evolutionary model in Section 3.8.

3.2 Sequence growth and average composition

Average sequence length Running the processes defined in (3.1) on sequences will change their lengths $N(t)$. The dynamics of $\langle N \rangle(t)$ averaged over an ensemble

of sequences is determined by the following differential equation

$$\frac{\partial}{\partial t}\langle N\rangle(t) = \left[\sum_{\ell=1}^{\langle N\rangle(t)} \ell\sigma(\delta_\ell - \gamma_\ell^-) + \sum_{\ell=1}^{\ell_{\max}} \ell\gamma_\ell^+\right]\langle N\rangle(t). \tag{3.5}$$

The finite size correction factor $\sigma = 1 - (\ell-1)/\langle N\rangle(t)$ accounts for the fact that in a sequence of length $N(t)$ there are only $N(t) - \ell + 1$ possibilities to duplicate or delete a segment of length ℓ. Using the initial condition $N(t=0) = N_0$, the solution of (3.5) in the asymptotic regime $\langle N\rangle(t) \gg \ell_{\max}$ is then given by

$$\langle N\rangle(t) = N_0 \exp(\lambda t) \tag{3.6}$$

with the asymptotic growth rate λ, as defined in Eq. (3.2). The distribution of sequence lengths in the ensemble can be obtained by mapping our dynamical model on a standard *branching process*. For a detailed discussion, see e. g. [53, 106].

Average composition bias The average composition of a sequence element s_k is measured by the expectation value $\langle s_k\rangle(t)$, and in the following we will show that any initial bias decays due to mutations and random insertions. In our binary model, $\langle s_k\rangle(t)$ can be written as the difference

$$\langle s_k\rangle(t) = P_k^+(t) - P_k^-(t). \tag{3.7}$$

$P_k^+(t)$ and $P_k^-(t)$ denote the probabilities of finding $s_k = +1$ or $s_k = -1$ at time t. The Master equations for $P_1^\pm(t)$ of the first sequence site s_1 are

$$\frac{\partial}{\partial t}P_1^\pm(t) = \mu[P_1^\mp - P_1^\pm] + \sum_{\ell=1}^{\ell_{\max}} \gamma_\ell^- \,[P_{1+\ell}^\pm - P_1^\pm]. \tag{3.8}$$

The first term on the right hand side specifies the rate of change of $P_1^\pm$ due to mutation of the first site. The second term results from deletions of segments $(s_1, \ldots, s_\ell)$, which will cause $s_{1+\ell}$ to become the new first site of the sequence. Omitting deletion $(\gamma_\ell^- = 0)$ and starting with a single site $\vec{s}(t=0) = (+1)$, we obtain

$$\langle s_1\rangle(t) = \exp(-2\mu t). \tag{3.9}$$

If one additionally allows deletions, any initial bias of s_1 will decay even faster.

Sequence sites s_k at positions $k > 1$ are also affected by duplications and insertions, and the Master equations for the probabilities $P_k^\pm(t)$ take the form

$$\begin{aligned}\frac{\partial}{\partial t}P_k^\pm(t) &= \mu\,[P_k^\mp - P_k^\pm] + \sum_{\ell=1}^{\ell_{\max}} \min(k-1,\ell)\,\gamma_\ell^+\,(1/2 - P_k^\pm) \\ &+ \sum_{\ell=1}^{k-2}(k-l-1)\,\gamma_\ell^+\,[P_{k-l}^\pm - P_k^\pm] + \sum_{\ell=1}^{k-1}(k-l)\,\delta_\ell\,[P_{k-l}^\pm - P_k^\pm] \\ &+ \sum_{\ell=1}^{\ell_{\max}} k\,\gamma_\ell^-\,[P_{k+l}^\pm - P_k^\pm].\end{aligned} \tag{3.10}$$

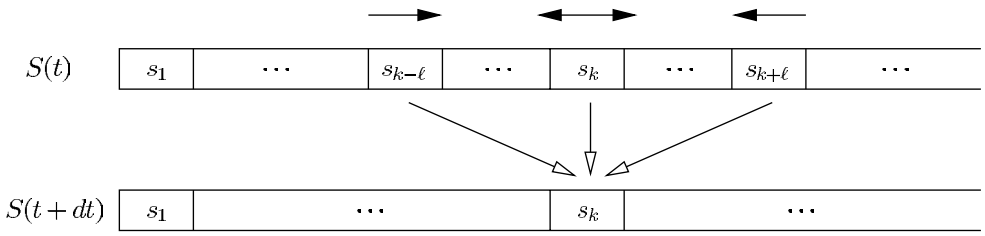

Figure 3.2: Illustration of the different mechanisms contributing to $\partial P_k^{\pm}(t)/\partial t$.

The different mechanisms contributing to $\partial P_k^{\pm}(t)/\partial t$ are illustrated in Fig. 3.2. Any bias at site s_k is again diminished due to single-site mutations, as specified by the first term on the r.h.s. of Eq. (3.10), but also by insertions of random segments $(x_i, \ldots, x_{i+\ell-1})$ of length ℓ at positions $i = k-\ell+1, \ldots, k$, which effectively randomize s_k (second term). In addition, there is a "shift" of composition bias from preceding sequence positions $s_{k-\ell}$ due to insertions of random segments $(x_i, \ldots, x_{i+\ell-1})$ of length ℓ at positions $i = 2, \ldots, k-\ell$ (third term), or duplications of existing sequence segments $(s_i, \ldots, s_{i+\ell-1})$ with $i = 1, \ldots, k-\ell$ (fourth term). Transport of bias from sites $s_{k+\ell}$ to s_k, on the other hand, occurs due to deletion of existing segments $(s_i, \ldots, s_{i+\ell-1})$ with $i = 1, \ldots, k$ (last term).

In order to reveal the large-distance asymptotics of this dynamics for $k \gg \ell_{\max}$ and in large sequences with $N(t) \gg \ell_{\max}$, we carry out a continuum limit of Eq. (3.10), i.e., we replace the discrete index k by a continuous variable and write $\langle s(k,t)\rangle \equiv \langle s_k\rangle(t)$. Using Eq. (3.7) we obtain a differential equation describing the asymptotic dynamics,

$$\frac{\partial}{\partial t}\langle s(k,t)\rangle = -2\mu_{\text{eff}}\langle s(k,t)\rangle - \lambda k\frac{\partial}{\partial k}\langle s(k,t)\rangle, \tag{3.11}$$

with the asymptotic growth rate λ and the effective mutation rate μ_{eff} defined in Eq. (3.2) and Eq. (3.3). The transport of composition bias due to the net exponential expansion of the sequences thereby gets incorporated in a dilatation operator of the functional form $k\partial/\partial k$; all finite size effects vanish in this regime. Eq. (3.11) has a solution of the form

$$\langle s(k,t)\rangle = e^{-2\mu_{\text{eff}}t}\mathcal{S}(ke^{-\lambda t}). \tag{3.12}$$

$\mathcal{S}(x)$ is a scaling function. This solution describes two different regimes of the expectation value, depending on the boundary condition chosen. (a) With fixed initial condition $s_1(t=0) = 1$, we have for any fixed k

$$\langle s(k,t)\rangle \propto \exp(-2\mu_{\text{eff}}t), \tag{3.13}$$

as shown in Fig. 3.3 (a) for different values of k and a given set of process rates. Thus, $\langle s(k,t)\rangle = 0$ for all k in the limit $t \to \infty$. (b) With fixed boundary condition

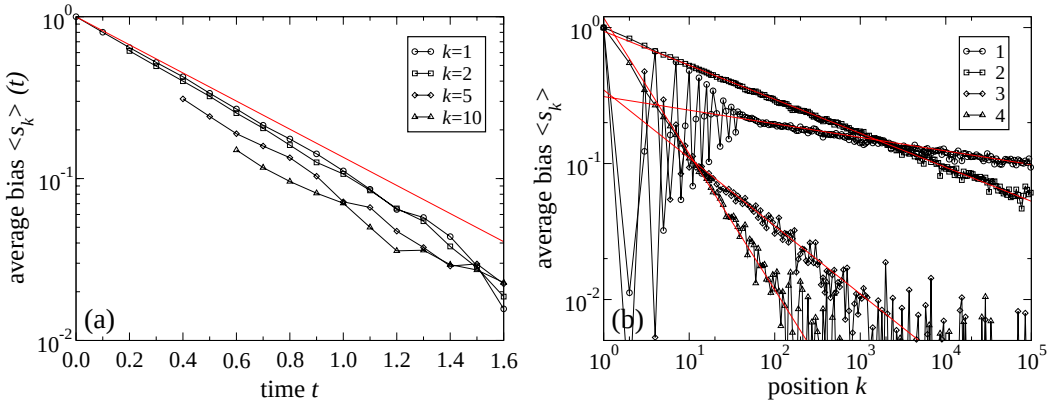

Figure 3.3: Average composition bias $\langle s_k \rangle(t)$: (a) Decay of $\langle s_k \rangle(t)$ in time for $k = 1, 2, 5, 10$. Rates of the processes are: $\mu = 1.0, \delta_1 = 4.0, \gamma_5^+ = 0.2, \gamma_2^- = 0.5$. The red line is the analytic upper bound on the rate of convergence (3.13).(b) Stationary $\langle s_k \rangle$ with fixed $\langle s_1 \rangle = +1$ at different rates of the elementary processes: (1) $\mu = 1.0, \delta_3 = 15.0, \gamma_2^+ = 1.0, \gamma_7^- = 1.0$; (2) $\mu = 1.0, \delta_1 = 16.0, \gamma_2^+ = 1.0, \gamma_1^- = 2.0$; (3) $\mu = 1.0, \delta_2 = 6.0, \gamma_3^+ = 2.0, \gamma_4^- = 0.5$; (4) $\mu = 1.0, \delta_1 = 4.0, \gamma_2^+ = 1.0, \gamma_4^- = 0.5$. Red lines denote the corresponding analytic asymptotics (3.14). All ensemble averages were obtained by averaging over 10^6 simulated sequences.

$\langle s_1 \rangle = +1$ for all t (i.e., suppressing mutations of the first element), we obtain a power-law decay of the composition bias along the sequence,

$$\langle s(k) \rangle \propto k^{-\chi} \quad \text{with } \chi = \frac{2\mu_{\text{eff}}}{\lambda}. \tag{3.14}$$

Numerical verification of the asymptotics (3.14) for this type of dynamics is presented in Fig. 3.3 (b), where we show the measured $\langle s_k \rangle$ in ensembles of sequences with different sets of rates using the simulation algorithm described in Section 3.5.

3.3 Stationary two-point correlations

Master equation The dynamics of the two-point composition correlation function $C(k, r, t) = \langle s_k s_{k+r} \rangle(t)$ between two sequence positions s_k and s_{k+r} can be derived by writing it as

$$C(k, r, t) = P_{\text{eq}}(k, r, t) - P_{\text{op}}(k, r, t). \tag{3.15}$$

$P_{\text{eq/op}}(k, r, t)$ denote the joint probabilities of simultaneously finding two equal or opposite symbols, respectively, at sequence positions k and $k + r$ and time t. For simplicity, we start with a restricted sequence evolution model where all processes are limited to single sequence sites ($\ell_{\max} = 1$). The Master equation for $P_{\text{eq}}(k, r, t)$

in the single-site model takes the form

$$\begin{aligned}
\frac{\partial}{\partial t}P_{\mathrm{eq}}(k,r,t) &= 2\mu\left[P_{\mathrm{op}}(k,r) - P_{\mathrm{eq}}(k,r)\right] && \text{(3.16a)}\\
&+ 1/2\,\gamma_1^+\left[P_{\mathrm{op}}(k,r) - P_{\mathrm{eq}}(k,r)\right] && \text{(3.16b)}\\
&+ 1/2\,\gamma_1^+\left[P_{\mathrm{op}}(k-1,r) - P_{\mathrm{eq}}(k-1,r)\right] && \text{(3.16c)}\\
&+ 1/2\,\gamma_1^+\left[P_{\mathrm{eq}}(k-1,r) - P_{\mathrm{eq}}(k,r)\right] && \text{(3.16d)}\\
&+ \left[(r-1)\gamma_1^+ + r\delta_1\right]\left[P_{\mathrm{eq}}(k,r-1) - P_{\mathrm{eq}}(k,r)\right] && \text{(3.16e)}\\
&+ r\gamma_1^-\left[P_{\mathrm{eq}}(k,r+1) - P_{\mathrm{eq}}(k,r)\right] && \text{(3.16f)}\\
&+ \left[(k-2)\gamma_1^+ + (k-1)\delta_1\right]\left[P_{\mathrm{eq}}(k-1,r) - P_{\mathrm{eq}}(k,r)\right] && \text{(3.16g)}\\
&+ k\gamma_1^-\left[P_{\mathrm{eq}}(k+1,r) - P_{\mathrm{eq}}(k,r)\right]. && \text{(3.16h)}
\end{aligned}$$

The different mechanisms contributing to $\partial P_{\mathrm{eq}}(k,r,t)/\partial t$ are illustrated in Fig. 3.4 and will now be discussed in order. The term (3.16a) describes the change in $P_{\mathrm{eq}}(k,r,t)$ due to mutation of any of the two sites (therefore two possibilities) in a pair of equal or opposite symbols at positions k and $k+r$. Term (3.16b) treats the insertion of a random site at position $k+r$, which in half of the cases will switch a pair of equal symbols $s_k = s_{k+r}$ to opposing symbols $s_k = -s_{k+r}$, whereas two opposing symbols might be switched to equal symbols accordingly. A similar contribution arises from a random insertion at position k. However, such an event can be regarded as duplication of s_{k-1} with a successional mutation of the newly introduced element s_k in half of the cases. If such a mutation occurs, the event is equivalent to term (3.16b) with the difference that contributions of this processes to $\partial P_{\mathrm{eq}}(k,r,t)/\partial t$ do now depend on the joint probabilities $P_{\mathrm{eq/op}}(k-1,r,t)$ (3.16c). In the other half of the cases, where the newly inserted random element s_k is equal to s_{k-1}, the process causes a shift of joint probability from $P_{\mathrm{eq}}(k-1,r,t)$ to $P_{\mathrm{eq}}(k,r,t)$ (3.16d). Transport of joint probability at distance $r-1$ to such at distance r takes place if a random site is inserted at sequence positions $k+1,\ldots,k+r-1$, or if any site $s_k,\ldots,s_{k+r-1}$ is duplicated (3.16e). On the other hand, deletion of any $s_{k+1},\ldots,s_{k+r}$ produces a transport of joint probability from distance $r+1$ to r (3.16f). Despite this "expansion" and "contraction" transport of joint probability from distances $r+1$ or $r-1$ to r at fixed k, there is also a "horizontal" shift along the sequence: insertion of a random site at positions $2,\ldots,k-1$ or duplication of any site $s_1,\ldots,s_{k-1}$ shifts joint probability $P_{\mathrm{eq}}(k-1,r,t)$ to $P_{\mathrm{eq}}(k,r,t)$ (3.16g). Deletion of an $s_1,\ldots,s_k$ shifts $P_{\mathrm{eq}}(k+1,r,t)$ to $P_{\mathrm{eq}}(k,r,t)$ (3.16h).

Notice that in contrast to [111], the Master equation stated above is exact and does not yet make use of specific sequence-inherent symmetries, e. g. translational invariance of $P_{\mathrm{eq/op}}(k,r,t)$ along the sequence. Thus, in its general form it also holds for position-dependent rates if products of the form rates × distances in (3.16e)-(3.16h) are substituted by integrals over the corresponding intervals.

Because we are interested in a stationary solution of this dynamics, we have to consider the limit $t \to \infty$. It has already been shown in Section 3.2 that asymptotically

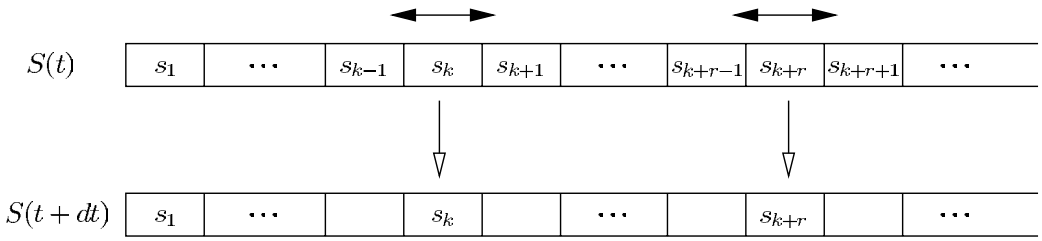

Figure 3.4: Illustration of the different mechanisms contributing to the dynamics of $P_{\rm eq}(k,r,t)$. Effectively mutational events are those that randomize either s_k, or s_{k+r}. "Expansion" or "contraction" transport of joint probability from $P_{\rm eq}(k,r\pm1)$ to $P_{\rm eq}(k,r)$ occurs due to duplication, insertion, or deletion events at sequence positions between s_k and s_{k+r}. "Horizontal" shift from $P_{\rm eq}(k\pm1,r)$ to $P_{\rm eq}(k,r)$ takes place if a duplication, insertion, or deletion occurs at sequence positions prior to s_k.

$\langle s_k\rangle(t)\to 0$ for large t at all k. Furthermore, all processes are acting homogeneously along the sequence, and therefore we expect the joint probabilities to be also independent of k in the long-time limit, i.e., $P_{\rm eq/op}(k,r)=P_{\rm eq/op}(k\pm1,r)$ (verification is given by our numerical simulations). Eq. (3.16h) then simplifies to

$$\begin{aligned}\frac{\partial}{\partial t}P_{\rm eq}(r,t) &= (2\mu+\gamma_1^+)\,[P_{\rm op}(r)-P_{\rm eq}(r)]\\ &+ [(r-1)\gamma_1^+ + r\delta_1]\,[P_{\rm eq}(r-1)-P_{\rm eq}(r)]\\ &+ r\gamma_1^-\,[P_{\rm eq}(r+1)-P_{\rm eq}(r)].\end{aligned}\tag{3.17}$$

By exchanging $P_{\rm eq}$ and $P_{\rm op}$, we can state an equivalent equation for $P_{\rm op}(r,t)$. Using (3.15), we obtain the dynamics of the correlation function $C(r,t)$ for large t

$$\begin{aligned}\frac{\partial}{\partial t}C(r,t) &= -(4\mu+2\gamma_1^+)\,C(r)\\ &+ [(r-1)\gamma_1^+ + r\delta_1]\,[C(r-1)-C(r)]\\ &+ r\gamma_1^-\,[C(r+1)-C(r)].\end{aligned}\tag{3.18}$$

This equation for the dynamics of $C(r,t)$ in the single-letter model ($\ell_{\rm max}=1$) is valid for all distances r in the limit $t\to\infty$. A corresponding dynamics can, in principle, be obtained analogously for the general model with $\ell_{\rm max}>1$, although it will be more complicated due to finite size effects coming into play for $r<\ell_{\rm max}$. However, for large distances $r\gg\ell_{\rm max}$, these finite size effects can be neglected, and the asymptotic dynamics of $C(r,t)$ in the general segmental model is then given by

$$\begin{aligned}\frac{\partial}{\partial t}C(r,t) &= -4\mu_{\rm eff}\,C(r)\\ &+ \sum\nolimits_{\ell=1}^{\ell_{\rm max}} [(r-\ell)\gamma_\ell^+ + (r-\ell+1)\delta_\ell]\,[C(r-\ell)-C(r)]\\ &+ \sum\nolimits_{\ell=1}^{\ell_{\rm max}} r\gamma_\ell^-\,[C(r+\ell)-C(r)]\end{aligned}\tag{3.19}$$

with the effective mutation rate μ_{eff}, as defined in Eq. (3.3). Note that the dynamics (3.18) of the single-letter model is a special case of the general dynamics (3.19) with $\ell_{\max} = 1$.

Stationary solutions In the following, we will derive analytic solutions of the stationary correlations $C(r)$ in our model. We start by shortly recapitulating the analytic results derived in [111] for the instructive special case of only single-site duplications and mutations ($\mu, \delta_1 > 0$, all other rates are zero). In this case, the solution of the dynamics (3.19) in the stationary state, $\partial C(r,t)/\partial t = 0$, obeys the recursion equation

$$C(r) = \frac{r}{\alpha + r}\, C(r-1) \quad \text{with} \quad \alpha = \frac{4\mu}{\delta_1}. \tag{3.20}$$

Using $C(0) \equiv 1$, the recursion can easily be solved, yielding

$$C(r) = \prod_{n=1}^{r} \frac{n}{\alpha + n}. \tag{3.21}$$

Introducing the gamma function and the beta function, defined by

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt, \qquad B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \tag{3.22}$$

$C(r)$ can finally be rewritten in the form

$$C(r) = \frac{\Gamma(r+1)\Gamma(1+\alpha)}{\Gamma(r+1+\alpha)} = \alpha B(r+1, \alpha). \tag{3.23}$$

To investigate the asymptotic regime, we evaluate the asymptotic behavior of $B(r,\alpha)$ for $r \gg 1$ which, in general, is given by

$$B(r,\alpha) \propto \Gamma(\alpha)\, r^{-\alpha} \left[1 - \frac{\alpha(\alpha-1)}{2r}\left(1 + O\left(\frac{1}{r}\right)\right)\right]. \tag{3.24}$$

Applying this asymptotics to Eq. (3.23) we obtain

$$C(r) \propto r^{-\alpha}. \tag{3.25}$$

Hence, we have proven the existence of long-range correlations in the simplified single-site duplication-mutation model. The exponent α is determined by a simple balance between the randomization processes (mutations) and the expansion processes (duplications) which create correlations between neighboring sites and transport these correlations to larger distances due to an overall expansion of the system.

We have performed extensive Monte Carlo simulations of this model using the algorithm presented at the end of Section 3.5. Fig. 3.5 (a) shows the numerical $C(r)$ for the duplication-mutation dynamics with various rates of δ_1 and μ, which is in excellent agreement with the analytic expression (3.23).

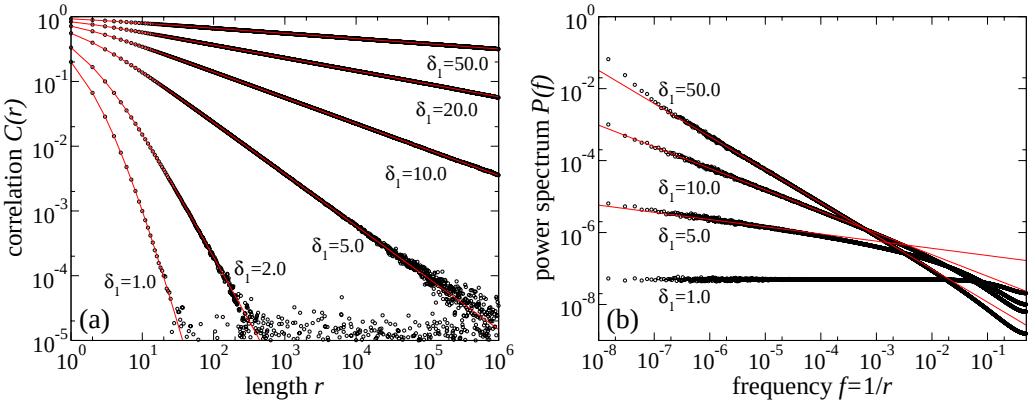

Figure 3.5: Single-site duplication-mutation model: (a) Stationary composition correlations $C(r)$ at different rates of the elementary processes; numerical results (circles) and the analytic form (3.23) (lines) for $\mu = 1.0$, δ_1 varying. $C(r)$ is averaged along the sequence. (b) Power spectra of simulated sequences for $\mu = 1.0$ and δ_1 varying: numerical results (circles) with the analytically predicted $P(f) \propto f^{-\beta}$ in those cases where $\delta_1 \geq 5$ (lines). The dynamics of the sequences was simulated until they reached a length of $N = 2^{27} \approx 10^8$. All data sets were obtained by averaging over 100 runs. Plots taken from [106].

For reasons of comparability with former studies [90, 91], we also calculated power spectra of the simulated sequences. In the stationary state, the power spectrum $P(f)$ is the Fourier transform of the correlation function $C(r)$. In our case, the large distance asymptotics of the correlation function is given by $C(r) \propto r^{-\alpha}$, and the power spectrum will therefore also decay algebraically, i.e., $P(f) \propto f^{-\beta}$ with exponent $\beta = 1 - \alpha$ [151]. The resulting data is shown in Fig. 3.5 (b). Due to the fact that $C(r) \propto r^{-\alpha}$ does only hold in the limit $r \gg 1$, the analytically estimated scaling $P(f) \propto f^{-\beta}$ is present at lower frequencies, but crosses over to a different behavior at higher ones. At values $\alpha > 1$, $C(r)$ decays below the fluctuation threshold $\Delta C = 1/\sqrt{N(t)}$ [173] before the scaling gets established, obviating the appearance of positive exponents β. In those cases, we measure a flat power spectrum in the low frequency part as one expects for random sequences. The finite size deviations of $C(r)$ at very large r show up in the low frequency part of the power spectra, too.

Obviously one cannot expect the stationary $C(r)$ of the general model to be described by a similar simple expression as has been obtained for the single-site duplication-mutation dynamics in (3.23). Consider, for example, a segmental duplication process, copying segments of length $\ell_1 = 50$. In case this is the only duplication process present, it will introduce a peak in $C(r)$ at a distance corresponding to its segment length $r = \ell_1$. If there is an additional duplication processes present, e.g. one with $\ell_2 = 1$, the peak in $C(r)$ established by the first duplication process will be shifted to larger distances by the second process. The functional form of $C(r)$ will thus show complex behavior on short scales reflecting the "microscopic" details of the elementary processes (see Fig 3.6). But what about the large-distance asymptotics of $C(r)$ for $r \gg \ell_{\max}$? In this regime, the dynamics of $C(r,t)$ is determined by Eq. (3.19). Carrying out a continuum limit, the difference equation (3.19) can again

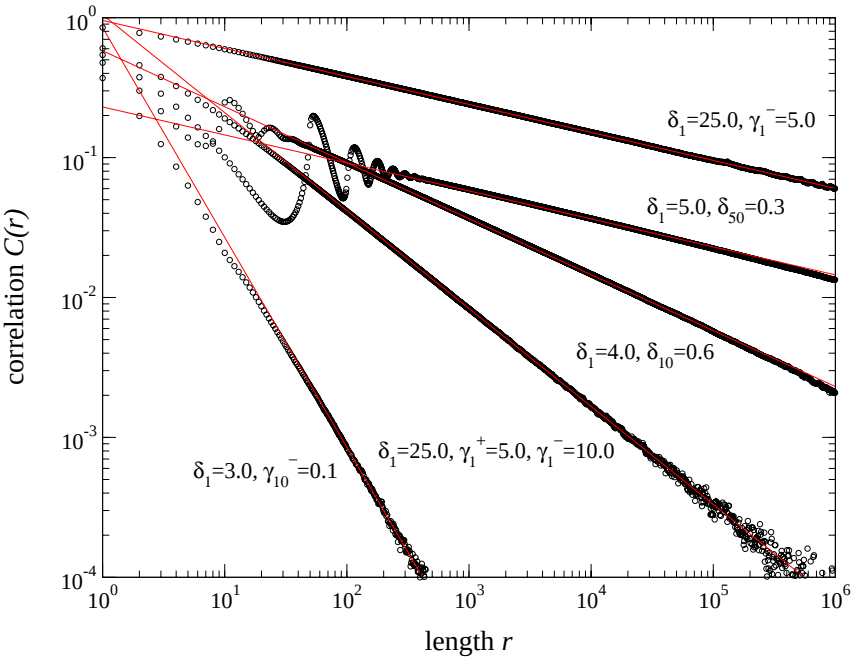

Figure 3.6: Stationary $C(r)$ for different rates of the elementary processes in the general model with various segmental processes present: numerical results (circles) coincide accurately with the analytic asymptotics (3.27) (lines) in the large distance regime. Mutations occurred at rate $\mu = 1.0$. Rates not specified in the plot are zero.

be written as a simple differential equation,

$$\frac{\partial}{\partial t}C(r,t) = -4\mu_{\text{eff}}C(r,t) - \lambda r\frac{\partial}{\partial r}C(r,t). \tag{3.26}$$

The stationary solution of Eq. (3.26) immediately yields the power-law decay

$$C(r) \propto r^{-\alpha} \quad \text{with} \quad \alpha = 2\chi = \frac{4\mu_{\text{eff}}}{\lambda}. \tag{3.27}$$

Hence, on macroscopic distances $r \gg \ell_{\max}$ our model universally produces long-range correlations in the sequences, irrespectively of the microscopic details of the individual processes. The decay exponent α depends on only two effective parameters, which are simple functions of the rates of the processes. Using these analytic results we can furthermore qualitatively classify the four different types of processes according to whether they increase α, or decrease it. Duplications are the only processes with $\partial\alpha/\partial\delta_\ell < 0$ because they increase the growth rate λ, but have no effectively mutational influence on large scales. All other processes, in contrast, will lead to larger values of α and thus to faster decaying correlations by increasing their rates.

To verify these analytic results, we show the measured correlation functions $C(r)$ of simulated sequences with all sorts of different processes present in Fig. 3.6. Whereas on short scales the correlations reveal the microscopic details of the particular processes, in the asymptotic regime long-range correlations are ubiquitous. Their func-

tional form is accurately described by our analytics (3.27) with the effective rates defined in Eq. (3.2) and Eq. (3.3).

3.4 Finite-size distribution of the composition bias

Up to this point, we have discussed correlation functions, which were defined as averages over an ensemble of sequences generated by the same stochastic dynamics. What can we say about the data of a single sequence, i.e., a single realization of the stochastic process? To address this question, we now consider the distribution of the composition bias evaluated in finite sequence intervals $k, \ldots, k+L-1$ of length L,

$$m = \frac{1}{L} \sum_{k'=k}^{k+L-1} s_{k'}. \tag{3.28}$$

Generalizing Eq. (3.11) and (3.26), we obtain the following differential equation for the distribution function $P(m, L, t)$,

$$\begin{aligned} \frac{\partial}{\partial t} P(m, L, t) = & - \lambda L \frac{\partial}{\partial L} P(m, L, t) \\ & + 2\mu_{\text{eff}} \frac{\partial}{\partial m}[mP(m, L, t)] + \frac{2\mu_{\text{eff}}}{L} \frac{\partial^2}{\partial m^2} P(m, L, t), \end{aligned} \tag{3.29}$$

which is valid again in a continuum approximation for $L \gg 1$. The three terms on the r.h.s. describe, in order, the transport of the composition bias due to the exponential dilatation of the sequence, its dissipative decay, and its stochastic fluctuations. Notice that the last two terms are caused by the same basic mutation process and are therefore both proportional to μ_{eff}.

We limit ourselves here to evaluating the equilibrium distribution $P(m, L)$ asymptotically for large values of L. The solution of Eq. (3.29) defines two different parameter regimes with transition point $\chi = 1/2$:

1. **Strong-correlation regime ($\chi < 1/2$):** The large-L asymptotics is determined by balancing dilatation and deterministic decay, i.e., the first two terms on the r.h.s. of Eq. (3.29). For this regime, we obtain

$$P(m, L) = L^{\chi} \mathcal{P}_{\chi}(x) \quad \text{with} \quad x = mL^{\chi}, \tag{3.30}$$

where $\mathcal{P}_{\chi}(x)$ is a scaling function (whose form is determined by the stochastic dynamics on smaller scales). We can verify the consistency of the solution (3.30) by checking that the third term on the r.h.s. of (3.29) gives a contribution which is subleading by a factor $L^{2\chi-1}$ for large L. This result is also verified by our numerics, as shown in Fig. 3.7 (a,b), where we present measured distributions $P(m, L)$ and the collapse into one scaling function $\mathcal{P}_{\chi}(x)$. The scaling of

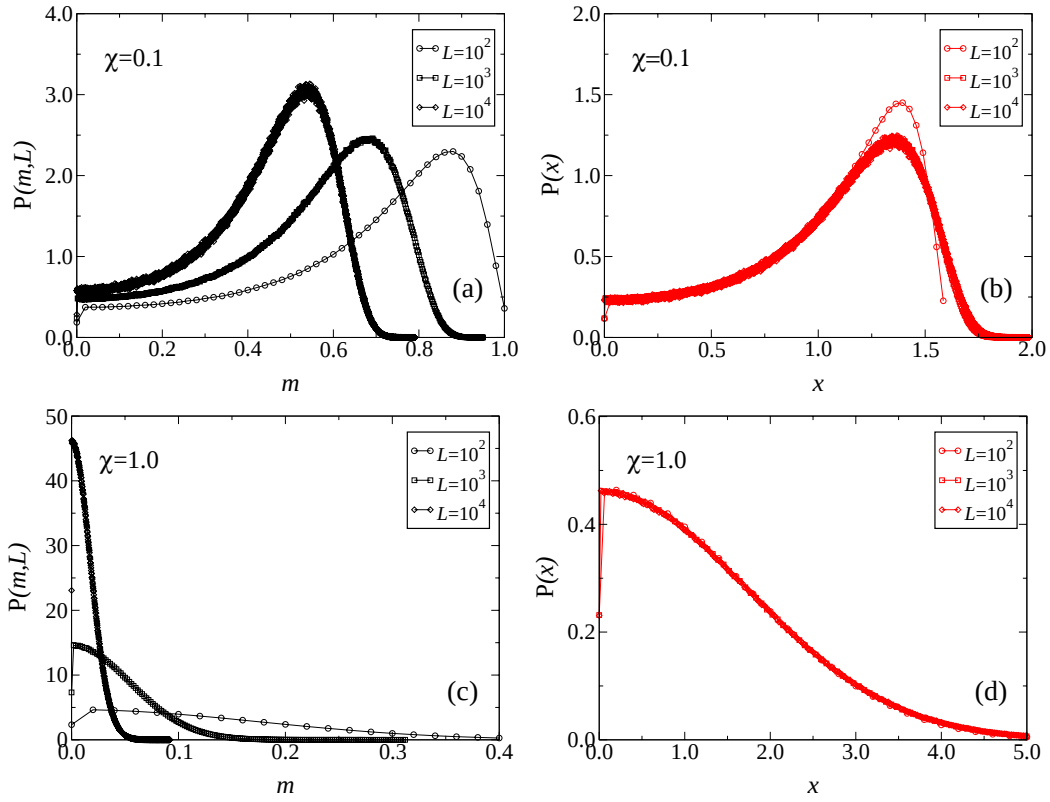

Figure 3.7: Numerically measured distribution functions $P(m, L)$ and the corresponding scaling functions $\mathcal{P}(x)$ for $L = 10^2, 10^3, 10^4$. (a,b) Regime 1. with $\chi = 0.1$ and $\mathcal{P}(x) = L^{-0.1}P(L^{-0.1}x, L)$. (c,d) Regime 2. with $\chi = 1.0$ and the Gaussian scaling function $\mathcal{P}(x) = L^{-1/2}P(L^{-1/2}x, L)$. The deviations for $L = 10^2$ for both regimes are due to the fact that the analytic asymptotics is only valid for large L. The ensemble averages were obtained by averaging over 10^7 sequence realizations for each parameter setting with random initial conditions, resulting in symmetric distributions (only positive values shown).

$P(m, L)$ also determines the scaling of its moments $\langle m^k \rangle(L) \equiv \int m^k P(m, L) dm$,

$$\langle m^k \rangle(L) \propto L^{-k\chi}. \tag{3.31}$$

This is consistent with the scaling of the one-point and two-point functions, obtained in Eq. (3.14) and (3.27).

2. **Weak-correlation regime ($\chi > 1/2$):** Eq. (3.29) has an exact solution of Gaussian form,

$$P(m, L) = \sqrt{\frac{L}{2\pi\xi(\chi)}} \exp\left[-\frac{(m - m_0 L^{-\chi})^2 L}{2\xi(\chi)}\right] \tag{3.32}$$

with $\xi(\chi) = \chi/(\chi - 1/2)$. This solution has the expectation value

$$\langle m \rangle(L) = m_0 L^{-\chi} \tag{3.33}$$

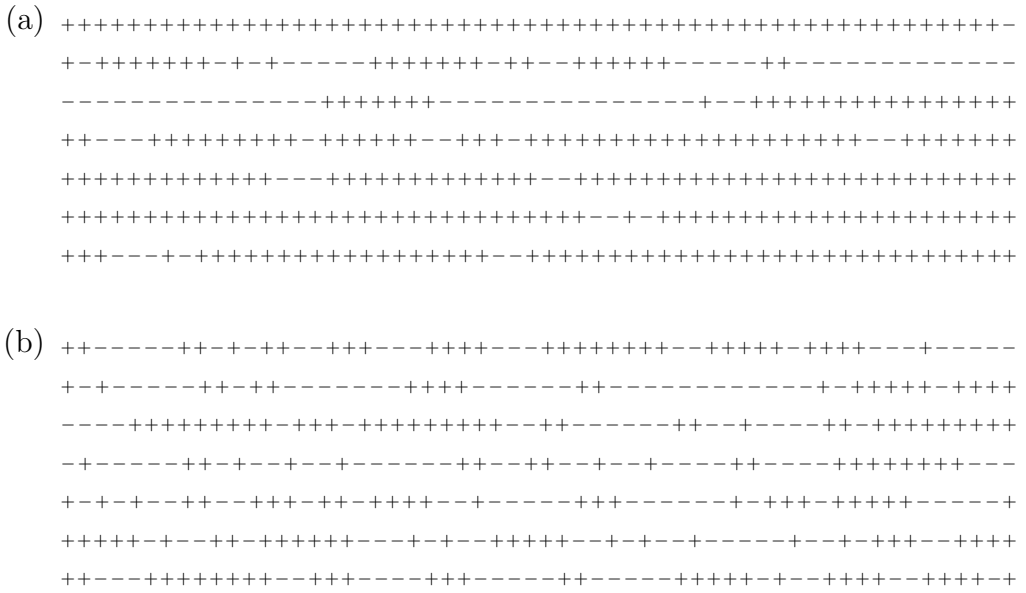

Figure 3.8: A single sequence of length $N = 400$ generated by the expansion-randomization process from an initial letter $+1$. (a) Strong-correlation regime ($\mu = 0.5$, $\delta_1 = 10.0$, i.e. $\chi = 0.1 < 1/2$): The sequence retains a net composition bias towards $+1$ in its entire length, i.e., the initial composition bias is detectable. Minority islands of -1 are found on all scales. (b) Weak-correlation regime ($\mu = 0.5$, $\delta_1 = 1.0$, i.e. $\chi = 1.0 > 1/2$): The sequence consists of strongly correlated islands of length $\xi \approx 5$ but looks random on larger scales. The initial composition bias is not detectable.

(with the coefficient m_0 determined by the initial condition) and the variance

$$\langle m^2\rangle(L) - \langle m\rangle^2(L) = \frac{\xi(\chi)}{L}. \tag{3.34}$$

It is thus of similar form as the fluctuation-dissipation equilibrium $\exp[-m^2/2L]$ for $\lambda = 0$, obtained from the last two terms on the r.h.s. of (3.29). The transport term generates an additional length scale ξ because individual sites are not completely independent of each other but are strongly correlated on scales smaller than ξ due to duplications. This reduces the number of effectively independent fluctuating sequence segments to L/ξ. Numerical measurements of the distribution $P(m, L)$ in this regime for random initial conditions ($m_0 = 0$) and the corresponding scaling function $\mathcal{P}_\chi(x) \propto \exp[-x^2/2\xi(\chi)]$ with $x \equiv mL^{1/2}$ are shown in Fig 3.7 (c,d).

3. **Transition point ($\chi = 1/2$):** The solution is still of Gaussian form,

$$P(m, L) = \sqrt{\frac{L}{2\pi \log L}} \exp\left[-\frac{(m - m_0 L^{-\chi})^2 L}{2 \log L}\right]. \tag{3.35}$$

The existence of two different scaling regimes has direct consequences for the detectability of correlations from data of a single sequence on large scales. In the

strong-correlation regime ($\chi < 1/2$), the composition bias on arbitrary large scales L is determined primarily by the ancestral bias, while the mutational fluctuations can be neglected asymptotically. In the weak-correlation regime, the ancestral bias can only be detected on scales $L < L^*$. The mutational noise is dominant on larger scales. The scale L^* can be estimated by equating the average $\langle m \rangle(L^*)$ with the root mean square deviation $[\langle (m - \langle m \rangle)^2 \rangle (L^*)]^{1/2}$ from Eq. (3.33) and (3.34).

The difference between the strong- and weak-correlation regime is illustrated in Fig. 3.8, where we show two single sequences generated from an ancestor letter $+1$. In the strong-correlation regime (a), the entire sequence has a detectable bias towards $+1$, with islands of -1 tracing back to their ancestors generated by mutation events. In the weak-correlation regime (b), the sequence is seen to consist of strongly correlated segments of length $\xi \approx 5$, but it looks random on larger scales.

We stress again that the existence of two different scaling regimes with a transition at $\chi = 1/2$ is a feature of the full distribution $P(m, L)$ in the asymptotic regime $L \gg 1$. Expectation values such as the composition bias (3.14) and the correlation function (3.27) have a universal form in both regimes and no transition at $\chi = 1/2$.

3.5 Symmetry breaking and universality

Biased insertions In the following, we will investigate generalizations of the dynamical model and thereby demonstrate the universality of our approach. For simplicity, we again start with a single-letter model ($\ell_{\max} = 1$). In contrast to the original model of Section 3.1, where random insertions were defined as insertions of random letters $x = \pm 1$ at position $k + 1$ independent of the preceding sequence element s_k, we now want to consider biased insertions. This extension is biologically well motivated. There is ample evidence by now that the rates of segmental insertions into the genome, e.g. those of interspersed repeats, are biased by the local GC-content of the genomic region [68]. Formally, the biased insertion process in our model is defined by

$$(\cdots, s, \cdots) \quad \to \quad (\cdots, s, y[s], \cdots) \qquad \text{insertion rate } \eta, \tag{3.36}$$

where $y[s]$ denotes a randomly chosen letters $y[s] = \pm 1$ with an average bias depending on the value of the preceding sequence element s,

$$\langle y[s] \rangle = \nu s, \qquad \nu \in [-1, 1]. \tag{3.37}$$

The degree of dependence can be tuned by a parameter ν. In fact, the random insertions of the original model are the special case of this generalized process using $\nu = 0$, $\nu = 1$ corresponds to duplications.

The contributions of this process to the dynamics of the joint-probabilities $P_{\text{eq/op}}(r, t)$ can still be calculated exactly. Terms (3.16a) and (3.16e)-(3.16h) will not be affected

because the biased insertion process will neither change the effect of single-site mutations, nor the "shift" and "transport" of joint-probability. However, an additional multiplicative factor $(1-\nu)$ has to be incorporated in terms (3.16b) and (3.16c). Effects on (3.16d) are described by an additional factor $(1+\nu)$. Concerning the Master equation for $C(r)$ in the continuum limit (3.26), this biased insertion process does therefore not affect the asymptotic growth rate λ. The effective mutation rate though, is now given by

$$\mu_{\text{eff}} = \mu + \frac{1}{2}(1-\nu)\eta. \tag{3.38}$$

We want to mention that the biased insertion of single letters can generically be extended to the biased insertion of segments $(y[s])_\ell$ at a rate η_ℓ with an average bias of their elements $\langle y_i[s]\rangle = \nu_\ell s$. In this case, one might actually have $\nu_\ell = \nu(\ell)$, and asymptotically for the effective mutation rate we obtain

$$\mu_{\text{eff}} = \mu + \frac{1}{2}\sum_{\ell=1}^{\ell_{\max}}(1-\nu_\ell)\eta_\ell. \tag{3.39}$$

Symmetry breaking The model considered so far was symmetric concerning $s_k \to -s_k$. It is known that this symmetry is not granted for genomic evolution. Distinct mutation rates of different nucleotides, for example, lead to unequal frequencies of the four different nucleotides along genomic DNA [12]. In the following, we will show that the restriction to symmetric processes is not crucial for the emergence of long-range correlations and the universal scaling of our model. A simple scenario breaking the model's Z_2 symmetry is the choice of asymmetric mutation rates,

$$(\cdots,+1,\cdots) \quad \to \quad (\cdots,-1,\cdots) \qquad \text{rate } \mu^+ \tag{3.40}$$
$$(\cdots,-1,\cdots) \quad \to \quad (\cdots,+1,\cdots) \qquad \text{rate } \mu^-, \tag{3.41}$$

with $\mu^+ \neq \mu^-$. In this scenario, the Master equations of the probabilities $P_k^\pm(t)$ are

$$\begin{aligned}\frac{\partial}{\partial t}P_k^\pm(t) = \quad & \pm\mu^- P_k^\mp \mp \mu^+ P_k^\pm + \sum_{\ell=1}^{\ell_{\max}} \min(k-1,\ell)\,\gamma_\ell^+\,(1/2 - P_k^\pm) \\ & + O\left(\sum_{\ell=-\ell_{\max}}^{\ell_{\max}} P_{k+\ell}^\pm - P_k^\pm\right),\end{aligned} \tag{3.42}$$

and we have already shown in Section 3.2 that $P_k^\pm$ is asymptotically independent of k if all sequence sites s_k are allowed to mutate. Thus, for the asymptotic stationary average composition bias $\langle s_k\rangle = P^+ - P^-$ in the asymmetric model we obtain

$$\langle s_k\rangle = \frac{\mu^- - \mu^+}{\mu^- + \mu^+ + 2\gamma_{\text{eff}}^+}. \tag{3.43}$$

Concerning the dynamics of the joint probabilities $P_{\mathrm{eq/op}}(r,t)$, the introduction of asymmetric mutation rates will only change the mutational term. The contributions of duplications, random insertions, and deletions will not be affected. In the asymmetric model, the Master equations for $P_{\mathrm{eq/op}}(r,t)$ are now given by

$$\frac{\partial}{\partial t}P_{\mathrm{eq}}(r,t) = \quad +(\mu^+ + \mu^-)P_{\mathrm{op}}(r) - 2\mu^+ P^{++}(r) - 2\mu^- P^{--}(r) + Q_{\mathrm{eq}}(r,t) \tag{3.44}$$

$$\frac{\partial}{\partial t}P_{\mathrm{op}}(r,t) = \quad -(\mu^+ + \mu^-)P_{\mathrm{op}}(r) + 2\mu^+ P^{++}(r) + 2\mu^- P^{--}(r) + Q_{\mathrm{op}}(r,t). \tag{3.45}$$

where $P^{++/--}(r)$ are the joint probabilities of simultaneously finding $s_k = s_{k+r} = +1$ and $s_k = s_{k+r} = -1$, respectively. $Q_{\mathrm{eq}}(r,t)$ denotes the terms (3.16b)-(3.16h) with the k-dependence of $P_{\mathrm{eq/op}}(r,t)$ already dropped. $Q_{\mathrm{op}}(r,t)$ is obtained by exchanging P_{eq} and P_{op}. The dynamics of $C(r,t)$ in the asymmetric model is therefore

$$\frac{\partial}{\partial t}C(r,t) = -2(\mu^+ + \mu^- + \gamma^+_{\mathrm{eff}})\,[C(r) + \langle s_k\rangle^2] + [Q_{\mathrm{eq}}(r,t) - Q_{\mathrm{op}}(r,t)], \tag{3.46}$$

where we used (3.43) and $\langle s_k\rangle = P^+ - P^- = P^{++}(r) + P^{+-}(r) - P^{-+}(r) - P^{--}(r)$ with $P^{+-}(r) = P^{-+}(r)$. Defining the effective mutation rate of the asymmetric model,

$$\tilde{\mu}_{\mathrm{eff}} = \frac{1}{2}(\mu^+ + \mu^- + \gamma^+_{\mathrm{eff}}), \tag{3.47}$$

the stationary solution of this dynamics in the continuum limit is now given by

$$C(r) \propto r^{-\alpha} + \langle s_k\rangle^2 \quad \text{with} \quad \alpha = 2\chi = \frac{4\tilde{\mu}_{\mathrm{eff}}}{\lambda}. \tag{3.48}$$

The magnitude of the segmental composition bias (3.28) scales as

$$\langle |m(L)|\rangle \propto L^{-\chi} + \langle s_k\rangle. \tag{3.49}$$

Hence, breaking the Z_2 symmetry by introducing asymmetric mutation rates will not change the long-range correlations and the general scaling of the model. It is obvious from Eq. (3.48) and (3.49) that the scaling still holds for the connected correlation function $C^c(r) \equiv \langle s_k s_{k+r}\rangle - \langle s_k\rangle^2$ and the shifted segmental composition bias $\langle 1/L|\sum_{k'=k}^{k+L-1} s_{k'}|\rangle - \langle s_k\rangle$.

Universality The structure of Eq. (3.26) reveals the basic mechanisms generating long-range correlations in a very general class of expansion-randomization systems that share three fundamental characteristics of their dynamics. The first feature is an overall exponential expansion of the system transporting correlations from shorter to larger sequence distances (combined effects of duplications, insertions, and deletions in our model). Mathematically this transport is described by a dilatation operator $r\partial/\partial r$ (second term on the r.h.s. of (3.26)). On the other hand, all correlations are counteracted by local processes randomizing the sequence (mutations + random insertions) and therefore trying to diminish $C(r)$ (first term of (3.26)). The competition

between expansion and randomization results in an algebraically decaying $C(r) \propto r^{-\alpha}$ in the stationary state with α determined by a simple ratio of effective growth rate to effective mutation rate. Calculation of these two fundamental parameters for any set of processes constituting such systems determines the large-distance asymptotics of the correlations in the generated sequences. However, $C(r) = 0$ for all r, is also a stationary solution of Eq. (3.26). For long-range correlations to be established, a third necessary feature of the dynamics is hence the presence of a mechanism continuously producing correlations on short scales. They serve as an ongoing reservoir for the transport of correlations to larger sequence distances and ensure the existence of a non-zero value $C(r_0) > 0$ for a specific $r_0 \geq 1$ (in our model, these initial correlations on short-scales are produced by duplications). As an intuitive example for the necessity of this third condition, consider an expansion-randomization system with mutations and insertions of single random letters, but no duplications. This system features exponential expansion, as well as local randomization. But the insertion process is not capable of producing $C(1) > 0$, and therefore no long-range correlations can be established in the generated sequences.

As expected from standard scaling theory, the decay of the two-point function has twice the exponent as the corresponding decay of the one-point function. The value χ can be interpreted as the scaling dimension of the variable s_k in this universality class. There is a one-parameter family of decay exponents as, for example, in the Gaussian model in two dimensions. This universal behavior is unaffected by the breakdown of the Z_2 symmetry, which manifests itself only in the non-universal constants in Eq. (3.49) and (3.48).

Numerical implementation Numerical simulation of the stochastic sequence dynamics (3.1) was implemented using a Monte Carlo procedure. During each discrete time step

$$\Delta t = \epsilon \cdot [(\mu + \sum\nolimits_{\ell}[\delta_\ell + \gamma_\ell^+ + \gamma_\ell^-])N(t)]^{-1} \tag{3.50}$$

with a tunable parameter $\epsilon \leq 1$, we choose a random site and randomly let a process act on it. The probability p_α of a process α being executed on the drawn site is

$$p_\alpha = \text{rate}(\alpha) \cdot \Delta t. \tag{3.51}$$

The overall probability of executing any process on the drawn site therefore depends on the parameter ϵ. Choosing $\epsilon = 1$ assures exactly one process being executed. For small $\epsilon \ll 1$, on he other hand, no process will be chosen to act on the drawn sites in most of the cases. We use $\epsilon = 0.1$ for our numerical simulations.

For a single realization of the stochastic dynamics, the average segmental composition bias and the correlation function are well approximated by sequence averages,

$$\langle |m| \rangle(L) \approx \frac{1}{N-L} \sum_{k=1}^{N-L} \frac{1}{L} \left| \sum_{k'=k}^{k+L-1} s_{k'} \right|, \tag{3.52}$$

$$C(r) \approx \frac{1}{N-r} \sum_{k=1}^{N-r} s_k s_{k+r} \tag{3.53}$$

if values of r and L are sufficiently small to allow efficient averaging. Averages over 100 sequence realizations reduce the noise further and produce very accurate measurements of $\langle |m| \rangle(L)$ and $C(r)$.

If the dynamics obeys Z_2 symmetry, we can directly infer the decay exponent α from these measurements, according to Eq. (3.31) and (3.25). However, if the Z_2 symmetry is violated, these power laws have to be disentangled from the additional constants $\langle s_k \rangle$ respectively $\langle s_k \rangle^2$, see Eq. (3.49) and (3.48). If the microscopic processes are known, these non-universal constants can be calculated. A numerical problem arises though in the analysis of genomic DNA sequences, where the Z_2 symmetry is broken by an unknown amount. In that case, we can self-consistently fit the data to the form $\langle |m| \rangle(L) = aL^{-\chi} + c$ and $C(r) = br^{-2\chi} + c^2$. Hence, the link between the finite-size scaling of $\langle |m| \rangle(L)$ and the scaling of the correlation function $C(r)$ dictated by universality is of practical importance for data analysis. In particular, it is not justified in general to approximate the constant c by $1/N \sum_{k=1}^{N} s_k$ for sequences of finite length N in the strong correlation regime $\chi < 1/2$, as it is often done in the literature. Furthermore, we can check consistency with the exponent $\beta = 1 - 2\chi$ of the GC power spectrum. Power spectra can easily be obtained using the *Fast Fourier Transform* algorithm [133].

3.6 Dynamical correlations

Up to now, results for the correlations $C(r)$ in our model have only been obtained for the stationary state reached in the limit $t \to \infty$. In this section, we want to focus explicitly on the dynamical behavior of $C(r, t)$ if process rates in our sequence evolution model are time-dependent. All results we are going to present in this section have already been established in [106], but a thorough understanding of the mechanisms of correlation build-up in growing sequences and decay of previously established correlations in sequences of constant length due to mutations will turn out to be crucial in Section 3.8 for investigating a possible connection between the theoretical results derived so far, and long-range correlations in real genomic sequences.

Correlation build-up When starting our sequence evolution model (3.1) with an initial sequence $S(t = 0) = (x)$, where $x = \pm 1$ denotes a uniformly distributed random letter, correlations are found to be present right from the beginning. Fig. 3.9 (a)

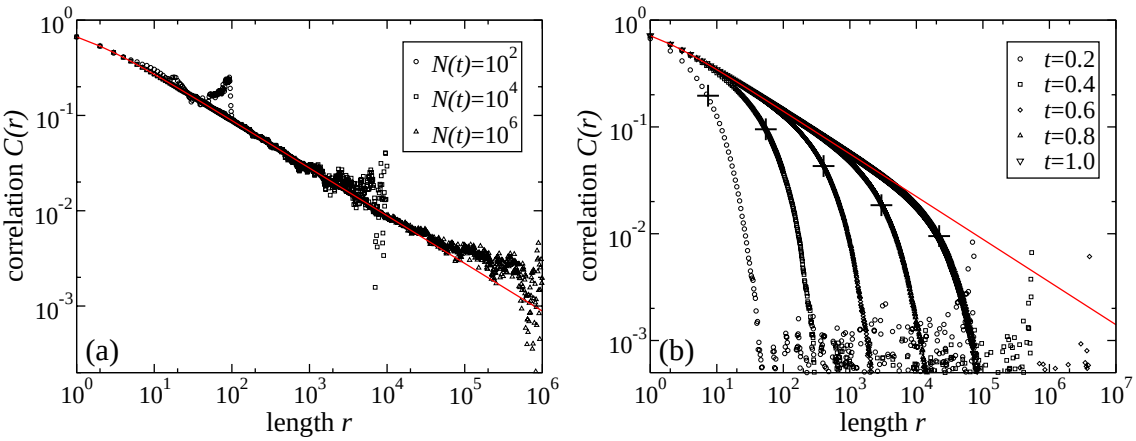

Figure 3.9: Time-dependent correlations $C(r,t)$. (a) Build-up of long-range correlations by stationary growth. Measured $C(r,t)$ at various intermediate lengths $N(t) = 10^2, 10^4, 10^6$ (symbols) together with the stationary form (3.23) for $\mu = 1.0$, $\delta_1 = 8.0$ (line). (b) Correlation build-up from a random sequence of length $N_0 = 10^4$. At $t = 0$ the processes started acting on the sequence with rates $\mu = 1.0$, $\delta_1 = 10.0$. $C(r,t)$ (symbols) was measured in simulated sequences after various times t (averages over 100 realizations). Black crosses denote the corresponding analytic cutoff sizes $r^*(t) = \exp(\lambda t)$. Correlations have been established in the sequences according to their analytic stationary form (red line) in the regime $r < r^*(t)$, whereas they vanish for $r > r^*(t)$.

gives examples for $C(r)$ measured along short single sequence realizations of length $N(t) = 10^2$, 10^4, and 10^6.

However, correlations cannot be present on all scales right away if we use a sequence $S(t = 0) = (s_1, \ldots, s_{N_0})$ with length $N_0 > 1$ as initial condition, whose letters are randomly chosen (and thus uncorrelated). All the processes of our model are local processes. A single step can introduce correlations only up to a microscopic length-scale $\ell_{\max}$. There will be a cutoff-length $r^*(t)$ up to which correlations can have been established at time $t > 0$. It is determined by the average distance, two copies of a duplication event at $t = 0$ are separated from each other along the sequence at time t,

$$r^*(t) = \exp(\lambda t). \tag{3.54}$$

Fig 3.9 (b) shows that $r^*(t)$ marks the range where $C(r)$ will start to deviate significantly from its stationary form.

Distinct dynamical regimes and correlation decay There is ample evidence that the rates of local evolutionary processes are not constant in time [12]. We mimic this non-stationarity of the individual process rates by the succession of several distinct dynamical phases. For each individual phase n, the rates of the elementary processes are constant during the time interval $t_{n-1} < t < t_n$ and result in specific values of $\lambda^{(n)}$ and $\mu_{\text{eff}}^{(n)}$ for that particular phase. Between different phases, however,

the complete set of rates may change,

$$\begin{array}{lll} \text{phase 1:} & (\mu^{(1)},\, \delta_1^{(1)},\, \cdots) & \text{for} \quad t_0 < t < t_1 \\ \text{phase 2:} & (\mu^{(2)},\, \delta_1^{(2)},\, \cdots) & \text{for} \quad t_1 < t < t_2 \\ \vdots & \vdots & \vdots \\ \text{phase } n\text{:} & (\mu^{(n)},\, \delta_1^{(n)},\, \cdots) & \text{for} \quad t_{n-1} < t < t_n \\ \vdots & \vdots & \vdots \end{array} \tag{3.55}$$

Using the findings derived above, we can generalize our dynamics with respect to varying rates during sequence evolution. We start with the following simple two-stage scenario: sequence growth with rate $\lambda^{(1)} > 0$ for $0 < t < t_1$, followed by a second phase with $\lambda^{(2)} = 0$ and therefore $\langle N \rangle(t) = N^{(1)}$ for $t > t_1$. It is obvious from Eq. (3.26) that stationary long-range correlations only emerge as long as the sequence grows, i.e. for $\lambda^{(n)} > 0$. The time-dependent solution of Eq. (3.26) for the asymptotics of $C(r)$ during the second phase ($t > t_1$) then takes the form

$$C(r,t) = C(r,t_1)\, e^{-4\mu_{\text{eff}}^{(2)} \Delta t} \propto r^{-4\mu_{\text{eff}}^{(1)}/\lambda^{(1)}} e^{-4\mu_{\text{eff}}^{(2)} \Delta t} \tag{3.56}$$

with $\Delta t = t - t_1$. The long-range tails of the correlations established during the first phase are preserved in the second phase, but their amplitude decays exponentially with a characteristic time scale $\tau = (4\mu_{\text{eff}}^{(2)})^{-1}$.

In the short range part, correlations may still be present depending on the particular set of process rates chosen to assure $\lambda^{(2)} = 0$. If, for example, all rates $\delta_\ell^{(2)}$, $\gamma_\ell^{+(2)}$, $\gamma_\ell^{-(2)}$ are zero in the second phase, the only process acting will be mutation which exponentially destroys correlations uniformly along the sequence, and thus the amplitude of $C(r)$ will decay according to Eq. (3.56) for all lengths r. The situation becomes more complex if $\lambda^{(2)} = 0$ is accomplished in the presence of duplications by a compensatory increase of the deletion rate. In this case, the duplication process will keep correlations present at short lengths because there is always a finite probability that a site s_k recently originated by a duplication of s_{k-1} (which again might be a duplication of s_{k-2}, and so on.) and was not yet affected by a mutation event. Numerical results for this type of two-phase dynamics are shown in Fig. 3.10 (a), verifying the exponential decay of the long-range tail, predicted by Eq. (3.56).

In a general evolutionary scenario, with several distinct dynamical phases and arbitrary values of $\lambda^{(n)}$ and $\mu_{\text{eff}}^{(n)}$ for each particular phase, the functional characteristics of the correlations in the generated sequences will be shaped by a combination of correlation build-up and decay, according to the mechanisms which have been revealed above. During phase n with $\lambda^{(n)} > 0$, correlations will be established with $\alpha^{(n)} = 4\mu_{\text{eff}}^{(n)}/\lambda^{(n)}$. They will approximately range over a length scale $r = 1, \ldots, r_{\max}$ with $r_{\max} = \exp(\lambda^{(n)} \Delta t_n)$. The correlations already present from the previous phases will be transported to larger sequence distances. If they ranged across an interval $r = 1, \ldots, N(t_{n-1})$ at the end of phase $n-1$, they will be shifted to the interval $r = N(t_{n-1}), \ldots, N(t_n)$ during phase n. The long-range tails, however, will still obey the same exponent corresponding to the effective rates of the original growth phase

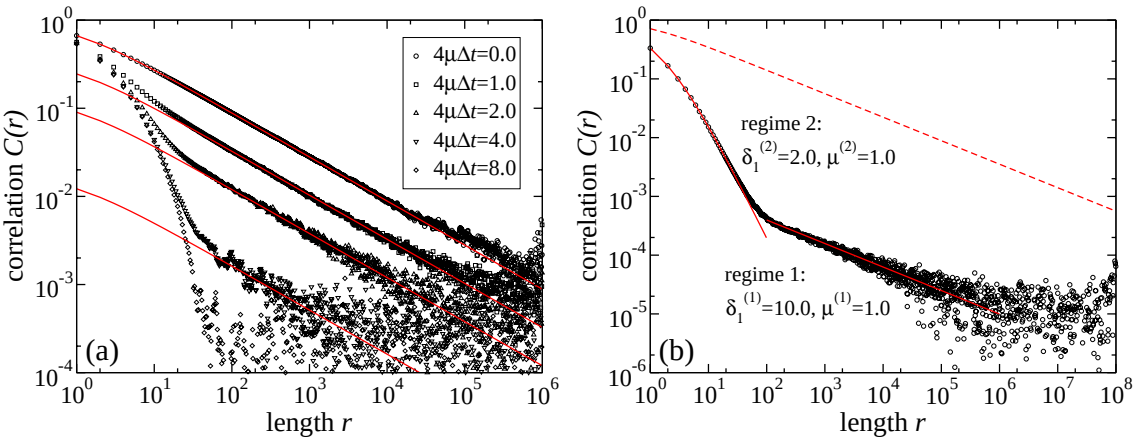

Figure 3.10: (a) Decay of correlations during sequence evolution at stationary length $N_0 = 10^6$. Measured $C(r,t)$ at various times Δt (symbols) together with the analytic decay of the long-range tail given by Eq. (3.56). In the previous growth phase for $t < t_0$, correlations have been established by a single-letter duplication-mutation dynamics with $\mu = 1.0$ and $\delta_1 = 8.0$ until the sequences reached the length $N_0 = 10^6$. For $\Delta t = t - t_1 > 0$, a single-letter deletion process with $\gamma_1^- = 8.0$ was introduced. Note that the correlations on short scales are preserved during the second phase. (b) $C(r)$ with two scaling regimes 1 and 2 (symbols). Process rates are: $\mu^{(1)} = 1.0$, $\delta_1^{(1)} = 10.0$ and $\mu^{(2)} = 1.0$, $\delta_1^{(2)} = 2.0$. The dashed red line is the analytical $C(r,t)$ for the parameters of phase 1. The second phase lasted over a period of time that on average allowed the sequences to increase their length by a factor of 100. For each scaling regime ($n = 1, 2$), $C(r)$ obeys the predicted algebraic decay with exponent $\alpha^{(n)} = 4\mu_{\text{eff}}^{(n)}/\lambda^{(n)}$. The transition between both regimes is sharp and its position agrees with the value predicted by Eq. (3.54).

they have originated from. Additionally, they are at the mercy of mutations. Their amplitude will therefore decay exponentially on all scales according to Eq. (3.56) with the effective mutation rate $\mu_{\text{eff}}^{(n)}$. A numerical example of a two-stage dynamics with two distinct scaling regimes is shown in Fig. 3.10 (b).

Given the chronology of the rates for all phases, we thus can in principle predict the different scaling regimes of $C(r)$. Furthermore, given the measured $C(r)$ of a sequence generated under the influence of our processes, we might be able to reconstruct the chronology of the ratio of the effective rates λ and μ_{eff} back throughout its evolutionary history. In practice, however, such an attempt will be confined by two major constraints: First, all of the above statements only apply to the long-range tails of $C(r)$. In order to perspicuously identify the decay exponent α of a certain rate regime, the net expansion during that regime must hence have been sufficiently large. Moreover, the ratio λ/μ_{eff} of the succeeding phases should be high, because correlations of the previous phases decay exponentially on a time-scale $\tau = (4\mu_{\text{eff}})^{-1}$. Otherwise previously established correlations will rapidly decay below the fluctuation threshold $\Delta C = 1/\sqrt{N(t)}$, and thus cannot be measured any longer.

3.7 General four-letter model and web service CorGen

The general four-letter model Our sequence evolution model defined in (3.1) operates on sequences with letters taken from a binary alphabet $s_k = \pm 1$. This simplification was originally motivated by the presumption of complementary strand symmetry, which has been shown to hold in genomes for first and also higher-order symmetries if investigated length scales are large enough [16]. Initially, we adopted an even stronger assumption by postulating Z_2 symmetry, meaning that the rates of all processes in our model are independent of s_k. However, breaking this symmetry does not change the general scaling features of the two-letter model as has been shown in Section 3.5. When it comes to a quantitative application of our results to genomic sequences, we have to generalize our findings to a full four-letter model with $s_k \in \{A, C, G, T\}$. Furthermore, we need to allow for arbitrary 4×4 mutation rate matrices. In the following, we will demonstrate how this can be accomplished in a systematic manner and thereby prove the emergence of long-range correlations in a generalized four-letter model with tunable GC-content.

By generalizing Eq. (3.17) in a continuum limit, we obtain Master equations for the 16 two-point functions $P_{ij}(r)$, which measure the joint probabilities of finding nucleotides i and j at a distance of r base pairs along the genome

$$\frac{\partial}{\partial t} P_{ij}(r) = \sum_a \left[\mu_{a \to i} P_{aj}(r) + \mu_{a \to j} P_{ia}(r) \right] - \lambda r \frac{\partial}{\partial r} P_{ij}(r). \tag{3.57}$$

To reveal the basic principles of our approach we will start with a simplified HKY-type [64] mutation matrix that has a tunable stationary GC-content g. A generalization to arbitrary 4×4 rate matrices is straightforward and will be discussed below. Here we define mutation rates $\mu_{i \to j}$ by the matrix

$$\mathbf{q} = \begin{pmatrix} \cdot & \mu_{T\to A} & \mu_{C\to A} & \mu_{G\to A} \\ \mu_{A\to T} & \cdot & \mu_{C\to T} & \mu_{G\to T} \\ \mu_{A\to C} & \mu_{T\to C} & \cdot & \mu_{G\to C} \\ \mu_{A\to G} & \mu_{T\to G} & \mu_{C\to G} & \cdot \end{pmatrix} = \frac{\mu}{2} \begin{pmatrix} \cdot & 1-g & 1-g & 1-g \\ 1-g & \cdot & 1-g & 1-g \\ g & g & \cdot & g \\ g & g & g & \cdot \end{pmatrix} \tag{3.58}$$

Elements on the diagonal are negative sums over columns, $\mathbf{q}_{ii} = -\sum_{j \neq i} \mathbf{q}_{ji}$. For our choice of $\mathbf{q}$, symmetry still holds for the rates $\mu_{A\to T} = \mu_{T\to A}$ and $\mu_{C\to G} = \mu_{G\to C}$, whereas rates $\mu_{A\to C}, \mu_{A\to G}, \mu_{T\to C}, \mu_{T\to G}$, are different from those of the backward processes. To solve Eq. (3.57) in the stationary state, we use a power-law Ansatz:

$$P_{ij}(r) = p_{ij}\, r^{-\alpha} + \pi_i \pi_j, \quad \text{where} \quad \pi_x = \begin{cases} (1-g)/2 & x = A, T \\ g/2 & x = C, G \end{cases} \tag{3.59}$$

denote stationary single nucleotide frequencies with respect to the rate matrix $\mathbf{q}$. Products $\pi_i \pi_j$ are hence joint-probabilities for the uncorrelated case, and we expect

$p_{i\neq j} < 0$ and $p_{ii} > 0$ in accordance with the two-letter model. For convenience, we can combine all $P_{ij}(r)$ in a 16-dimensional vector of the form

$$\begin{aligned}\vec{P}(r) &= (P_{AA}, P_{AT}, P_{AC}, P_{AG}, P_{TA}, P_{TT}, \cdots, P_{GC}, P_{GG})^\top \\ &= \vec{p}\, r^{-\alpha} + \vec{P}_0. \end{aligned} \tag{3.60}$$

Due to extensive structural symmetries in Eq. (3.57) and our choice of $\mathbf{q}$ there are various dependencies between the components of $\vec{P}(r)$. In particular, we have

$$\begin{aligned} &P_{AA} = P_{TT}, \quad P_{CC} = P_{GG}, \quad P_{AT} = P_{TA}, \quad P_{CG} = P_{GC} \\ &P_{AC} = P_{AG} = P_{TC} = P_{TG} = P_{CA} = P_{CT} = P_{GA} = P_{GT}. \end{aligned} \tag{3.61}$$

Furthermore, it has to hold that $\sum_j P_{ij} = \sum_j P_{ji} = \pi_i$ for all i, which also assures that $\vec{P}(r)$ is a valid probability vector with $\sum_{ij} P_{ij} = 1$. This leaves 3 independent parameters for $\vec{p}$. To solve the system of differential equations (3.57) with power-law Ansatz (3.60), we define a 16×16 matrix $\mathbf{Q}$ by

$$\mathbf{Q} = \mathbf{I}_4 \otimes \mathbf{q} + \mathbf{q} \otimes \mathbf{I}_4, \tag{3.62}$$

where $\mathbf{I}_4$ is the 4×4 identity matrix and $\otimes$ denote Kronecker products. All 16 Master equations (3.57) can then be combined in a matrix equation:

$$\begin{aligned}\frac{\partial}{\partial t}\vec{P}(r) &= \mathbf{Q}\vec{P}(r) - \lambda r \frac{\partial}{\partial r}\vec{P}(r) \\ &= \mathbf{Q}\, \vec{p}\, r^{-\alpha} + \lambda\alpha\, \vec{p}\, r^{-\alpha}. \end{aligned} \tag{3.63}$$

The last identity holds due to $\mathbf{Q}\vec{P}_0 = \vec{0}$. In the stationary state, we have $\partial\vec{P}(r)/\partial t = 0$ and obtain the Eigenvalue equation

$$\mathbf{Q}\, \vec{p} = -\alpha\lambda\, \vec{p} \tag{3.64}$$

Spectral decomposition of $\mathbf{Q}$ yields eigenvalues $(-2\mu, -\mu, 0)$ and three corresponding sets of eigenvectors spanning the orthogonal subspaces $\mathcal{M}_{-2\mu}$, $\mathcal{M}_{-\mu}$, and $\mathcal{M}_0$. However, for admissible vectors $\vec{p} = r^\alpha[\vec{P}(r) - \vec{P}_0]$ meaning that $\vec{P}(r)$ fulfills all conditions in Eq. (3.61) and is also a valid probability vector, we have always $\vec{p} \in \mathcal{M}_{-2\mu}$. Hence, all feasible solutions of Eq. (3.64) have Eigenvalue -2μ, resulting in $\alpha = 2\mu/\lambda$. If we use definition (1.5) of the general two-point correlation function for nucleotide sequences, i. e. $C(r) = \sum_i [P_{ii}(r) - \pi_i^2]$, we again obtain algebraically decaying long-range correlations in the asymptotic regime,

$$C(r) = r^{-\alpha}\left(\sum_i p_{ii}\right) \quad \text{with} \quad \alpha = \frac{2\mu}{\lambda}. \tag{3.65}$$

As expected from universality, long-range correlations are generic in the four-letter model too. The characteristic decay-exponent α does not depend on the GC-content g.

Notice however that α in the four-letter model differs from the two-letter case by a factor of two. The reason for this disparity becomes obvious when reducing our four-letter rate matrix (3.58) to the binary case by grouping together C/G to $+1$, and A/T to -1. Transitions $+1 \to -1$ then occur at rate $\mu(1-g)$ in the four-letter model. For the symmetric case $g = 1/2$, we thus have $\mu_{+\to-} = \mu_{-\to+} = \mu/2$.

Our approach can be extended to more general rate matrices in a straightforward manner. The only requirement we demand from the mutational dynamics is stationarity, which is crucial for Eq. (3.64). In its most general form, stationarity is fulfilled if $\mathbf{q}(\pi_A, \pi_T, \pi_C, \pi_G)^\top = 0$ holds for the mutation matrix $\mathbf{q}$ and the corresponding single nucleotide distribution. From Eq. (3.57) one can then construct a 16×16 matrix $\mathbf{Q}$ according to (3.62) with its elements constituted by simple functions of the elements of $\mathbf{q}$. Solving the resulting eigenvalue equation (3.64) and thereby taking into account all constraints resulting from symmetries in the chosen rate matrix $\mathbf{q}$ and the probability nature of $\vec{P}(r)$, will yield α and the solution space for $\vec{p}$.

The web server CorGen In computational biology one often requires an appropriate null model of DNA sequences, reflecting our assumptions about the "background" statistical features of the sequence under consideration. The need for a realistic null model arises from the fact that the statistical significance of a computational prediction derived by bioinformatics methods is often characterized by a p-value, which specifies the likelihood that the prediction could have arisen by chance. Popular null models are random sequences with letters drawn independently from an identical distribution, or kth order Markov models specifying the transition probabilities $P(s_{i+1}|s_{i-k+1}, \cdots, s_i)$ in a genomic sequence [48]. However, both models are incapable of incorporating long-range correlations in the sequence composition.

The widespread presence of long-range correlations in eukaryotic genomes raises the question whether they should be incorporated in a realistic null model of DNA. For example, we will show in Chapter 4 that such correlations substantially change the p-values of sequence alignment scores if the the standard iid model is replaced by a null model with long-range correlated sequences. To establish a realistic null model that incorporates long-range correlations in the sequences we need to specify its precise correlation parameters (amplitude and decay exponent α) reflecting the values measured in the genomic sequence. Often one also wants to have an ensemble of random sequence realizations from the null model, for example in those cases, where p-values can only be calculated by numerical simulations. For these purposes we developed the web server CorGen [109], which can measure long-range correlations in DNA sequences and generate random sequences with the same (or user-specified) correlation and composition parameters. CorGen is publicly available at `http://corgen.molgen.mpg.de`.

The generation of random DNA sequences with long-range correlated nucleotide composition has long been regarded as quite intricate. However, we have shown in this chapter that such sequences can efficiently be generated by simple dynamical models of sequence evolution including tandem duplication and mutation processes. In contrast to previously proposed methods to produce long-range correlated

sequences [103, 164, 39], the duplication-mutation model combines the following advantages: (a) exact analytic results for the correlation function of the generated sequences have been derived; (b) the method allows to generate sequences with any user-defined value of the decay exponent $\alpha > 0$, desired GC-content $0 < g < 1$, and length N; (c) the correlation amplitude is high enough to keep up with strong genomic correlations and can easily be reduced; (d) the dynamics can be implemented by a simple algorithm with runtime $O(N)$; (e) duplication and mutation processes are well known processes of molecular evolution.

In CorGen, we use a simple single-site duplication-mutation algorithm to generate long-range correlated sequences. We start with a sequence of one random nucleotide. The dynamics is defined by a single-site duplication process occurring at rate $\gamma_1^+ = 1.0$, and mutations specified by a rate matrix (3.58) with tunable GC-content g and rate-parameter μ. No other processes are acting on the sequence. As derived in Eq. (3.23), this dynamics generates sequences with correlations

$$C(r) = \frac{3}{4}\alpha B(r+1, \alpha) \tag{3.66}$$

The additional factor 3/4 compared to Eq. (3.23) results from extending the two-letter model of Section 3.3 to a four-letter model here. Asymptotically we have long-range correlations $C(r) \propto r^{-\alpha}$ for large r. The decay exponent is determined by $\alpha = 2\mu$, as derived in Eq. (3.65) using $\lambda = \gamma_1^+ = 1$. By varying the mutation parameter μ we can hence tune α to any desired positive value.

The correlations $C(r)$ of the generated sequences define the maximal amplitude obtainable by our dynamics for the specific settings of α and g. However, for the generation of long-range correlated sequences with correlation parameters comparable to those measured in genomic sequences, the correlation amplitude typically has to be reduced to the particular genomic amplitude. This can be accomplished according to the results derived in Section 3.6 by a simple procedure. After the sequence has reached its desired length, the duplication process is stopped. Subsequent mutation of M randomly drawn sites with mutation probabilities

$$\text{Prob}(X \to Y) = \begin{cases} (1-g)/2 & Y = A, T \\ g/2 & Y = C, G \end{cases} \tag{3.67}$$

will uniformly decrease the correlation amplitude to $C^*(r) = C(r)\exp(-2M/N)$ without changing the exponent α and the GC-content g.

CorGen offers three different types of services: (a) measuring long-range correlations of a given DNA sequence, (b) generating long-range correlated random sequences with the same statistical parameters as the query sequence, and (c) generating sequences with specific user-defined long-range correlations. The first two tasks require the user to upload a query DNA sequence in FASTA or EMBL format. For long-range correlations to be detectable, the sequences need to be sufficiently long (we recommend at least 1000 bp). The distance interval where a power-law is fitted to the measured $C(r)$ can be specified by the user.

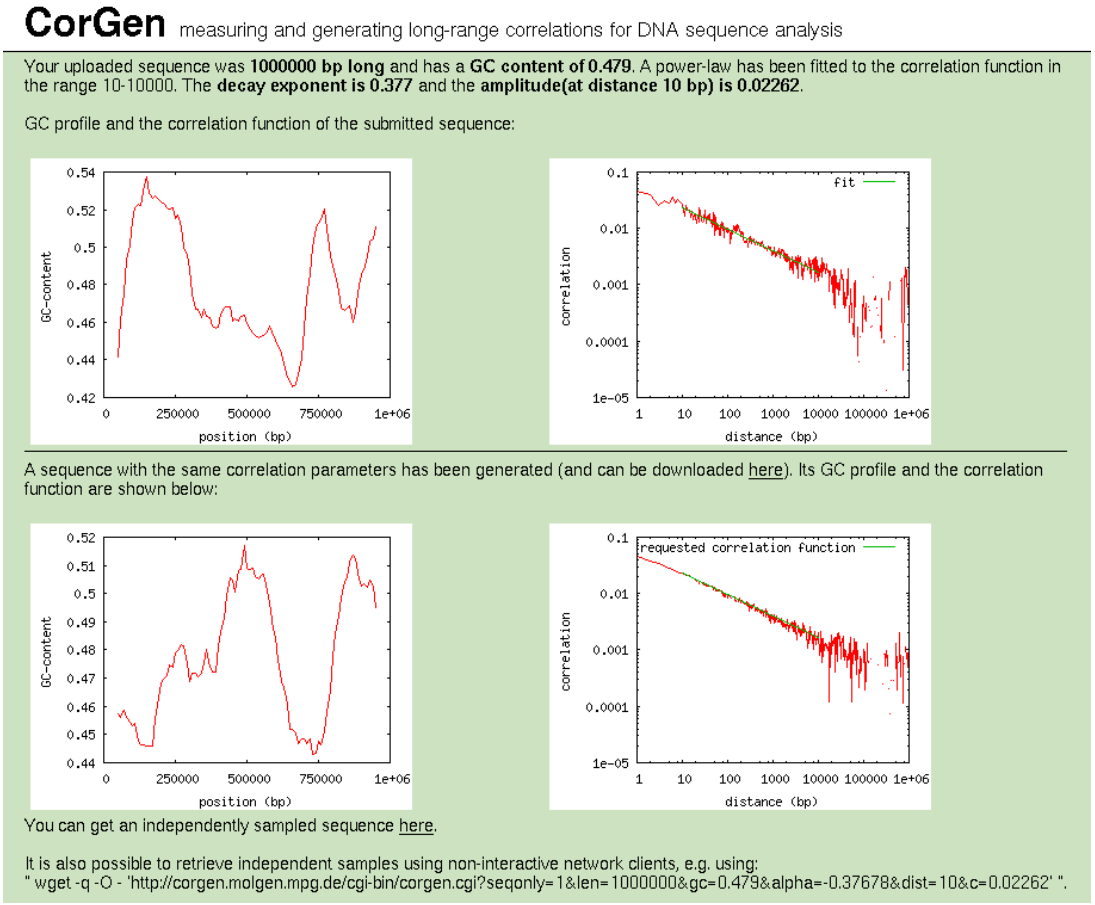

Figure 3.11: CorGen analysis of a 1 Mbp region of human chromosome 1 (position: 25 Mbp - 26 Mbp). The two plots in the top part show the measured GC-profile (left) and correlation function (right) of the chromosomal region. The fitting to $C(r) \propto r^{-\alpha}$ has been performed in the range $10 < r < 10000$, and the obtained parameters are $\alpha = 0.377$ and $C(10) = 0.0226$ (green line). A corresponding random sequence of length 1 Mbp with the measured long-range correlation parameters and average GC-content of the query sequence has been generated and can be downloaded by the user. Its composition profile and correlation function are shown in the two plots at the bottom.

Upon submission of a query DNA sequence, CorGen will generate plots with the measured GC-profile and correlation function, as defined by Eq. (1.5). Unsequenced or ambiguous sites are thereby excluded from the analysis. The user can specify a distance interval where a power-law should be fitted to the measured correlation function. The obtained values for the decay exponent α and the correlation amplitude will be reported by CorGen. If a long-range correlated random sequence with the same statistical features in the specified fitting interval has been requested, it will be generated and its corresponding composition and correlation plots will also be shown. See Fig. 3.11, for an example CorGen output page. The generated random sequences can be downloaded by the user. If large sequence ensembles are needed, independent realizations of the sequences can directly be obtained via non-interactive network clients, e. g. `wget`. CorGen can also be used to generate long-range correlated

random sequences with specific user-defined correlation parameters. In this case, the user needs to specify the decay exponent α, the correlation amplitude $C(r^*)$ at a reference distance r^*, the desired GC-content g, and the sequence length N. Notice that there is a generic limit for the correlation amplitude depending on the values of α and g. As a typical example, the measurement of $C(r)$ for human chr. 22 takes ~ 65 seconds. A random sequence of length 1 Mbp with the same correlation parameters can be generated in less than 5 seconds.

3.8 Origin of genomic correlations

Long-range correlations in base composition characterized by an asymptotic power-law decay $C(r) \propto r^{-\alpha}$ of the autocorrelation function are a widespread feature among the genomes of most eukaryotic genomes [131, 94, 163, 13, 23, 92, 93]. In this chapter, we have analytically shown that long-range correlations generically emerge in sequences that evolve under the influence of nucleotide mutation and tandem duplication processes. These processes are also regarded as the major local processes acting on genomic DNA sequences during evolutionary history. We have demonstrated in Chapter 2 that the vast majority of recent short DNA insertions in the human genome indeed resulted from tandem duplications of existing adjacent sequence segments, and supposably this predominance is universal throughout large parts of eukaryotic evolution.

Our findings raise the question whether – and if so to what extent – there is a causal connection between the elementary mutational processes of molecular evolution and the observed long-range correlations in genomic base composition? The theory developed in the previous sections provides a promising approach to this question by quantitatively linking the decay exponent α, which is the determining property of long-range correlations, to only two effective parameters of the evolutionary dynamics: effective growth rate λ and effective mutation rate μ_{eff}. With a comprehensive record of these two parameters for a particular genomic region and over a sufficiently long evolutionary period at hand, we could in principle check compatibility with the observed correlations in this region. However, such an approach is hardly feasible considering the extremely long time-scales we would have to take into account. According to our model, present-day correlations $C(r)$ at distance $r = 10^6$ bp, for example, originated from correlations present at distance $r = 10^4$ bp when the investigated region was 100 times shorter compared with today. It is also very unlikely that such extensive expansions have occurred with constant growth rate in time and along the genome. Rather there might have been "bursts" of rapid expansion, e. g. by acquisitions of new classes of transposable elements, followed by long periods of approximately constant sequence length. In a punctuated growth process, strong long-range correlations with small exponents α are produced and transported to larger distances during the rapid expansion phases. During the stationary phases, previously established correlations will decay uniformly on all scales as given by Eq. (3.56) without changing α.

Such punctuated growth scenarios also pose a possible solution to the key problem of exponential sequence growth inherent to our sequence evolution model (3.1) if $\lambda > 0$. As has already been pointed out in [111], evolutionary scenarios with constant growth rate in time and along the genome can clearly be rejected assuming values of λ compatible with those predicted from the observed long-range correlations according to our sequence evolution model. This becomes obvious by the following simple estimation for the human lineage. The correlation function $C(r)$ along human chromosomes shows a rather slow algebraic decay on "mesoscopic" distance scales $10^2 < r < 10^5$ with typical effective exponents $\alpha \approx 0.5$ [23, 66, 93] (see also Fig. 3.12). A lower bound of the effective mutation rate in mammals is $\mu \approx 2 \cdot 10^{-9}\mathrm{a}^{-1}$ per site [10]. Assuming stationary growth, we can use these values of α and μ to derive a lower bound on the genomic growth rate λ, resulting in a minimum value $\lambda \approx 10^{-8}\mathrm{a}^{-1}$ per site according to Eq. (3.27). This rate is much too high. The human genome contains $N \approx 3 \cdot 10^9$ base pairs and – assuming the above rate of genome expansion – would have contained only about $4 \cdot 10^6$ base pairs at the time of mammalian radiation about 90 million years ago. This is obviously incommensurate with the fact that approximately 40% of the human genome can still be aligned to the mouse genome, representing most of the orthologous sequences that remain in both lineages from the common ancestor [169]. However, if we assume a punctuated growth process, this discrepancy can be resolved. In mammals the last likely period of rapid expansion has been the mammalian radiation, and the characteristic time scale of correlation decay is $\tau \approx 100$ Myr according to Eq. (3.56). Correlations present or generated at the time of the mammalian radiation would hence still persist today. 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 [23]. Thus, a punctuated expansion-randomization process may be compatible with the correlations observed in mammals.

Compared to variations in time, spatial fluctuations of λ along different regions of the genome are presumably even more pronounced [100, 116, 86, 70]. The effect of such regional fluctuations in λ on long-range correlation characteristics essentially depends on the spatial scale these fluctuations occur on. As an immediate consequence of universality, process rates on microscopic length-scales will enter the composition correlations in the mesoscopic range only via the average growth rate and effective mutation rate. Variations in λ on small scales can hence be straightened out by mesoscopic averages. If, on the the other hand, growth rates averaged over mesoscopic windows differ considerably between such windows, the correlations estimated in individual windows will also vary between windows. As long as $\lambda > 0$ holds throughout a long enough evolutionary history of a given genomic segment, long-range correlations will have been established in that segment with decay exponent $\alpha = 4\mu_{\mathrm{eff}}/\lambda$ according to the particular effective rates in the segment. In segments with $\lambda < 0$, long-range correlations cannot emerge and the amplitude of previously established correlations will decay uniformly on all scales. Hence, different segments may exhibit different α, which is exactly what we observe in genomic sequences. In Fig. 3.12 for example, we show $C(r)$ measured separately for 10 Mbp long non-overlapping

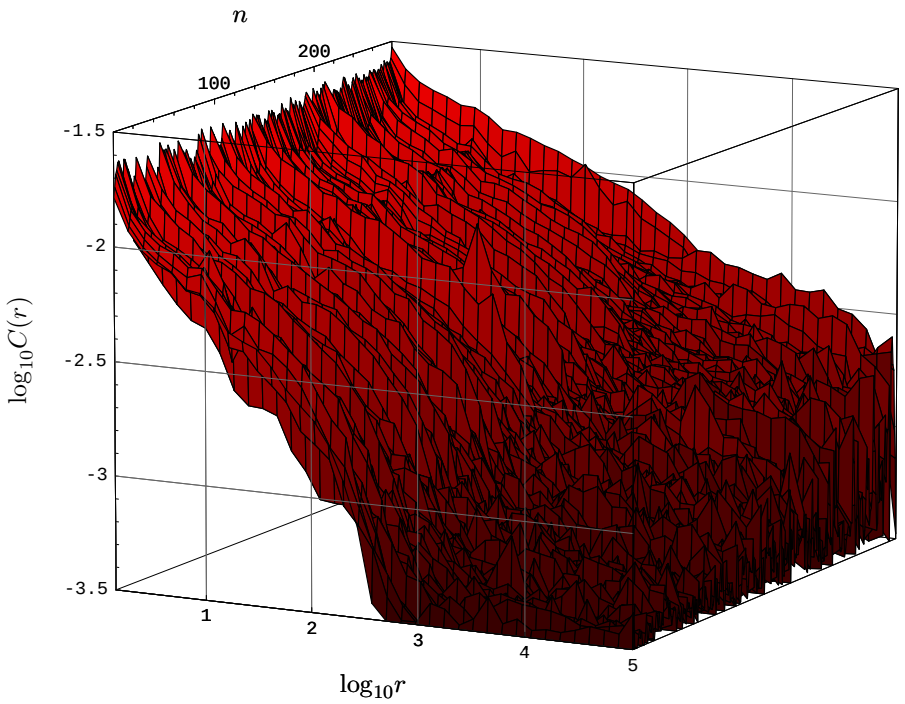

Figure 3.12: Regional variation in correlation characteristics along the human genome. Correlation functions $C(r)$ were estimated according to definition (1.5) for 280 non-overlapping windows of length 10 Mbp covering most of the human genome. They are ordered along the y-axis according to the averaged correlation strength in the interval $10^3 < r < 10^4$. Our regional analysis reveals clear differences in the composition correlations between distinct genomic regions. While the emergence of power-law correlations appears to be universal, the decay exponent is not. For large values of α, correlations rapidly decay below the fluctuation and long-range tails cannot be measured any longer.

windows of the human genome. Decay exponents in the mesoscopic regime vary strongly between strong long-range correlations with $\alpha \approx 0.1$, and windows with no measurable long-range correlations on mesoscopic scales.

Regional variations in λ on large spatial scales also pose a possible solution to the problem of exponential sequence growth because high growth rates in expanding regions can be counterbalanced by appropriate negative growth rates in other regions. We numerically demonstrate that by this mechanism long-range correlations can indeed be generated in sequences of constant overall length in Fig. 3.13. A particular class of evolutionary processes likely to play a crucial role in this context are genomic rearrangements. Rearrangements of genomic segments are effectively neutral regarding the overall growth rate of the genome, but they can obviously have a substantial effect on local growth and deletion rates. For long-range correlations to be generated by this process, insertions of translocated segments in expanding regions additionally have to occur in a GC-biased fashion according to Eq. (3.37). Otherwise their contribution to the effective mutation rate will exceed the effect of an increased local growth rate, as has been discussed in Section 3.5.

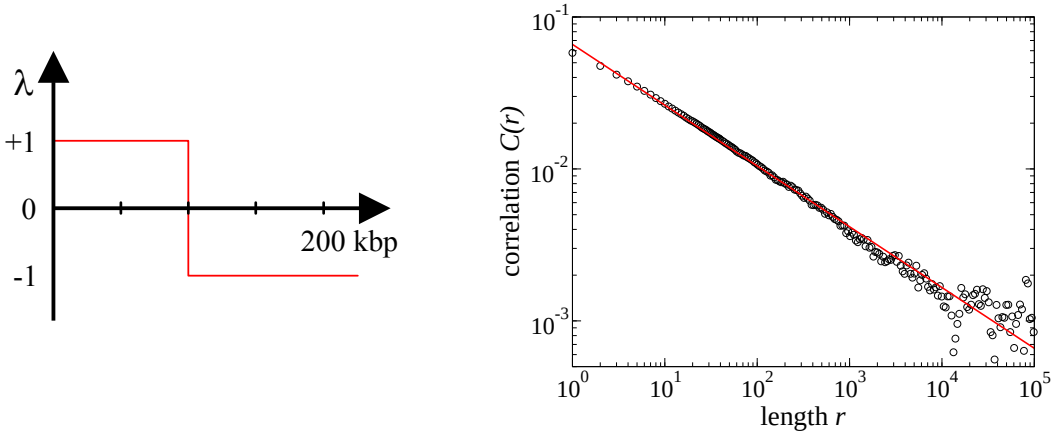

Figure 3.13: Emergence of long-range correlations by expansion-randomization processes in a simple two-regime scenario with constant average sequence length. Binary sequences of length 200 kbp initially consisting of independent random letters were evolved under a single-site duplication-deletion-mutation dynamics until correlations reached stationarity. The rates of duplication and deletion were thereby chosen according to the spatial profile defined in the left plot. In the left sequence regime for sequence positions $0 < x < 100$ kbp, $\lambda = 1$ was achieved by single-site duplications with rate $\delta_1 = 1.0$ and no deletions. In the right regime for $x > 100$ kbp, duplications were switched off and single-site deletions occurred at rate $\gamma_1^- = 1.0$. Mutations acted in both regimes at constant rate $\mu = 0.1$. In this scenario, the overall sequence length is approximately constant and fluctuates around 200 kbp. In the right plot we show the measured correlation function $C(r)$ estimated over the entire sequences and averaged over 100 runs. $C(r)$ features clear long-range correlations with theoretically predicted exponent $\alpha = 0.4$ (red line).

We conclude that the observed long-range correlations in eukaryotic genomes are in principle compatible with local expansion-randomization processes. The problem of overall exponential sequence growth can be resolved by assuming strong variations of the local growth rate λ on large genomic scales if elongation of rapidly expanding regions is compensated for by other contracting regions. Long-range correlations are then produced in the currently expanding regions of the genome whenever growth is driven by tandem duplication events or, more general, insertions that are biased towards the local GC-content. In order to generate correlations of genomic magnitude $\alpha \approx 0.5$ by an expansion-randomization dynamics, our theory implies that for some genomic regions the average growth rate has at least been of the same order of magnitude as the single nucleotide mutation rate over sufficiently long evolutionary periods. Clearly, further analysis of genomic data is needed to corroborate or refute possible causes of the observed correlations. Advanced comparative genomics approaches facilitated by the rapidly growing availability of whole-genome sequence data (e. g. the recently sequenced genomes of 12 Drosophila species [47]) will hopefully help us to elucidate the mutational dynamics of chromosomes on long evolutionary timescales in more detail. If genomic expansion proves to be a significant contribution, composition correlations could become the "background radiation" of genomics, allowing us to trace the expansion history of genomes far back in evolutionary time.