A Stochastic Model for Genomic Interspersed Duplication

Farzad Farnoud
Electrical Engineering
California Institute of Technology
Pasadena, CA 91125, USA
Email: farnoud@caltech.edu

Moshe Schwartz
Electrical and Computer Engineering
Ben-Gurion University of the Negev
Beer Sheva 8410501, Israel
Email: schwartz@ee.bgu.ac.il

Jehoshua Bruck
Electrical Engineering
California Institute of Technology
Pasadena, CA 91125, USA
Email: bruck@paradise.caltech.edu

***Abstract*—Mutation processes such as point mutation, insertion, deletion, and duplication (including tandem and interspersed duplication) have an important role in evolution, as they lead to genomic diversity, and thus to phenotypic variation. In this work, we study the expressive power of interspersed duplication, i.e., its ability to generate diversity, via a simple but fundamental stochastic model, where the length and the location of the subsequence that is duplicated and the point of insertion of the copy are chosen randomly. In contrast to combinatorial models, where the goal is to determine the set of possible outcomes regardless of their likelihood, in stochastic systems, we investigate the properties of the set of high-probability sequences. In particular we provide results regarding the asymptotic behavior of frequencies of symbols and short words in a sequence evolving through interspersed duplication. The study of such a systems is an important step towards the design and analysis of more realistic and sophisticated models of genomic mutation processes.**

I. Introduction

It is estimated that there are about 8.7 million species on earth [5]. Of course, individuals within each species are also different from each other. Thus there is a vast amount of biological diversity on Earth. This diversity is, for the most part, the result of genomic mutation. The types of mutation include point mutation, insertion/deletion, and duplication. Duplication mutations, where a segment of DNA is copied and inserted elsewhere in the genome, may in turn be of the tandem or interspersed type. In tandem duplication, the copy is inserted immediately after the original, while in interspersed repeats, the copy may be inserted far from the original DNA segment. Interspersed duplications are caused by transposons, which are segments of DNA that can copy and insert themselves into new positions of the genome.

The general goal of this work is to move towards better understanding the effects of genomic interspersed duplication on generating novel sequences and creating biological diversity. Among the aforementioned mutation processes, interspersed duplication is of particular interest as it leads to interspersed repeated sequences, which form 45% of the human genome [4]. Here, we take a *probabilistic* approach to model interspersed duplications and investigate their ability to generate novelty and diversity. This complements our previous work [2], [3], in which we considered the same problem from a *combinatorial* point of view, with the goal of studying the set of *possible* sequences arising from duplication systems. In contrast, the probabilistic view is concerned with identifying the *probable* outcomes of stochastic duplication systems. Particularly, we seek to find certain properties which the outcome of an interspersed-duplication system will possess with high probability.

In our interspersed-duplication model, a string evolves through random interspersed-duplication events, i.e., in each step, a random segment of the string is duplicated and then inserted in a random position in the string, independent of the position of the original segment. To avoid complications arising from boundary cases, we consider *circular strings*. It is worth noting that in fact many bacteria have circular chromosomes. While in practice, different mutation processes work together to create novel sequences, the scope of this work is limited to analyzing interspersed duplications in isolation. This helps us to obtain a better understanding of the properties of this type of mutation. We leave the study of more complex systems that evolve through more than one mutation type to future work.

Our analysis starts by considering how the frequencies (multiplicities divided by the length of the evolving string) of the alphabet symbols change as duplications occur. We show that under general conditions, the frequencies are martingales and thus converge almost surely. The same argument does not apply to the frequencies of strings of length larger than one. To analyze such frequencies, we use the stochastic approximation method which enables modeling of a discrete dynamic system by a corresponding continuous model described by ordinary differential equations. We show then that for interspersed-duplication systems, the frequencies of strings of length larger than one are, in the limit, consistent with those of iid sequences; implying that in a certain sense, a sequence evolving through interspersed duplication is unrecognizable from an iid sequence. Note that an iid sequence has the maximum entropy among sequences with a given symbol distribution.

The rest of the paper is organized as follows. Notation and preliminaries are given in the next section. Section III contains the analysis of the evolution of symbol frequencies in the evolving string. In Section IV, we present the necessary background and preliminaries for the use of stochastic approximation in this work. Sections V is devoted to the analysis of

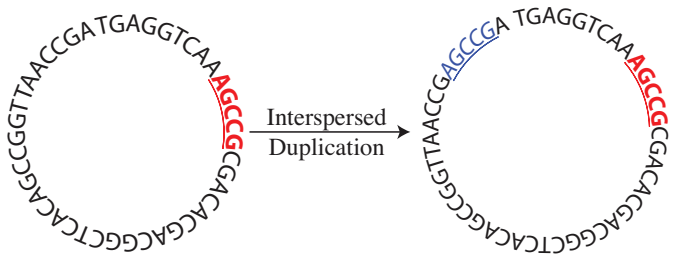

Figure 1. Interspersed duplication in a circular string, where the **original segment** and *copy* are shown in different styles.

substring frequencies in interspersed-duplication systems. We close the paper with concluding remarks in Section VI.

II. Preliminaries and Notation

Let s be a circular string over some alphabet $\mathcal{A}$, that evolves over time through interspersed duplication. In our examples, we typically let $\mathcal{A} = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$. In each step of the process, first a length ℓ is chosen according to a fixed and bounded distribution $\boldsymbol{q} = (q_i)_{i=1}^{K-1}$, where q_i denotes the probability of the event $\ell = i$. Then an ℓ-substring is chosen randomly and uniformly among all $|s|$ substrings of s that have length ℓ (recall that s is viewed as a circular string and thus contains $|s|$ such substrings). Then a copy of s' is inserted at a random position in s. For simplicity of notation, the dependence of s on n is implicit. If there is possibility for confusion, we denote the string s at time n with $s^{(n)}$. An example of this process is given in Figure 1. As another example, suppose $s^{(n)} = \underline{\mathsf{A}}\mathsf{GTT}\underline{\mathsf{CAC}}$, $\ell = 3$, and $s' = \mathsf{ACA}$, where s' is underlined in s. Two possibilities for $s^{(n+1)}$, after a single interspersed duplication, are $\underline{\mathsf{A}}\mathsf{GTT}\overline{\mathsf{ACA}}\underline{\mathsf{CAC}}$ and $\underline{\mathsf{A}}\mathsf{GTT}\underline{\mathsf{CA}}\overline{\mathsf{ACA}}\underline{\mathsf{C}}$, where the inserted copy of s' is over-lined.

The length of s at time n (that is, after n duplications) is denoted by $|s^{(n)}| = L_n$, with L_0 being the length of the initial string. For a sequence u, we use $u_{i,j}$ to denote the length-j subsequence of u starting at u_i. Furthermore, the concatenation of two strings u and v is represented by uv. In this paper, vectors are denoted by boldface letters such as $\boldsymbol{x}_n$, and scalars by normal letters such as x_n^u.

III. Evolution of Symbol Frequencies

In this section, we study the evolution of symbol frequencies as the evolving string s undergoes interspersed duplications. The results of this section are in fact valid not only for interspersed duplication, but also for any duplication process in which for each i, all i-substrings of s have the same chance of being duplicated.

For a symbol u of $\mathcal{A}$, let the number of appearances of u in s at time n be denoted by μ_n^u, and its frequency by x_n^u, where $x_n^u = \mu_n^u / L_n$. We let $\{\mathcal{F}_n\}$ be the filtration generated by the random variables $\{x_n^u, L_n\}$. Recall that we assume $q_0 = 0$.

Theorem 1. *The random variables x_n^u, where u is a symbol of the alphabet, are martingales and converge almost surely.*

Proof: Suppose $u \in \mathcal{A}$. We have

$$\mathbb{E}\big[x_{n+1}^u | \mathcal{F}_n\big] = \mathbb{E}\left[\frac{\mu_{n+1}^u}{L_{n+1}}\middle|\mathcal{F}_n\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\mu_{n+1}^u}{L_{n+1}}\middle|\mathcal{F}_n, \ell\right]\middle|\mathcal{F}_n\right]$$
$$= \mathbb{E}\left[\frac{\mu_n^u + \ell x_n^u}{L_n + \ell}\middle|\mathcal{F}_n\right] = x_n^u.$$

We thus have $E\big[x_{n+1}^u | \mathcal{F}_n\big] = x_n^u$ and so x_n^u is a Martingale. Since $\mu_n^u \le L_n$, we clearly have that $\mathbb{E}[|x_n^u|] \le 1$ and so by the martingale convergence theorem, x_n^u converges almost surely. ∎

The above theorem does not in fact require the distribution q to be bounded. If q is bounded, we can also obtain the following result on the probability of x_n^u deviating from its starting value.

Theorem 2. *If there exists a positive integer K such that $q_i = 0$ for $i \ge K$, then we have*

$$P(|x_n^u - x_0^u| \ge \lambda) \le 2e^{-\lambda^2 L_0^2/(2K^4)}$$

for all $u \in \mathcal{A}$ and $n \ge 1$.

Proof: Note that $\frac{\mu_{n-1}^u}{L_{n-1}+K} \le \frac{\mu_n^u}{L_n} \le \frac{\mu_{n-1}^u + K}{L_{n-1}+K}$. Thus

$$-\frac{K\mu_{n-1}^u}{L_{n-1}(L_{n-1}+K)} \le \frac{\mu_n^u}{L_n} - \frac{\mu_{n-1}^u}{L_{n-1}} \le \frac{K(L_{n-1}-\mu_{n-1}^u)}{L_{n-1}(L_{n-1}+K)},$$

implying that

$$\big|x_n^u - x_{n-1}^u\big| \le \frac{K\max\big\{L_{n-1}-\mu_{n-1}^u, \mu_{n-1}^u\big\}}{L_{n-1}(L_{n-1}+K)}$$
$$\le \frac{K}{L_{n-1}+K} \le \frac{K}{L_0+n-1+K} \le \frac{K}{L_0+n}.$$

Let $c_n = \frac{K}{L_0+n}$ so that $\big|x_n^u - x_{n-1}^u\big| \le c_n$ and note that

$$\sum_{i=1}^n c_i^2 = K^2 \sum_{i=1}^n \frac{1}{(L_0+i)^2} \le K^2 \int_0^n \frac{dt}{(L_0+t)^2}$$
$$= \frac{K^2}{L_0} - \frac{K^2}{L_0+n} = \frac{K^2 n}{L_0(L_0+n)} \le \frac{K^2}{L_0}.$$

By the Hoeffding-Azuma inequality, we have $P(|x_n^u - x_0^u| \ge \lambda) \le 2\exp\left(\frac{-\lambda^2 L_0^2}{2K^4}\right)$. ∎

The preceding theorem implies that it is unlikely for the composition of a long DNA sequence to change dramatically through uniformly random duplication events. Such changes, if observed, are likely the result of context-dependent duplications or other biased mutations.

Unfortunately, this simple martingale argument does not extend to more complex cases, for example when u is a substring of length more than 1. Therefore, for analyzing such cases, we use the more flexible technique of stochastic approximation as described in the sequel.

An illustration of the change in frequencies of symbols and short strings versus the number of duplications is given in Figure 2. It can be observed that the frequencies of A, C, G, and T vary less as the number n of duplications increases. In particular for large values of n, they become almost constant. The frequencies of strings of length larger than 1 is discussed in the following sections.

IV. Stochastic Approximation for Duplication Systems

In this section, we present a brief overview of the stochastic approximation method adapted to duplication systems. For an ordered set U, let $\boldsymbol{\mu}_n = (\mu_n^u)_{u\in U}$ be a vector representing the number of appearances of objects $u \in U$ in the string s at time n and let $\boldsymbol{x}_n = \frac{\boldsymbol{\mu}_n}{L_n}$ be the normalized version of $\boldsymbol{\mu}_n$. For example, U can be the set of all strings over $\mathcal{A}$ with length at most three. We also let $\{\mathcal{F}_n\}$ be the filtration generated by the random variables $\{\boldsymbol{x}_n, L_n\}$. Our goal is to find out how $\boldsymbol{x}_n$ changes with n by finding a differential equation whose solution approximates $\boldsymbol{x}_n$.

We state a set of conditions that must be satisfied for our analysis. Let $\mathbb{E}_\ell[\,\cdot\,]$ denote the expected value conditioned on the fact that the length of the duplicated substring is ℓ and let $\boldsymbol{\delta}_\ell = \mathbb{E}_\ell\big[\boldsymbol{\mu}_{n+1}|\mathcal{F}_n\big] - \boldsymbol{\mu}_n$. We consider the following conditions. Among them, we assume (A1) and, for now, accept the others without a proof.

(A1) There exists $K \in \mathbb{N}$ such that $q_i = 0$ for $i = 0$ or $i \geq K$.

(A2) $\boldsymbol{\mu}_{n+1} - \boldsymbol{\mu}_n$, and thus $\boldsymbol{\delta}_\ell$, are bounded.

(A3) $\boldsymbol{x}_n$ is bounded.

(A4) For each ℓ, $\boldsymbol{\delta}_\ell$ is a function of $\boldsymbol{x}_n$ only, so we can write $\boldsymbol{\delta}_\ell = \boldsymbol{\delta}_\ell(\boldsymbol{x}_n)$.

(A5) The function $\boldsymbol{\delta}_\ell(\boldsymbol{x}_n)$ is Lipschitz.

To understand how $\boldsymbol{x}_n$ varies, our starting point is its difference sequence $\boldsymbol{x}_{n+1} - \boldsymbol{x}_n$. We note that

$$\boldsymbol{x}_{n+1} - \boldsymbol{x}_n = \mathbb{E}[\boldsymbol{x}_{n+1} - \boldsymbol{x}_n|\mathcal{F}_n] + (\boldsymbol{x}_{n+1} - \mathbb{E}[\boldsymbol{x}_{n+1}|\mathcal{F}_n]). \tag{1}$$

For the first term of the right side of (1), we have[1]

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}_{n+1} - \boldsymbol{x}_n|\mathcal{F}_n] &= \sum_\ell q_\ell(\mathbb{E}_\ell[\boldsymbol{x}_{n+1}|\mathcal{F}_n] - \boldsymbol{x}_n) \\
&= \sum_\ell q_\ell\left(\frac{\boldsymbol{\mu}_n + \boldsymbol{\delta}_\ell(\boldsymbol{x}_n)}{L_n + \ell} - \frac{\boldsymbol{\mu}_n}{L_n}\right) \\
&= \frac{1}{L_n}\sum_\ell q_\ell \boldsymbol{h}_\ell(\boldsymbol{x}_n)\big(1 + O\big(L_n^{-1}\big)\big) \\
&= \frac{1}{L_n}\boldsymbol{h}(\boldsymbol{x}_n)\big(1 + O\big(L_n^{-1}\big)\big), \qquad (2)
\end{aligned}$$

where $\boldsymbol{h}_\ell(\boldsymbol{x}) = \boldsymbol{\delta}_\ell(\boldsymbol{x}) - \ell\boldsymbol{x}$, $\boldsymbol{h}(\boldsymbol{x}) = \sum_\ell q_\ell \boldsymbol{h}_\ell(\boldsymbol{x})$, and where we have used $1/(L_n + \ell) = \big(1 + O\big(L_n^{-1}\big)\big)/L_n$ which follows from the boundedness of ℓ (see (A1)).

Furthermore, for the second term of the right side of (1), we have

$$\begin{aligned}
\boldsymbol{x}_{n+1} - \mathbb{E}[\boldsymbol{x}_{n+1}|\mathcal{F}_t] &= \frac{\boldsymbol{\mu}_{n+1}}{L_{n+1}} - \mathbb{E}\left[\frac{\boldsymbol{\mu}_{n+1}}{L_{n+1}}|\mathcal{F}_n\right] \\
&= \frac{1 + O\big(L_n^{-1}\big)}{L_n}\big(\boldsymbol{\mu}_{n+1} - \mathbb{E}\big[\boldsymbol{\mu}_{n+1}|\mathcal{F}_n\big]\big) \\
&= \frac{1}{L_n}\big(1 + O\big(L_n^{-1}\big)\big)M_{n+1} \qquad (3)
\end{aligned}$$

[1]Here, we abuse the notation by using ℓ, which is a random variable, as an index of summations when the goal is finding expected values by conditioning on ℓ.

where $\boldsymbol{M}_{n+1} = \boldsymbol{\mu}_{n+1} - E\big[\boldsymbol{\mu}_{n+1}|\mathcal{F}_n\big]$. Note that $\boldsymbol{M}_n$ is a bounded martingale difference sequence.

From (1), (2), and (3), we find $\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \frac{1}{L_n}\big(\boldsymbol{h}(\boldsymbol{x}_n) + \boldsymbol{M}_{n+1} + O\big(L_n^{-1}\big)\big)$, where we have used the fact that $\boldsymbol{h}(\boldsymbol{x}_n)\big(1 + O\big(L_n^{-1}\big)\big) = \boldsymbol{h}(\boldsymbol{x}_n) + O\big(L_n^{-1}\big)$. This follows from the boundedness of $\boldsymbol{h}(\boldsymbol{x}_n)$, which in turn follows from the boundedness of $\boldsymbol{\delta}(\boldsymbol{x}_n)$.

The following theorem relates the discrete system describing $\boldsymbol{x}_n$ to a continuous system.

Theorem 3. *[1, Theorem 2] The sequence $\{\boldsymbol{x}_n\}$ converges almost surely to a compact connected internally chain transitive invariant set of the ode $d\boldsymbol{x}_t/dt = \boldsymbol{h}(\boldsymbol{x}_t)$.*

Note the dual use of the symbol $\boldsymbol{x}$ in the theorem; the meaning is however clear from the subscript. Recall that a set A is an *invariant* set of an ode $d\boldsymbol{z}_t/dt = \boldsymbol{f}(\boldsymbol{z}_t)$ if it is closed and $\boldsymbol{z}_{t'} \in A$ for some $t' \in \mathbb{R}$ implies that $\boldsymbol{z}_t \in A$ for all $t \in \mathbb{R}$. The invariant set A is *internally chain transitive* with respect to the ode $d\boldsymbol{z}_t/dt = \boldsymbol{f}(\boldsymbol{z}_t)$, provided that for every $\boldsymbol{y}, \boldsymbol{y}' \in A$ and positive reals T and ϵ, there exist $N \geq 1$ and a sequence $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_N$ with $\boldsymbol{y}_i \in A$, $\boldsymbol{y}_0 = \boldsymbol{y}$, and $\boldsymbol{y}_N = \boldsymbol{y}'$ such that for $0 \leq i < n$, if $\boldsymbol{z}_0 = \boldsymbol{y}_i$, then for some $t \geq T$, $\boldsymbol{z}_t$ is in the ϵ-neighborhood of $\boldsymbol{y}_{i+1}$.

V. Interspersed Duplication

Now, we use the technique presented in Section IV to extend the results of Section III by analyzing the frequencies of strings $u \in \mathcal{A}^*$ in an interspersed-duplication system.

Let $\ell'_{max} \in \mathbb{N}$ and let U be an ordered set consisting of all strings of length at most ℓ'_{max}. The vectors $\boldsymbol{x}_n$ and $\boldsymbol{\mu}_n$ are defined as before using U. Consider $u \in U$. To illustrate, we assume $\mathcal{A} = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ and will use $u = \mathsf{ACT}$ and $\ell = 1$. In an interspersed-duplication system, for $\ell < |u|$, we have

$$\begin{aligned}
\delta_\ell^u = &-(|u| - 1)x_n^u + \sum_{i=1}^{\ell} x_n^{u_{1,i}} x_n^{u_{i+1,|u|-i}} \\
&+ \sum_{i=1}^{\ell} x_n^{u_{1,|u|-i}} x_n^{u_{|u|-i+1,i}} + \sum_{i=1}^{|u|-\ell-1} x_n^{u_{1,i}u_{i+\ell+1,|u|-\ell-i}} x_n^{u_{i+1,\ell}}.
\end{aligned}$$

Here, the term $-(|u| - 1)x_n^u$ accounts for the expected number of lost occurrences of u in s as a result of inserting the duplicate substring. For example, an occurrence of $u = \mathsf{ACT}$ will be lost if the symbol G is duplicated and inserted after A in this occurrence of u, since it becomes AGCT. The probability that a certain occurrence is lost equals $\frac{|u|-1}{L_n}$. Since there are μ_n^u such occurrences, the expected number of lost occurrences of u equals $\mu_n^u \frac{|u|-1}{L_n} = x_n^u(|u| - 1)$. Note that if the symbol T is duplicated and inserted after C in an occurrence of ACT, we still count the original occurrence as lost, but count a new occurrence in the resulting ACTT, as seen in what follows. We now explain the first summation above. This summation represents the newly created occurrences of u where the first i symbols come from the duplicate and the next $|u| - i$ are from the substring that starts after the point of insertion of the duplicate. There are $\mu_n^{u_{1,i}}$ occurrences of $u_{1,i}$. The duplicate

starts with one of these with probability $\frac{\mu_n^{u_{1,i}}}{L_n} = x_n^{u_{1,i}}$. Furthermore, the duplicate is inserted before an occurrences of $u_{i+1,|u|-i}$ with probability $x_n^{u_{i+1,|u|-i}}$. Hence, the probability of a new occurrence created in this way is $x_n^{u_{1,i}} x_n^{u_{i+1,|u|-i}}$, and so is the expected number of such new occurrences. The role of the second summation is similar, except that the duplicate provides the second part of u. The last summation accounts for new occurrences of u in which the duplicate substring forms a middle part of u of length ℓ and previously existing substrings contribute a prefix of length i and a suffix of length $|u|-\ell-i$. In terms of our running example with $u = \mathsf{ACT}$ and $\ell = 1$, one such new occurrence is created if C is duplicated and inserted after A in an occurrence of AT. The probability of such an event is $x_n^{u_{1,i}u_{i+\ell+1,|u|-\ell-i}} x_n^{u_{i+1,\ell}} = x_n^{\mathsf{AT}} x_n^{\mathsf{C}}$, where $i = 1$.

For $\ell \geq |u|$, we have

$$\delta_\ell^u = -(|u|-1)x_n^u + \sum_{i=1}^{|u|-1} x_n^{u_{1,|u|-i}} x_n^{u_{|u|-i+1,i}} + \sum_{i=1}^{|u|-1} x_n^{u_{1,i}} x_n^{u_{i+1,|u|-i}} + (\ell - |u| + 1)x_n^u$$

where the first two summations are similar to the first two summations for the case of $\ell < |u|$, but a term corresponding to the third summation is not present. The term $(\ell - |u| + 1)x_n^u$ corresponds to the cases in which a new occurrence of u is created as a substring of the duplicate substring. Note that δ_ℓ^u is bounded, depends only on $\boldsymbol{x}_n$, and is Lipschitz since $\boldsymbol{x}_n \in [0,1]^{|U|}$.

Since $h_\ell^u(\boldsymbol{x}) = \delta_\ell^u(\boldsymbol{x}) - \ell x^u$, we have

$$h_\ell^u(\boldsymbol{x}) = -(\ell + |u| - 1)x^u + \sum_{i=1}^{\ell} x_n^{u_{1,i}} x_n^{u_{i+1,|u|-i}} + \sum_{i=1}^{\ell} x_n^{u_{1,|u|-i}} x_n^{u_{|u|-i+1,i}} + \sum_{i=1}^{|u|-\ell-1} x_n^{u_{1,i}u_{i+\ell+1,|u|-\ell-i}} x_n^{u_{i+1,\ell}} \tag{4}$$

for $\ell < |u|$, and

$$h_\ell^u(\boldsymbol{x}) = -2(|u|-1)x^u + 2\sum_{i=1}^{|u|-1} x^{u_{1,i}} x^{u_{i+1,|u|-i}} \tag{5}$$

for $\ell \geq |u|$. Recall that $\boldsymbol{h}_\ell(\boldsymbol{x}) = (h_\ell^v(\boldsymbol{x}))_{v\in U}$. So from (4) and (5), we can find the ode $d\boldsymbol{x}_t/dt = \boldsymbol{h}(\boldsymbol{x}_t) = \sum_\ell q_\ell \boldsymbol{h}_\ell(\boldsymbol{x}_t)$. As an example, if $\ell'_{max} = 2$ and $\mathcal{A} = \{\mathsf{A}, \mathsf{C}\}$, then $U = (\mathsf{A}, \mathsf{C}, \mathsf{AA}, \mathsf{AC}, \mathsf{CA}, \mathsf{CC}, \mathsf{AAA}, \ldots, \mathsf{CCC})$ and some of the equations of the ode system are

$$\begin{aligned}
&\frac{d}{dt}x_t^{\mathsf{A}} = \frac{d}{dt}x_t^{\mathsf{C}} = 0,\\
&\frac{d}{dt}x_t^{\mathsf{AA}} = -2x_t^{\mathsf{AA}} + 2\big(x_t^{\mathsf{A}}\big)^2, \quad \frac{d}{dt}x_t^{\mathsf{AC}} = -2x_t^{\mathsf{AC}} + 2x_t^{\mathsf{A}}x_t^{\mathsf{C}},\\
&\frac{d}{dt}x_t^{\mathsf{AAC}} = -(4-q_1)x_t^{\mathsf{AAC}} + 2x_t^{\mathsf{A}}x_t^{\mathsf{AC}} + (2-q_1)x_t^{\mathsf{C}}x_t^{\mathsf{AA}}.
\end{aligned} \tag{6}$$

For a vector $\boldsymbol{x}$ that contains the elements $(x^a)_{a\in\mathcal{A}}$ and for $v \in \mathcal{A}^*$, define $p(v,\boldsymbol{x}) = \prod_{a\in\mathcal{A}}(x^a)^{n_v(a)}$ and note that $p(vw,\boldsymbol{x}) = p(v,\boldsymbol{x})p(w,\boldsymbol{x})$. We now turn to find the solutions to the ode $d\boldsymbol{x}_t/dt = \boldsymbol{h}(\boldsymbol{x}_t)$.

Lemma 4. *Consider the ode $d\boldsymbol{x}_t/dt = \boldsymbol{h}(\boldsymbol{x}_t)$ where $\boldsymbol{h}(\boldsymbol{x}) = \sum_\ell q_\ell \boldsymbol{h}_\ell(\boldsymbol{x})$ and the elements of $\boldsymbol{h}_\ell(\boldsymbol{x})$ are given by (4) and (5). The solution to this ode is*

$$x_t^v = p(v,\boldsymbol{x}_0) + \sum_i b_i^v e^{-d_i^v t}, \qquad v \in U, \tag{7}$$

where $\boldsymbol{x}_0 = \boldsymbol{x}_t|_{t=0}$; the range of i in the summation is finite; and b_i^v and d_i^v are constants with $d_i^v > 0$.

Proof: We prove the lemma by induction. The claim (7) holds for $v \in \mathcal{A}$, since the equations for x_t^a, $a \in \mathcal{A}$, are of the form $dx_t^a/dt = 0$ and so $x_t^a = x_0^a$. Fix $u \in U$ such that $|u| > 1$, and assume that (7) holds for all $v \in U$ such that $|v| < |u|$. We show that it also holds for u, i.e., $x_t^u = p(u,\boldsymbol{x}_0) + \sum_i b_i^u e^{-d_i^u t}$. Using the assumption, we rewrite (4) and (5) as

$$h_\ell^u(\boldsymbol{x}_t) = -(\ell + |u| - 1)(x_t^u - p(u,\boldsymbol{x}_0)) + \sum_i b_i^{'} e^{-d_i^{'} t}$$

for $\ell < |u|$, and

$$h_\ell^u(\boldsymbol{x}_t) = -2(|u|-1)(x_t^u - p(u,\boldsymbol{x}_0)) + \sum_i b_i^{''} e^{-d_i^{''} t}$$

for $\ell \geq |u|$, where $b_i^{'}, d_i^{'}, b_i^{''}, d_i^{''}$ are constants with $d_i^{'}, d_i^{''} > 0$. Hence, $h^u(\boldsymbol{x}_t)$ can be written as

$$h^u(\boldsymbol{x}_t) = -c^u(x_t^u - p(u,\boldsymbol{x}_0)) + \sum_i b_i^{'''} e^{-d_i^{'''} t},$$

where $c^u = 2|u| - 2 - \sum_{\ell=1}^{|u|-1} q_\ell(|u| - 1 - \ell)$, and $b_i^{'''}, d_i^{'''}$ are constants with $d_i^{'''} > 0$. Thus the solution to the ode $dx_t^u/dt = h^u(\boldsymbol{x}_t)$ is

$$\begin{aligned}
x_t^u &= e^{-c^u t}\int e^{c^u t'}\Big(c^u p(u,\boldsymbol{x}_0) + \sum_i b_i^{'''} e^{-d_i^{'''} t'}\Big)dt' + \bar{b}e^{-c^u t}\\
&= p(u,\boldsymbol{x}_0) + \sum_i b_i^u e^{-d_i^u t},
\end{aligned}$$

where $\bar{b}, b_i^u, d_i^u$ are some constants, with $d_i^u > 0$ (note that $c^u > 0$ since $|u| > 1$). This completes the proof. ∎

For example, the solutions to (6) with $q_1 = 0$ are

$$\begin{aligned}
&x_t^{\mathsf{A}} = x_0^{\mathsf{A}}, \qquad x_t^{\mathsf{C}} = x_0^{\mathsf{C}},\\
&x_t^{\mathsf{AA}} = \big(x_0^{\mathsf{A}}\big)^2 + b_1^{\mathsf{AA}}e^{-2t}, \quad x_t^{\mathsf{AC}} = x_0^{\mathsf{A}}x_0^{\mathsf{C}} + b_1^{\mathsf{AC}}e^{-2t},\\
&x_t^{\mathsf{AAC}} = \big(x_0^{\mathsf{A}}\big)^2 x_0^{\mathsf{C}} + b_1^{\mathsf{AAC}}e^{-2t} + b_2^{\mathsf{AAC}}e^{-4t},
\end{aligned}$$

where $b_1^{\mathsf{AAC}} = x_0^{\mathsf{A}}b_1^{\mathsf{AC}} + x_0^{\mathsf{C}}b_1^{\mathsf{AA}}$.

In the next theorem, we use Lemma 4 to characterize the limits of the frequencies of substrings in interspersed-duplication systems.

Theorem 5. *Let U be an ordered set consisting of all strings over the alphabet $\mathcal{A}$ of a certain maximum length and let*

$\boldsymbol{x}_n = (x_n^u)_{u\in U}$ be the vector of frequencies of these strings at time n in an interspersed-duplication system. The vector $\boldsymbol{x}_n$ converges almost surely. Furthermore, its limit $\boldsymbol{x}_\infty$ satisfies

$$x_\infty^u = \prod_{a\in\mathcal{A}} (x_\infty^a)^{n_u(a)}, \quad \textit{for all } u \in U.$$

Proof: From Theorem 3, we know that the limit set of $\boldsymbol{x}_n$ is an internally chain transitive invariant set of the ode described by (4) and (5). Let this set, which consists of points of the form $\boldsymbol{y} = (y^v)_{v\in U}$, be denoted by A. Since for each $u \in U$, $x_n^u \in [0,1]$, we have that $A \subseteq [0,1]^{|U|}$. We now use these facts to show that for each $\boldsymbol{y} \in A$ and $u \in U$, we have $y^u = p(u, \boldsymbol{y})$.

Suppose to the contrary that there exist $\boldsymbol{y} \in A$ and $u \in U$ such that $y^u \neq p(u, \boldsymbol{y})$. Among all possible choices of such $\boldsymbol{y}$ and u, choose the ones where the length $|u|$ of u is minimum. Hence, $y^u \neq p(u, \boldsymbol{y})$ but $z^v = p(v, \boldsymbol{z})$ for all $v \in \mathcal{A}^*$ with $|v| < |u|$, and all $\boldsymbol{z} \in A$. Then, similar to the proof of Lemma 4, one can show that if $\boldsymbol{x}_0 = \boldsymbol{z} \in A$, then $x_t^u = p(u, \boldsymbol{z}) + be^{-c^u t}$, where $b = z^u - p(u, \boldsymbol{z})$ and $c^u \geq |u|$.

By the definition of internal chain transitivity, for any $\epsilon > 0$ and $T > 0$, there exist $N \geq 1$ and a sequence $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_N$ with $\boldsymbol{y}_i \in A$, $\boldsymbol{y}_0 = \boldsymbol{y}_N = \boldsymbol{y}$ such that for $0 \leq i < n$, if $\boldsymbol{x}_0 = \boldsymbol{y}_i$, then there exists $t \geq T$ such that $\boldsymbol{x}_t$ is in the ϵ-neighborhood of $\boldsymbol{y}_{i+1}$. Suppose $\boldsymbol{x}_0 = \boldsymbol{y}_i$ and suppose for $t' \geq T$, $\boldsymbol{x}_t$ is in the ϵ-neighborhood of $\boldsymbol{y}_{i+1}$. We have

$$\begin{aligned} y_{i+1}^u &\leq x_{t'}^u + \epsilon \leq p(u, \boldsymbol{y}_i) + (y_i^u - p(u, \boldsymbol{y}_i))e^{-c^u t'} + \epsilon \\ &\leq p(u, \boldsymbol{y}_i) + e^{-c^u T} + \epsilon, \end{aligned}$$

where we have used the fact that $y_i^u \leq 1$. Furthermore, since $y_{i+1}^a \leq y_i^a + \epsilon$ for $a \in \mathcal{A}$, we find

$$p\big(u, \boldsymbol{y}_{i+1}\big) - p(u, \boldsymbol{y}_i) \geq (1+\epsilon)^{|u|} - 1,$$

where we have again used the fact that $y_i^a \leq 1$ for $a \in \mathcal{A}$. It thus follows that

$$y_{i+1}^u - p\big(u, \boldsymbol{y}_{i+1}\big) \leq e^{-c^u T} + \epsilon + (1+\epsilon)^{|u|} - 1.$$

In particular, this holds for $i = n-1$, i.e.,

$$y^u - p(u, \boldsymbol{y}) \leq e^{-c^u T} + \epsilon + (1+\epsilon)^{|u|} - 1.$$

But we can make the right side of the above inequality arbitrary small by choosing T large enough and ϵ small enough. Thus $y^u = p(u, \boldsymbol{y})$, which is a contradiction. Hence, for each $\boldsymbol{y} \in A$ and $u \in U$, we have $y^u = p(u, \boldsymbol{y})$, and the theorem follows. ∎

In words, the theorem shows that for $u \in \mathcal{A}^*$, the frequency of u converges to the frequency of same in an iid sequence where the probability of $a \in \mathcal{A}$ equals x_∞^a. Figure 2 illustrates an example, obtained via simulation, where the system starts with $s^{(0)} = \mathsf{AGCGTATGCG}$ and duplications of lengths 4 and 6 occur with equal probability. As the number n of duplications increases, the frequency vector $\boldsymbol{x}_n$ becomes more compatible with that of an iid sequence. For example for $n = 15000$, we have $x_n^{\mathsf{AC}} = 0.0251 \simeq x_n^{\mathsf{A}} x_n^{\mathsf{C}} = 0.0266$, $x_n^{\mathsf{GT}} = 0.0872 \simeq x_n^{\mathsf{G}} x_n^{\mathsf{T}} = 0.0880$, and $x_n^{\mathsf{GGG}} = 0.0992 \simeq \big(x_n^{\mathsf{G}}\big)^3 = 0.1084$

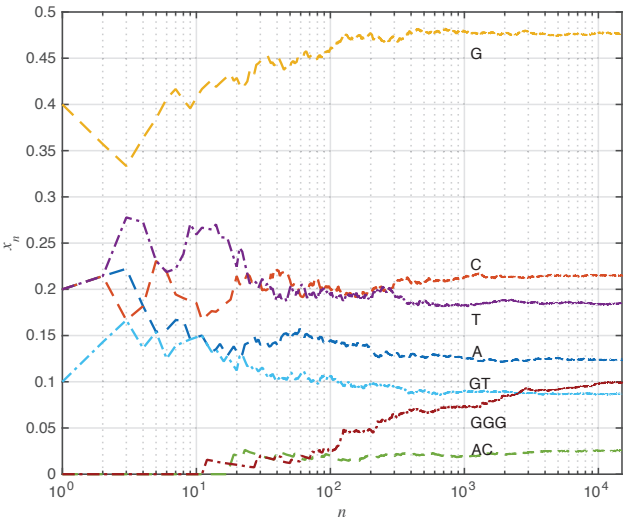

Figure 2. Symbol frequencies vs the number of duplications in an interspersed-duplication system, with $s^{(0)} = \mathsf{AGCGTATGCG}$, and $q_4 = q_6 = 1/2$.

VI. Conclusion

We studied the limiting behavior of stochastic interspersed-duplication systems in order to evaluate their ability in creating biological diversity. We showed that the composition of long sequences does not vary greatly in random duplication systems, given that all substrings of the same length are duplicated with equal probability. We also established that frequencies of sequences in interspersed-duplication systems tend to the corresponding probabilities in sequences generated by iid sources, which have the highest possible entropy for given symbol probabilities. It thus seems plausible that diversity may arise from random interspersed-duplication events. Since this work was limited to the asymptotic analysis of these systems, further research is required to quantify their finite-time behavior. Furthermore, in this work we did not consider the more realistic scenarios in which different duplications have different probabilities.

References

[1] V. S. Borkar, "Stochastic approximation," *Cambridge Books*, 2008.

[2] F. Farnoud, M. Schwartz, and J. Bruck, "The capacity of string-duplication systems," in *Proc. IEEE Int. Symp. Information Theory*, Honolulu, HI, USA, Jun. 2014, pp. 1301–1305.

[3] ——, "The capacity of string-replication systems," *Submitted to IEEE Trans. Inf. Theory, arXiv preprint: http://arxiv.org/abs/1401.4634*, 2014.

[4] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[5] C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm, "How many species are there on earth and in the ocean?" *PLoS Biol*, vol. 9, no. 8, Aug. 2011. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pbio.1001127