Evolution of k-mer Frequencies and Entropy in Duplication and Substitution Mutation Systems

Hao Lou, *Student Member, IEEE,* Moshe Schwartz, *Senior Member, IEEE,*
Jehoshua Bruck, *Fellow, IEEE,* and Farzad Farnoud (Hassanzadeh), *Member, IEEE*

***Abstract*—Genomic evolution can be viewed as string-editing processes driven by mutations. An understanding of the statistical properties resulting from these mutation processes is of value in a variety of tasks related to biological sequence data, e.g., estimation of model parameters and compression. At the same time, due to the complexity of these processes, designing tractable stochastic models and analyzing them are challenging. In this paper, we study two kinds of systems, each representing a set of mutations. In the first system, tandem duplications and substitution mutations are allowed and in the other, interspersed duplications. We provide stochastic models and, via stochastic approximation, study the evolution of substring frequencies for these two systems separately. Specifically, we show that k-mer frequencies converge almost surely and determine the limit set. Furthermore, we present a method for finding upper bounds on entropy for such systems.**

***Index Terms*—String-duplication systems, substitution mutation, entropy**

I. Introduction

DUE to advances in DNA sequencing, vast amounts of biological sequence data are available nowadays. Developing efficient methods for the analysis and storage of this type of data will benefit from gaining a better mathematical understanding of the structure of these sequences. Biological sequences are formed by genomic mutations, which alter the sequence in each generation to create a new sequence in the next generation. These processes can be viewed as stochastic string editing operations that shape the statistical properties of sequence data.

In this paper, our goal is to gain a better understanding of the evolution of sequences under random mutations. We represent the evolutionary process as a stochastic system in which an arbitrary initial string evolves through random mutation events. In such systems, we will study the evolution of the frequencies of words of length k, i.e., k-mers, as the sequence evolves. The analysis of k-mers has various applications, including identifying functions and evolutionary features [48]. Alignment-free sequence comparison also relies on k-mer frequencies [54]. Their analysis is also of interest because other statistical properties can be computed from k-mer frequencies.

From an information-theoretic point of view, stochastic sequence generation process through mutation can be viewed as a *source* of information. We study the entropy of such sources, which can be viewed as representing the complexity of sequences generated by the source. Sequence complexity measures, including entropy, have been used to determine the origin and/or the role of DNA sequences [16], [43], [53], for example to classify protein-coding and non-coding regions of a genome. The entropy of a source also determines how well the sequences it produces can be compressed, an increasingly important problem given the growth of biological data.

Several types of mutations exist, including substitution, duplication, insertion, and deletion. Substitution refers to changing a symbol in the sequence, e.g., $\mathsf{ACGTCT} \to \mathsf{ACG\underline{C}CT}$. Duplication mutations, where a segment of DNA (called the template) is copied and inserted elsewhere in the genome, may be of the *tandem* or *interspersed* type. In tandem duplication, the copy is inserted immediately after the template. For example, from ACGTCT, we may obtain $\mathsf{AC\overline{GT}\underline{GT}CT}$, where the template is overlined and the copy is underlined. For interspersed duplication, there is generally no relationship between where the template is located and where the copy is inserted. As an example, two possibilities for AGTTC after a single interspersed duplication are $\mathsf{\overline{AGT}T\underline{AGT}C}$ and $\mathsf{\overline{AG}\underline{AGT}\overline{T}TC}$. Our focus will be on duplication mutations, which are thought to play a critical role in generating new genetic material [42].

Tandem duplication is generally thought to be caused by slipped-strand mispairings [41], where during DNA synthesis, one strand in a DNA duplex becomes misaligned with the other. Tandem duplications and substitutions, along with other mutations, lead to tandem repeats, i.e., stretches of DNA in which the same pattern is repeated many times. Tandem repeats are known to cause important phenomena such as chromosome fragility [50]. Interspersed duplications are caused by transposons, or "jumping genes", which are elements in the genome that can "copy/paste" themselves into different locations. Interspersed duplication is of interest as it leads to interspersed repeats, which form 45% of the human genome [29].

We will analyze two systems involving the types of duplica-

This paper was presented in part at ISIT 2018 [36] and ISIT 2015 [17].

Hao Lou is with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA, 22903, USA, (email: hl2nu@virginia.edu).

Moshe Schwartz is with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer Sheva 8410501, Israel (e-mail: schwartz@ee.bgu.ac.il).

Jehoshua Bruck is with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125, USA (e-mail: bruck@paradise.caltech.edu).

Farzad Farnoud (Hassanzadeh) is with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA, 22903, USA, (email: farzad@virginia.edu).

This work was supported in part by a United States–Israel Binational Science Foundation (BSF) grant no. 2017652, and by NSF grants under grant nos. CCF-1816409, CCF-1755773, CCF-1816965, and CCF-1717884.

tions discussed. The first system models a sequence evolving through tandem duplications and substitutions (TDS) and the second system represents interspersed duplications (ID). Along with duplications, other types of mutations occur. But for simplicity, our attention is limited to the aforementioned systems, and we leave more comprehensive analysis to future work. Furthermore, the significantly more complex effect of natural selection is not considered.

In TDS systems, in each step, i) a randomly chosen substring of the sequence is duplicated and inserted in tandem, or ii) a position is chosen at random and the symbol in that position is changed to one of the other symbols. In ID systems, 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.

Our analysis starts by considering how k-mer frequencies evolve as mutations occur. To analyze their evolution, we use the stochastic-approximation method, which enables modeling of a discrete dynamic system by a corresponding continuous system described by an ordinary differential equation (ODE). For the TDS model, our approach allows us to compute the limit for the frequency of any k-mer as a function of model parameters. We will then use these results to provide bounds on the entropy of sequences generated by tandem duplications and substitutions. For the ID model, we show that 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 structure of the limit set for k-mer frequencies in ID systems, however, leads to trivial upper bounds on the entropy. However, in certain cases these bounds are satisfied with equality. Parts of the paper have been presented at the International Symposium on Information Theory [17], [36]. Relative to those, the current paper presents omitted proofs along with additional examples and illustrations.

In previous work, the related problem of finding the combinatorial capacity of duplication systems has been studied. The combinatorial capacity is related to entropy but is defined based on the size of the set of sequences that can be generated by the system, without considering their probabilities. The combinatorial capacity is studied by [19], [25], for duplication systems (without allowing other types of mutations) and by [24] for systems with both tandem duplication and substitution. Compared to combinatorial capacity, entropy, which is studied in this paper, provides a more accurate measure of the complexity and compressibility of sequences generated by the system. For duplication systems and duplication/substitution systems, entropy has been studied by [12]. While this work considers a wider range of systems, it only allows duplications involving single symbols. Furthermore, it does not study k-mer frequencies. The stochastic-approximation framework has been used for estimation of model parameters in tandem duplication systems [18]. Estimating the entropy of DNA sequences has been studied in [16], [35], [47]. However these works focus on estimating the entropy from a given sequence, rather than computing the entropy of a stochastic sequence generation system that models evolution. Duplication systems have also been studied in the context of designing error-correcting codes [6], [11], [26], [31].

From a broader perspective, information theory has natural applications in biology since the processing and transmission of information are ubiquitous in living organisms, from genetic to ecological inheritance mechanisms [52]. Research towards the intersection of information theory and biology can be traced back to the paper "The information content and error rate of living things" [9] in 1949 (just one year after Shannon's seminal paper on information theory). Since then, efforts have been made to address many problems in biology with information-theoretic methods, and have been successful in areas such as predicting the correlation between DNA mutations and disease, identifying protein binding sequences in nucleic acids, and analyzing neural spike trains and higher functionalities of cognitive systems [39]. Recently, due to the symbolism of biological sequences, information theory has found various applications in molecular biology, regarding which [1], [2], [21] serve as excellent surveys. For example, [33] introduced a universal sequence distance based on the information theoretical concept of Kolmogorov complexity and applied it in constructing genome phylogeny; [21] studied the possibility of using mutual information for gene mapping and marker clustering; and [40] studied the the minimum number of reads required for an assembly DNA sequencing algorithm to reconstruct the original sequence. Moreover, two essential areas of information theory, data compression and channel coding, both have direct and practical applications in biology. Compressing biological data has become an inevitable need as the amount of biological sequencing data grows explosively. Many compression algorithms have been designed targeting DNA/RNA sequences [4], [5], [7], [8], [28], [44]. This paper is related to this line of research since determining the entropy of sequences provides bounds on the performance of compression methods. On the other hand, DNA storage is also attracting increased attention due to the longevity and enormous information density of DNA. With challenges arising from the existence of diverse error types in DNA synthesis, replication, and sequencing, many techniques in information theory, especially coding schemes, have been studied and used to enhance the reliability of DNA storage system [22], [27], [30], [32], [45], [49].

The rest of the paper is organized as follows. Notation and preliminaries are given in the next section. In Section III, we present the framework for the application of stochastic approximation to our string-duplication systems. Section IV contains the analysis of the evolution of k-mer frequencies in tandem duplication systems and the proof of entropy bounds. Section V is devoted to the analysis of k-mer frequencies in strings undergoing random interspersed duplications. We close the paper with concluding remarks in Section VI.

II. Notation and Preliminaries

For a positive integer m, let $[m] = \{1, \dots, m\}$. For a finite alphabet $\mathcal{A}$, the set of all finite strings over $\mathcal{A}$ is denoted $\mathcal{A}^*$, and the set of all finite non-empty strings is denoted $\mathcal{A}^+$. Also, let $\mathcal{A}^k$ denote the set of k-mers, i.e., length-k strings, over $\mathcal{A}$. The elements in strings are indexed starting from 1, e.g., $\boldsymbol{s} = s_1 \cdots s_m$, where $|\boldsymbol{s}| = m$ is the length of $\boldsymbol{s}$. For a string $\boldsymbol{u} \in \mathcal{A}^*$, $\boldsymbol{u}_{i,j}$ denotes the length-j substring of $\boldsymbol{u}$ starting at u_i. Furthermore, the concatenation of two strings $\boldsymbol{u}$ and $\boldsymbol{v}$ is denoted by $\boldsymbol{uv}$. For a non-negative integer j, and $\boldsymbol{u} \in \mathcal{A}^*$, $\boldsymbol{u}^j$ is a concatenation of j copies of $\boldsymbol{u}$. Vectors and strings are denoted by boldface letters such as $\boldsymbol{x}$, while scalars and symbols by normal letters such as x.

Consider an initial string $\boldsymbol{s}_0$ and a process where in each step a random transform, or "mutation", is applied to $\boldsymbol{s}_n$, resulting in $\boldsymbol{s}_{n+1}$. To avoid the complications arising from boundaries, we assume the strings $\boldsymbol{s}_n$ are circular, with a given origin and direction. Let the length of $\boldsymbol{s}_n$ be denoted by L_n. To a duplication of length ℓ, which may be tandem or interspersed depending on the model under study, we assign probability q_ℓ. For TDS systems, in which substitutions are present, we denote the probability of substitution with q_0. For ID systems, we let $q_0 = 0$. The position of the template in duplication mutations is chosen at random among the $|\boldsymbol{s}_n|$ possible options. For interspersed duplication, the position at which the copy is inserted is also chosen randomly. Furthermore, for substitution mutations, the position of the symbol that is substituted is chosen randomly. We assume there exists M such that $q_\ell = 0$ for all $\ell \geqslant M$. Hence, we have $\sum_{\ell=0}^{M-1} q_\ell = 1$.

For a string $\boldsymbol{u} \in \mathcal{A}^+$, denote the number of appearances of $\boldsymbol{u}$ in $\boldsymbol{s}_n$ as $\mu_n^{\boldsymbol{u}}$, and its frequency as $x_n^{\boldsymbol{u}}$, where $x_n^{\boldsymbol{u}} = \mu_n^{\boldsymbol{u}}/L_n$. For example, if $\boldsymbol{s}_n = \mathsf{ACGAC}$, then $\mu_n^{\mathsf{AC}} = 2, x_n^{\mathsf{AC}} = \frac{2}{5}$. Furthermore, for any set $U \subseteq \mathcal{A}^+$, we define $\boldsymbol{\mu}_n = (\mu_n^{\boldsymbol{u}})_{\boldsymbol{u} \in U}$, and $\boldsymbol{x}_n = (x_n^{\boldsymbol{u}})_{\boldsymbol{u} \in U}$. Thus $\boldsymbol{\mu}_n$ is a vector representing the number of appearances of $\boldsymbol{u} \in U$ in the string $\boldsymbol{s}$ at time n and $\boldsymbol{x}_n$ is the normalized version of $\boldsymbol{\mu}_n$.

We now provide an informal review of some concepts from probability theory that will be of use in this paper. For further detail, we refer the reader to [10]. For a sequence of random variables $y_n, n = 0, 1, 2, \ldots$, the filtration $\mathcal{F}_n$ associated with the process represents the information provided by $y_0, \ldots, y_n$. Formally, $\mathcal{F}_n$ is the sigma-algebra $\sigma(y_0, \ldots, y_n)$. The process y_n is a martingale if $\mathbb{E}[y_{n+1}|\mathcal{F}_n] = y_n$. Intuitively, this says that given knowledge of what has happened so far, the expected value of y in the future is equal to its current value. The process y_n is called a martingale difference sequence if $\mathbb{E}[y_{n+1}|\mathcal{F}_n] = 0$. Moreover, we introduce two important results about martingales in the following, Doob's convergence theorem and the Hoeffding-Azuma inequality. Doob's martingale convergence theorem states that if a martingale y_n satisfies $\sup_n \mathbb{E}[|y_n|] < \infty$, then almost surely $y_\infty = \lim_n y_n$ exists and is finite in expectation. The Hoeffding-Azuma inequality states that for a martingale y_n, if $|y_n - y_{n-1}| \leqslant c_n$ almost surely, then for all positive integers N and all positive reals λ,

$$\Pr(|y_n - y_0| \geqslant \lambda) \leqslant 2 \exp\left(\frac{-\lambda^2}{2 \sum_{n=1}^{N} c_n^2} \right).$$

We let $\{\mathcal{F}_n\}$ be the filtration [10] generated by the random variables $\{\boldsymbol{x}_n, L_n\}$, which roughly speaking represents the information contained in all $\boldsymbol{x}_i$ and L_i with $0 \leqslant i \leqslant n$.

Before proceeding to the analysis of k-mer frequencies, we present two results for the evolution of symbol frequencies (1-mers). These results can be viewed as extensions of results for Pólya urn models [37]. In such models, a random ball is chosen from an urn containing balls of different colors. The chosen ball is returned to the urn, along with a predetermined number of balls of the same color. It is known that, conditioned on the present state, the expected ratio of the balls of each color (equivalent to symbol frequencies) in the future is equal to the present value and therefore by definition is a martingale and converges almost surely. While strings are more complex objects than urns, we describe similar results that are valid for any duplication process in which for each i, all i-substring of $\boldsymbol{s}$ have the same chance of being duplicated. In particular, these results hold both for TDS systems with $q_0 = 0$ and for ID systems.

Theorem 1. *In a duplication system with $q_0 = 0$, the random variables x_n^a, $a \in \mathcal{A}$, are martingales and converge almost surely.*

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

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

We thus have $\mathbb{E}\left[x_{n+1}^a|\mathcal{F}_n\right] = x_n^a$ and so x_n^a is a martingale. Since it is nonnegative, by the martingale convergence theorem, it converges almost surely. ■

Remark. The above theorem does not in fact require the distribution q to be constant and bounded. Under our assumption that q is so, we can in addition obtain the following result on the probability of x_n^a deviating from its starting value.

Theorem 2. *For all $a \in \mathcal{A}$ and $n \geqslant 1$ we have*

$$\Pr(|x_n^a - x_0^a| \geqslant \lambda) \leqslant 2e^{-\lambda^2 L_0/(2M^2)} .$$

Proof: Since $q_i = 0$ for $i \geqslant M$ or $i \leqslant 0$, $\frac{\mu_{n-1}^a}{L_{n-1}+M} \leqslant \frac{\mu_n^a}{L_n} \leqslant \frac{\mu_{n-1}^a + M}{L_{n-1}+M}$. Thus

$$-\frac{M\mu_{n-1}^a}{L_{n-1}(L_{n-1}+M)} \leqslant \frac{\mu_n^a}{L_n} - \frac{\mu_{n-1}^a}{L_{n-1}} \leqslant \frac{M(L_{n-1} - \mu_{n-1}^a)}{L_{n-1}(L_{n-1}+M)} ,$$

implying that

$$\begin{aligned}\left|x_n^a - x_{n-1}^a\right| &\leqslant \frac{M \max\left\{L_{n-1} - \mu_{n-1}^a, \mu_{n-1}^a\right\}}{L_{n-1}(L_{n-1}+M)} \\ &\leqslant \frac{M}{L_{n-1}+M} \leqslant \frac{M}{L_0 + n - 1 + M} \leqslant \frac{M}{L_0 + n} .\end{aligned}$$

Let $c_n = \frac{M}{L_0+n}$ so that $\left|x_n^a - x_{n-1}^a\right| \leqslant c_n$ and note that

$$\begin{aligned}\sum_{i=1}^{n} c_i^2 &= M^2 \sum_{i=1}^{n} \frac{1}{(L_0+i)^2} \leqslant M^2 \int_0^n \frac{dt}{(L_0+t)^2} \\ &= \frac{M^2}{L_0} - \frac{M^2}{L_0+n} = \frac{M^2 n}{L_0(L_0+n)} \leqslant \frac{M^2}{L_0}.\end{aligned}$$

By the Hoeffding-Azuma inequality [20], since $\{x_n^a : n = 0,1,2,\ldots\}$ is a martingale and $\left|x_n^a - x_{n-1}^a\right| \leqslant c_n$, we have

$$\begin{aligned}\Pr(|x_n^a - x_0^a| \geqslant \lambda) &\leqslant 2\exp\left(\frac{-\lambda^2}{2\sum_{i=1}^n c_i^2}\right)\\ &\leqslant 2\exp\left(\frac{-\lambda^2 L_0}{2M^2}\right).\end{aligned}$$

■

The preceding theorem implies that it is unlikely for the composition of a long DNA sequence to change dramatically through random duplication events of bounded length. 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 $x_n^{\boldsymbol{u}}$ when $|\boldsymbol{u}| > 1$. Therefore, for analyzing such cases, we use the more flexible technique of stochastic approximation as described in the sequel.

III. Stochastic Approximation for Duplication Systems

In this section, we present an overview of the application of stochastic approximation in the analysis of duplication systems. By using stochastic approximation, our goal is to study how the k-mer frequencies vector $\boldsymbol{x}_n$ changes with n by finding a differential equation whose solution approximates $\boldsymbol{x}_n$.

A. Preliminaries

We start by providing the definitions used in this section. For any positive integer d, a subset of $\mathbb{R}^d$ is said to be *closed* if it contains its boundary, and is said to be *compact* if it is both closed and bounded. Moreover, a subset of $\mathbb{R}^d$ is *connected* if it is not a union of two nonempty separated sets [46]. 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 $N \geqslant 1$ 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 $0 \leqslant i < n$, if $\boldsymbol{z}_0 = \boldsymbol{y}_i$, then for some $t \geqslant T$, $\boldsymbol{z}_t$ is in the ϵ-neighborhood of $\boldsymbol{y}_{i+1}$ [3].

We will also make use of the following theorem, which enables studying the behavior of a discrete dynamical system through a system of differential equations.

Theorem 3. *(Stochastic Approximation Theorem [3, Theorem 2].) Let $\{\boldsymbol{z}_n, n \geqslant 0\}$ be a bounded discrete stochastic process in $\mathbb{R}^d$ with*

$$\boldsymbol{z}_{n+1} = \boldsymbol{z}_n + a(n)[\boldsymbol{h}(\boldsymbol{z}_n) + \boldsymbol{M}_{n+1}], \quad n \geqslant 0,$$

where $\{\boldsymbol{M}_n, n \geqslant 0\}$ is a bounded martingale difference sequence in $\mathbb{R}^d$ with $\mathbb{E}[\boldsymbol{M}_{n+1}|\boldsymbol{z}_m, \boldsymbol{M}_m, m \leqslant n] = 0$ almost surely, $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^d$ is a Lipschitz map, and $\{a(n), n \geqslant 0\}$ are positive scalars satisfying $\sum_n a(n) = \infty, \sum_n a(n)^2 < \infty$. Then $\{\boldsymbol{z}_n, n \geqslant 0\}$ converges almost surely to a compact connected internally chain transitive invariant set of the ODE

$$\dot{\boldsymbol{z}}_t = \boldsymbol{h}(\boldsymbol{z}_t), \quad t \geqslant 0.$$

Note the dual use of the symbol $\boldsymbol{z}$; the meaning is however clear from the subscript.

B. Stochastic Approximation in Duplication Systems

We present a set of conditions that will allow us to adapt duplication systems to the stochastic approximation framework, described in Theorem 3. Let $\mathbb{E}_\ell[\,\cdot\,]$ denote the expected value conditioned on the fact that the length of the duplicated substring is ℓ and let $\boldsymbol{\delta}_\ell = \mathbb{E}_\ell[\boldsymbol{\mu}_{n+1}|\mathcal{F}_n] - \boldsymbol{\mu}_n$. In the case of substitution, we let $\ell = 0$. We consider the following conditions.

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

(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.

(A1) holds by assumption. From this follows (A2) since for each k-mer, a mutation can create or eliminate a bounded number of occurrences. Additionally, (A3) holds because each element of $\boldsymbol{x}_n$ is between 0 and 1. The correctness of (A4) and (A5) will be shown for each system.

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

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

For the first term of the right side of (III-B), we have

$$\begin{aligned}\mathbb{E}[\boldsymbol{x}_{n+1} - \boldsymbol{x}_n|\mathcal{F}_n] &= \sum_{\ell=0}^{M-1} q_\ell(\mathbb{E}_\ell[\boldsymbol{x}_{n+1}|\mathcal{F}_n] - \boldsymbol{x}_n)\\ &= \sum_{\ell=0}^{M-1} q_\ell\left(\frac{\boldsymbol{\mu}_n + \boldsymbol{\delta}_\ell(\boldsymbol{x}_n)}{L_n + \ell} - \frac{\boldsymbol{\mu}_n}{L_n}\right)\\ &= \sum_{\ell=0}^{M-1} q_\ell \frac{L_n\delta_\ell(\boldsymbol{x}_n) - \ell\boldsymbol{\mu}_n}{L_n(L_n + \ell)}\\ &= \sum_{\ell=0}^{M-1} q_\ell \frac{\delta_\ell(\boldsymbol{x}_n) - \ell\boldsymbol{x}_n}{L_n + \ell}\\ &= \frac{1}{L_n}\sum_{\ell=0}^{M-1} q_\ell \boldsymbol{h}_\ell(\boldsymbol{x}_n)\big(1 + O\big(L_n^{-1}\big)\big)\\ &= \frac{1}{L_n}\boldsymbol{h}(\boldsymbol{x}_n)\big(1 + O\big(L_n^{-1}\big)\big), \qquad (1)\end{aligned}$$

where $\boldsymbol{h}_\ell(\boldsymbol{x}_n) = \boldsymbol{\delta}_\ell(\boldsymbol{x}_n) - \ell\boldsymbol{x}_n$, $\boldsymbol{h}(\boldsymbol{x}_n) = \sum_{\ell=0}^{M-1} q_\ell\boldsymbol{h}_\ell(\boldsymbol{x}_n)$, 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 (III-B), we have

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

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

From (III-B), (1), and (2), 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 $\boldsymbol{\delta}_\ell(\boldsymbol{x}_n)$ for all $0 \leqslant \ell < M$. We note that $\boldsymbol{h}$ determines the overall expected behavior of the system.

In the rest of the paper, the element of $\boldsymbol{\delta}_\ell(\boldsymbol{x}_n)$ that corresponds to $\boldsymbol{u}$ is denoted by $\delta_\ell^{\boldsymbol{u}}(\boldsymbol{x}_n)$. More precisely, $\delta_\ell^{\boldsymbol{u}}(\boldsymbol{x}_n) = \mathbb{E}_\ell[\mu_{n+1}^{\boldsymbol{u}} - \mu_n^{\boldsymbol{u}}|\mathcal{F}_n]$. This notation also extends to $\boldsymbol{h}$.

An additional condition requires $\sum_n 1/|\boldsymbol{s}_n| = \infty$ and $\sum_n 1/|\boldsymbol{s}_n|^2 < \infty$, which can be proven using the Borel-Cantelli lemma [20] if $q_0 < 1$. Given these and our discussion above, the following theorem, which relates the discrete system describing $\boldsymbol{x}_n$ to a continuous system, follows directly from Theorem 3.

Theorem 4. *The vector of k-mer frequencies $\boldsymbol{x}_n$ converges almost surely to a compact connected internally chain transitive invariant set of the ODE $d\boldsymbol{x}_t/dt = \boldsymbol{h}(\boldsymbol{x}_t)$.*

IV. Tandem Duplication with Substitution

In this section, we study the behavior of a system that allows tandem duplication and substitution mutations. First, we will determine the limits of the frequencies of k-mers. Then, after presenting a theorem relating the limits to entropy, we find bounds on the entropy of these systems.

Let $U = \mathcal{A}^k$, so $\boldsymbol{\mu}_n$ is the vector of all k-mer occurrences, and $\boldsymbol{x}_n$ is the vector of all k-mer frequencies. From Section III we know that we can use the differential equation $d\boldsymbol{x}_t/dt = \boldsymbol{h}(\boldsymbol{x}_t)$ to determine the limit of k-mer frequencies. To find the differential equation, in Theorem 8, we determine $\delta_\ell^{\boldsymbol{u}}(\boldsymbol{x}_n)$ for ℓ with $q_\ell > 0$ and $\boldsymbol{u} \in U$, where it can be observed that (A.4) and (A.5) hold in our model

In the next subsection, we will give some necessary definitions. We will then prove that $\delta_\ell^{\boldsymbol{u}}(\boldsymbol{x}_n)$ is a linear function of $\boldsymbol{x}_n$, which leads to a linear first-order differential equation. This linear form facilitates determining the asymptotic behavior of the k-mer frequencies. We will then show that the entropy of stochastic string systems can be related to the capacity of semiconstrained systems defined by the limit set of the k-mer frequencies. Leveraging the simple form of the limits for systems with tandem duplications and substitutions, we will provide bounds on the entropy of these systems.

A. Preliminaries

The following definitions will be useful for finding $\boldsymbol{\delta}_\ell(\boldsymbol{x}_n)$.

Definition 1. For $\boldsymbol{u} \in \mathcal{A}^*$ and $m \in \mathbb{N}^+$, define $\boldsymbol{\phi}_m(\boldsymbol{u})$ to be a sequence of length $|\boldsymbol{u}|$ whose i-th element is determined by whether the symbol in position i of $\boldsymbol{u}$ equals the symbol in position $i - m$. More specifically, the i-th element of $\boldsymbol{\phi}_m(\boldsymbol{u})$ is

$$\boldsymbol{\phi}_m(\boldsymbol{u})_i = \begin{cases} 0, & m+1 \leqslant i \leqslant |\boldsymbol{u}|, u_i = u_{i-m} \\ \mathsf{X}, & \text{otherwise} \end{cases}$$

where X is a dummy variable. Let the lengths of the maximal runs of 0s immediately after the initial X^m and at the end of $\boldsymbol{\phi}_m(\boldsymbol{u})$ be denoted by $l_m^{\boldsymbol{u}}$ and $r_m^{\boldsymbol{u}}$, respectively.

Note that either of $l_m^{\boldsymbol{u}}$ or $r_m^{\boldsymbol{u}}$ may be equal to 0. If $\boldsymbol{\phi}_m(\boldsymbol{u}) = \mathsf{X}^m 0^{|\boldsymbol{u}|-m}$, then $l_m^{\boldsymbol{u}} = r_m^{\boldsymbol{u}} = |\boldsymbol{u}| - m$. Otherwise, we have $\boldsymbol{\phi}_m(\boldsymbol{u}) = \mathsf{X}^m 0^{l_m^{\boldsymbol{u}}} Y 0^{r_m^{\boldsymbol{u}}}$, for some Y that starts and ends with X.

Example. For $\mathcal{A} = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$, we have

$$\begin{aligned} \boldsymbol{u} =& \mathsf{ACA\,A\,CC\,ACC\,AA\,CAAC}, \\ \boldsymbol{\phi}_3(\boldsymbol{u}) =& \mathsf{XXX}\,0\;0\,\mathsf{X}\,0\;000\,\mathsf{X}\,0\;000, \end{aligned}$$

and $l_m^{\boldsymbol{u}} = 2$, and $r_m^{\boldsymbol{u}} = 4$.

Remark. A duplication of length m is equivalent to inserting m zeros into $\boldsymbol{\phi}_m(\boldsymbol{u})$. In the above example, $\boldsymbol{u}$ may come from $\boldsymbol{u}' = \mathsf{ACA}\overline{\mathsf{ACC}}\mathsf{AACAAC}$ after a length 3 tandem duplication with the overlined substring as the template and $\boldsymbol{\phi}_3(\boldsymbol{u})$ can be viewed as the result of inserting 3 zeros into $\boldsymbol{\phi}_3(\boldsymbol{u}') = \mathsf{XXX}00\mathsf{X}\overline{0\mathsf{X}}0000$ between the two overlined symbols.

To enable us to succinctly represent the results, we then define several functions. These functions relate $\boldsymbol{u}$ to the frequencies of other substrings that can generate $\boldsymbol{u}$ via appropriate duplication events. For example, consider the sequence $\boldsymbol{u} = \mathsf{ACACAGAG}$, for which $\boldsymbol{\phi}_2(\boldsymbol{u}) = \mathsf{XX}000\mathsf{X}00$. This sequence can be created through duplications of length 2 from ACAGAG (in two ways) and from ACACAG. These correspond to runs of 0 of length 2 in $\boldsymbol{\phi}_2(\boldsymbol{u})$.

Definition 2. For a sequence $\boldsymbol{u}$ and positive integers m, z with $m + z \leqslant |\boldsymbol{u}| + 1$, define

$$D_{z,m}(\boldsymbol{u}) = \boldsymbol{u}_{1,z-1}\boldsymbol{u}_{z+m,|\boldsymbol{u}|+1-z-m},$$

the sequence obtained from $\boldsymbol{u}$ by removing the subsequence $\boldsymbol{u}_{z,m}$, i.e., by removing symbols in positions $z, \ldots, z+m-1$.

Example. For $\boldsymbol{u} = \mathsf{ACGTA}, z = 3, m = 2$, we have $\boldsymbol{u}_{3,2} = \mathsf{GT}$ and $D_{3,2}(\mathsf{ACGTA}) = \mathsf{ACA}$.

Definition 3. For a string $\boldsymbol{u}$ and positive integers m, z with $m + z \leqslant |\boldsymbol{u}| + 1$, define

$$G_m^{\boldsymbol{u}}(\boldsymbol{x}) = \sum_z x^{D_{z,m}(\boldsymbol{u})}, \tag{3}$$

where the sum is over all z that are the indices of the start of (not necessarily maximal) runs of 0s in $\boldsymbol{\phi}_m(\boldsymbol{u})$, i.e., $(\boldsymbol{\phi}_m(\boldsymbol{u}))_{z,m} = 0^m$.

Example. For $\boldsymbol{u} = \mathsf{GACCACCA}, m = 3$, we have $\boldsymbol{\phi}_3(\boldsymbol{u}) = \mathsf{XXXX}0000$ and $(\boldsymbol{\phi}_3(\boldsymbol{u}))_{5,3} = (\boldsymbol{\phi}_3(\boldsymbol{u}))_{6,3} = 0^3$. Therefore $G_3^{\boldsymbol{u}}(\boldsymbol{x}) = 2x^{\mathsf{GACCA}}$.

There is a slight abuse of notation in the definition of G above (as well as the definitions of F and M below). While the argument of G is $\boldsymbol{x} = (x^{\boldsymbol{v}})_{\boldsymbol{v}\in\mathcal{A}^k}$, on the right side of (3), $x^{\boldsymbol{w}}$

for sequences $\boldsymbol{w}$ with $|\boldsymbol{w}| < k$ may appear. We note however that $x^{\boldsymbol{w}}$ can be obtained from $\boldsymbol{x}$ by summing over the elements of $\boldsymbol{x}$ corresponding to strings that include $\boldsymbol{w}$ as a prefix.

New occurrences of $\boldsymbol{u}$ can also be generated from strings that are not of the form $D_{z,m}(\boldsymbol{u})$. For example, consider the sequence $\boldsymbol{u} = \mathsf{ACGACTG}$, for which $\boldsymbol{\phi}_3(\boldsymbol{u}) = \mathsf{XXX00XX}$. This sequence can be created through a length-3 tandem duplication from $\overline{\mathsf{CGA}}\mathsf{CTG}$ and $\overline{\mathsf{GAC}}\mathsf{TG}$, where the part that is to be duplicated is overlined. The following definitions will be of use in the analysis of this type of duplication.

Definition 4. For a sequence $\boldsymbol{u}$ and a positive interger m, define

$$F^{\boldsymbol{u}}_{m,l}(\boldsymbol{x}) = \sum_{i=1}^{\min(l^{\boldsymbol{u}}_m, m-1)} x^{\boldsymbol{u}_{i+1,|\boldsymbol{u}|-i}},$$
$$F^{\boldsymbol{u}}_{m,r}(\boldsymbol{x}) = \sum_{i=1}^{\min(r^{\boldsymbol{u}}_m, m-1)} x^{\boldsymbol{u}_{1,|\boldsymbol{u}|-i}}.$$

In the special case where $\boldsymbol{\phi}_m(\boldsymbol{u}) = \mathsf{X}^m 0^{|\boldsymbol{u}|-m}$ and $|\boldsymbol{u}| \leqslant 2m-2$, we will benefit from the following definition.

Definition 5. For a sequence $\boldsymbol{u}$ and a positive integer m s.t. $\boldsymbol{\phi}_m(\boldsymbol{u}) = \mathsf{X}^m 0^{|\boldsymbol{u}|-m}$ and $|\boldsymbol{u}| \leqslant 2m-2$, define

$$M^{\boldsymbol{u}}_m(\boldsymbol{x}) = \sum_{b=|\boldsymbol{u}|-m+1}^{m-1} x^{\boldsymbol{u}_{b+1,m-b}\boldsymbol{u}_{1,b}}.$$

We define $M^{\boldsymbol{u}}_m(\boldsymbol{x}) = 0$ if $\boldsymbol{\phi}_m(\boldsymbol{u}) \neq \mathsf{X}^m 0^{|\boldsymbol{u}|-m}$.

Definition 6. For a sequence $\boldsymbol{u}$, define $\mathcal{B}_1(\boldsymbol{u})$ to be the set of sequences which are at Hamming distance 1 from $\boldsymbol{u}$.

Example. For $\mathcal{A} = \{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}$, $\boldsymbol{u} = \mathsf{AC}$, we have $\mathcal{B}_1(\boldsymbol{u}) = \{\mathsf{GC},\mathsf{CC},\mathsf{TC},\mathsf{AA},\mathsf{AG},\mathsf{AT}\}$.

Definition 7. For $\boldsymbol{u},\boldsymbol{v} \in \mathcal{A}^*$, define the indicator function $I(\boldsymbol{u},\boldsymbol{v})$ as following,

$$I(\boldsymbol{u},\boldsymbol{v}) = \begin{cases} 1, & \text{if } \boldsymbol{u} = \boldsymbol{v} \\ 0, & \text{otherwise} \end{cases}.$$

B. Evolution of k-mer Frequencies

We first find $\boldsymbol{\delta}_\ell(\boldsymbol{x}) = \left(\delta^{\boldsymbol{u}}_\ell(\boldsymbol{x})\right)_{\boldsymbol{u}\in U}$ for $\ell > 0$ (duplication) and then for $\ell = 0$ (substitution). When analyzing $\delta^{\boldsymbol{u}}_\ell(\boldsymbol{x})$, we only consider substrings $\boldsymbol{u}$ of length $|\boldsymbol{u}| > \ell$, which simplifies the derivation. The frequencies of shorter substrings can be found by summing over the frequencies of longer substrings.

We first analyze the case in which $\ell > 0$. We present three lemmas and then use them to prove a general form for $\boldsymbol{\delta}_\ell(\boldsymbol{x}), \ell > 0$. Suppose a duplication of length ℓ occurs in $\boldsymbol{s}_n$, resulting in $\boldsymbol{s}_{n+1}$. The number of occurrences of $\boldsymbol{u}$ may change due to the duplication event. To study this change, we consider the k-substrings of $\boldsymbol{s}_n$ that are eliminated (do not exist in $\boldsymbol{s}_{n+1}$) and the k-substrings of $\boldsymbol{s}_{n+1}$ that are new (do not exist in $\boldsymbol{s}_n$). Any new k-substring must intersect with both the template and the copy in $\boldsymbol{s}_{n+1}$. Likewise, an eliminated k-substring must include symbols on both sides of the template in $\boldsymbol{s}_n$, i.e., the template must be a strict substring of the k-substring that includes neither its leftmost symbol nor its rightmost symbol. As an example, suppose

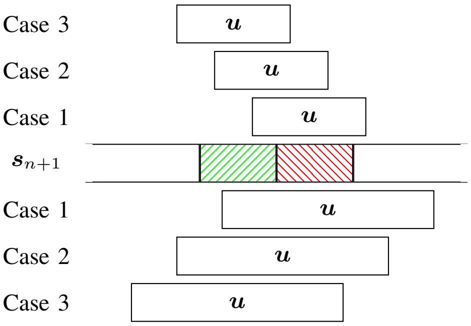

Figure 1. Possible cases for new occurrences of $\boldsymbol{u}$ in $\boldsymbol{s}_{n+1}$. Cases above and below $\boldsymbol{s}_{n+1}$ correspond to $\ell+1 \leqslant k < 2\ell$ and $k \geqslant 2\ell$, respectively. The hatched boxes, from left to right, are the template and the copy.

$$\boldsymbol{s}_n = \boldsymbol{v}\mathsf{ACGTAGATw}, \quad \boldsymbol{s}_{n+1} = \boldsymbol{v}\mathsf{ACG}\overline{\mathsf{TAG}}\underline{\mathsf{TAG}}\mathsf{AT}\boldsymbol{w}, \quad (4)$$

where $\ell = 3$, the (new) copy is underlined and the template is overlined, and $\boldsymbol{v},\boldsymbol{w} \in \mathcal{A}^*$. Let $k = 5$, the new 5-substrings are GTAGT, TAGTA, AGTAG, GTAGA and the eliminated substring is GTAGA. Note that here the two GTAGA substrings are counted as different. Formally, let

$$\boldsymbol{s}_n = a_1 \cdots a_i a_{i+1} \cdots a_{i+\ell} a_{i+\ell+1} \cdots a_{|s_n|},$$
$$\boldsymbol{s}_{n+1} = a_1 \cdots a_i a_{i+1} \ldots a_{i+\ell} a_{i+1} \ldots a_{i+\ell} a_{i+\ell+1} \ldots a_{|s_n|},$$

where the substring $a_{i+1} \cdots a_{i+\ell}$ is duplicated. The new k-substrings created in $\boldsymbol{s}_{n+1}$ are

$$\boldsymbol{y}_b = a_{i+\ell+1-b} a_{i+\ell+2-b} \ldots a_{i+\ell} a_{i+1} a_{i+2} \ldots a_{i+k-b},$$

for $1 \leqslant b \leqslant k-1$. Note that we have defined $\boldsymbol{y}_b$ such that the first element of the copy, a_{i+1}, is at position $b+1$ in $\boldsymbol{y}_b$. The k-substrings eliminated from $\boldsymbol{s}_n$ are $a_{i-c+1} \cdots a_{i+k-c}$, for $1 \leqslant c \leqslant k-\ell-1$.

For a given $\boldsymbol{u}$, let Y_b denote the indicator random variable for the event that $\boldsymbol{y}_b = \boldsymbol{u}$, that is, the duplication creates a new occurrence of $\boldsymbol{u}$ in $\boldsymbol{s}_{n+1}$ in which the first symbol of the copy is in position $b+1$. In example denoted by (4), if $\boldsymbol{u} = \mathsf{TAGTA}$, then $\boldsymbol{y}_3 = \boldsymbol{u}$ and thus $Y_3 = 1$.

Furthermore, let W denote the number of occurrences of $\boldsymbol{u}$ that are eliminated. We have

$$\delta^{\boldsymbol{u}}_\ell(\boldsymbol{x}) = \Big(\sum_{b=1}^{k-1} \mathbb{E}_\ell[Y_b|\mathcal{F}_n]\Big) - \mathbb{E}_\ell[W|\mathcal{F}_n]$$
$$= \Big(\sum_{b=1}^{k-1} \mathbb{E}_\ell[Y_b|\mathcal{F}_n]\Big) - (k-\ell-1)x^{\boldsymbol{u}}, \quad (5)$$

where the second equality follows from the fact that each of the $k-\ell-1$ eliminated k-substrings are equal to $\boldsymbol{u}$ with probability $x^{\boldsymbol{u}}$.

To find $\delta^{\boldsymbol{u}}_\ell$, it suffices to find $\mathbb{E}_\ell[Y_b|\mathcal{F}_n]$, or equivalently, $\Pr(Y_b = 1|\mathcal{F}_n, \ell)$. We consider different cases based on the value of b, which determines how $\boldsymbol{u}$ overlaps with the template and the copy. These cases are illustrated in Figure 1 and are considered in Lemmas 5–7, whose proofs are given in the Appendix.

Lemma 5 (Case 1)**.** *For* $1 \leqslant b < \min(\ell, k-\ell+1)$,

$$\mathbb{E}_\ell[Y_b|\mathcal{F}_n] = x^{\boldsymbol{u}_{b+1,k-b}} I(\boldsymbol{u}_{1,b}, \boldsymbol{u}_{1+\ell,b}).$$

Lemma 6 (Case 2)**.** *Suppose* $\min(\ell, k-\ell+1) \leqslant b < \max(k-\ell+1, \ell)$*. If* $k \geqslant 2\ell$*, then*

$$\mathbb{E}_\ell[Y_b|\mathcal{F}_n] = x^{\boldsymbol{u}_{1,b-\ell}\boldsymbol{u}_{b+1,k-b}} I(\boldsymbol{u}_{b-\ell+1,\ell}, \boldsymbol{u}_{b+1,\ell}),$$

and if $\ell+1 \leqslant k \leqslant 2\ell-2$*, then*

$$\mathbb{E}_\ell[Y_b|\mathcal{F}_n] = x^{\boldsymbol{u}_{b+1,\ell-b}\boldsymbol{u}_{1,b}} I(\boldsymbol{u}_{1,k-\ell}, \boldsymbol{u}_{\ell+1,k-\ell}).$$

Lemma 7 (Case 3)**.** *For* $\max(k-\ell+1, \ell) \leqslant b \leqslant k-1$,

$$\mathbb{E}_\ell[Y_b|\mathcal{F}_n] = x^{\boldsymbol{u}_{1,b}} I(\boldsymbol{u}_{b-\ell+1,k-b}, \boldsymbol{u}_{b+1,k-b}).$$

Based on Lemmas 5–7, we then prove the following Theorem. We will use the three lemmas above to break the summation of (5) into three parts and then simplify them to get a generalized expression.

Theorem 8. *For an integer* $\ell > 0$ *and a string* $\boldsymbol{u} = u_1 u_2 \cdots u_k$*, if* $\ell+1 \leqslant k < 2\ell$*, then*

$$\delta_\ell^{\boldsymbol{u}}(\boldsymbol{x}) = F_{\ell,l}^{\boldsymbol{u}}(\boldsymbol{x}) + F_{\ell,r}^{\boldsymbol{u}}(\boldsymbol{x}) + M_\ell^{\boldsymbol{u}}(\boldsymbol{x}) - (k-1-\ell)x^{\boldsymbol{u}},$$

and if $k \geqslant 2\ell$*,*

$$\delta_\ell^{\boldsymbol{u}}(\boldsymbol{x}) = F_{\ell,l}^{\boldsymbol{u}}(\boldsymbol{x}) + F_{\ell,r}^{\boldsymbol{u}}(\boldsymbol{x}) + G_\ell^{\boldsymbol{u}}(\boldsymbol{x}) - (k-1-\ell)x^{\boldsymbol{u}}. \quad (6)$$

Proof: From (5), we can write

$$\begin{aligned}\delta_\ell^{\boldsymbol{u}}(\boldsymbol{x}) &= \Big(\sum_{b=1}^{k-1} \mathbb{E}_\ell[Y_b|\mathcal{F}_n]\Big) - (k-\ell-1)x^{\boldsymbol{u}} \\ &= \sum_{b=1}^{\min(\ell-1,k-\ell)} \mathbb{E}_\ell[Y_b|\mathcal{F}_n] \quad + \sum_{b=\min(\ell,k-\ell+1)}^{\max(k-\ell,\ell-1)} \mathbb{E}_\ell[Y_b|\mathcal{F}_n] \\ &\quad + \sum_{b=\max(k-\ell+1,\ell)}^{k-1} \mathbb{E}_\ell[Y_b|\mathcal{F}_n] - (k-\ell-1)x^{\boldsymbol{u}}. \quad (7)\end{aligned}$$

By Lemma 5, we have

$$\begin{aligned}\sum_{b=1}^{\min(\ell-1,k-\ell)} \mathbb{E}_\ell[Y_b|\mathcal{F}_n] &= \sum_{b=1}^{\min(\ell-1,k-\ell)} x^{\boldsymbol{u}_{b+1,k-b}} I(\boldsymbol{u}_{1,b}, \boldsymbol{u}_{1+\ell,b}) \\ &= \sum_{b=1}^{\min(\ell-1,k-\ell)} x^{\boldsymbol{u}_{b+1,k-b}} I(\boldsymbol{\phi}_\ell(\boldsymbol{u})_{\ell+1,b}, 0^b) \\ &= \sum_{b=1}^{\min(\ell-1,k-\ell,l_\ell^{\boldsymbol{u}})} x^{\boldsymbol{u}_{b+1,k-b}} \\ &= \sum_{b=1}^{\min(\ell-1,l_\ell^{\boldsymbol{u}})} x^{\boldsymbol{u}_{b+1,k-b}} \\ &= F_{\ell,l}^{\boldsymbol{u}}(\boldsymbol{x}), \quad (8)\end{aligned}$$

where the fourth equality follows from the fact that $l_\ell^{\boldsymbol{u}} \leqslant k-\ell$.

Similarly, using Lemma 7, it can be shown that

$$\begin{aligned}&\sum_{b=\max(k-\ell+1,\ell)}^{k-1} \mathbb{E}_\ell[Y_b|\mathcal{F}_n] \\ &= \sum_{b=\max(k-\ell+1,\ell)}^{k-1} x^{\boldsymbol{u}_{1,b}} I(\boldsymbol{u}_{b-\ell+1,k-b}, \boldsymbol{u}_{b+1,k-b}) \\ &= \sum_{b=\max(k-\ell+1,\ell)}^{k-1} x^{\boldsymbol{u}_{1,b}} I(\boldsymbol{\phi}_\ell(\boldsymbol{u})_{b+1,k-b}, 0^{k-b}) \\ &= \sum_{b=\max(k-\ell+1,\ell,k-r_\ell^{\boldsymbol{u}})}^{k-1} x^{\boldsymbol{u}_{1,b}} \\ &= \sum_{b=\max(k-\ell+1,k-r_\ell^{\boldsymbol{u}})}^{k-1} x^{\boldsymbol{u}_{1,b}} \\ &= \sum_{i=1}^{\min(r_\ell^{\boldsymbol{u}},\ell-1)} x^{\boldsymbol{u}_{1,k-i}} \\ &= F_{\ell,r}^{\boldsymbol{u}}(\boldsymbol{x}), \quad (9)\end{aligned}$$

where the fourth equality follows from $r_\ell^{\boldsymbol{u}} \leqslant k-\ell$ and the fifth equality comes from setting $i = k-b$.

To complete the proof, we need to show that $\mathbb{E}_\ell[Y_b|\mathcal{F}_n]$ summed over the range $\min(\ell, k-\ell+1) \leqslant b \leqslant \max(k-\ell, \ell-1)$ reduces to $G_\ell^{\boldsymbol{u}}(\boldsymbol{x})$ or $M_\ell^{\boldsymbol{u}}(\boldsymbol{x})$ as appropriate.

From Lemma 6, if $\ell+1 \leqslant k \leqslant 2\ell-2$, then

$$\begin{aligned}\sum_{b=\min(\ell,k-\ell+1)}^{\max(k-\ell,\ell-1)} \mathbb{E}_\ell[Y_b|\mathcal{F}_n] &= \sum_{b=k-\ell+1}^{\ell-1} \mathbb{E}_\ell[Y_b|\mathcal{F}_n] \\ &= \sum_{b=k-\ell+1}^{\ell-1} x^{\boldsymbol{u}_{b+1,\ell-b}\boldsymbol{u}_{1,b}} I(\boldsymbol{u}_{1,k-\ell}, \boldsymbol{u}_{\ell+1,k-\ell}) \\ &= \sum_{b=k-\ell+1}^{\ell-1} x^{\boldsymbol{u}_{b+1,\ell-b}\boldsymbol{u}_{1,b}} I(\boldsymbol{\phi}_\ell(\boldsymbol{u})_{\ell+1,k-\ell}, 0^{k-\ell}) \\ &= M_\ell^{\boldsymbol{u}}(\boldsymbol{x}), \quad (10)\end{aligned}$$

and if $k = 2\ell-1$, also

$$\sum_{b=\min(\ell,k-\ell+1)}^{\max(k-\ell,\ell-1)} \mathbb{E}_\ell[Y_b|\mathcal{F}_n] = 0 = M_\ell^{\boldsymbol{u}}(\boldsymbol{x}). \quad (11)$$

Finally, if $k \geqslant 2\ell$, from the same lemma, we find

$$\begin{aligned}\sum_{b=\min(\ell,k-\ell+1)}^{\max(k-\ell,\ell-1)} \mathbb{E}_\ell[Y_b|\mathcal{F}_n] &= \sum_{b=\ell}^{k-\ell} \mathbb{E}_\ell[Y_b|\mathcal{F}] \\ &= \sum_{b=\ell}^{k-\ell} x^{\boldsymbol{u}_{1,b-\ell}\boldsymbol{u}_{b+1,k-b}} I(\boldsymbol{u}_{b-\ell+1,\ell}, \boldsymbol{u}_{b+1,\ell}) \\ &= \sum_{b=\ell}^{k-\ell} x^{\boldsymbol{u}_{1,b-\ell}\boldsymbol{u}_{b+1,k-b}} I(\boldsymbol{\phi}_\ell(\boldsymbol{u})_{b+1,\ell}, 0^\ell) = G_k(\boldsymbol{u}), \quad (12)\end{aligned}$$

where the last step follows from the definition of G_k.

Summing over the expressions provided by (8)-(12) provides the desired result. ■

In the case of $\ell = 0$, $\boldsymbol{\delta}_\ell(\boldsymbol{x})$ is given by the following theorem.

Theorem 9. *For a string $\boldsymbol{u}$ of length k, we have*

$$\delta_0^{\boldsymbol{u}}(\boldsymbol{x}) = \frac{1}{|\mathcal{A}|-1}\sum_{\boldsymbol{v}\in\mathcal{B}_1(\boldsymbol{u})} x^{\boldsymbol{v}} - kx^{\boldsymbol{u}}. \tag{13}$$

Before proving the theorem, we give an example for $\mathcal{A} = \{1,2,3\}$:

$$\delta_0^{123}(\boldsymbol{x}) = \frac{1}{2}(x^{223}+x^{323}+x^{113}+x^{133}+x^{121}+x^{122}) - 3x^{123}$$

Proof: A new occurrence of $\boldsymbol{u}$ results from an appropriate substitution in some $\boldsymbol{v} \in \mathcal{B}_1(\boldsymbol{u})$, which has probability $x^{\boldsymbol{v}}/(|\mathcal{A}|-1)$. On the other hand, an occurrence of $\boldsymbol{u}$ is eliminated if a substitution occurs in any of its k positions. So the expected number occurrences that vanish is $kx^{\boldsymbol{u}}$. ■

C. ODE and the Limits of Substring Frequencies

Theorems 8 and 9 provide expressions for $\boldsymbol{\delta}_\ell(\boldsymbol{x})$ for $0 \leqslant \ell \leqslant M-1$. With these results in hand, we can formulate an ordinary differential equation (ODE) whose limits are the same as those of the substring frequencies of interest, $\boldsymbol{x} = (x^{\boldsymbol{u}})_{\boldsymbol{u}\in\mathcal{A}^k}$, where $k \geqslant M$.

We first show that $\delta_\ell^{\boldsymbol{u}}(\boldsymbol{x})$ can be written as a linear combination of the elements of $\boldsymbol{x}$, i.e., a linear combination of $x^{\boldsymbol{v}}, \boldsymbol{v} \in \mathcal{A}^k$. To see this, note that on the right side in expressions for $\delta_\ell^{\boldsymbol{u}}$ in Theorems 8 and 9, terms of the form $x^{\boldsymbol{w}}$ appear where $|\boldsymbol{w}| \leqslant k$. We can replace $x^{\boldsymbol{w}}$ with $\sum_{\boldsymbol{v}} x^{\boldsymbol{v}}$, where the summation is over all strings $\boldsymbol{v}$ of length k such that $\boldsymbol{w}$ is a prefix of $\boldsymbol{v}$. For example, consider the alphabet $\{1,2,3\}$ and $k=3$. From Theorem 8, we have

$$\begin{aligned}\delta_2^{121}(\boldsymbol{x}) &= x^{12} + x^{21}\\ &= x^{121}+x^{122}+x^{123}+x^{211}+x^{212}+x^{213}.\end{aligned}$$

For $0 \leqslant \ell < M$, let A_ℓ be the matrix satisfying $\boldsymbol{\delta}_\ell(\boldsymbol{x}) - \ell\boldsymbol{x} = A_\ell\boldsymbol{x}$. Based on the argument above, such a matrix exists and can be computed from Theorems 8 and 9. Furthermore, let

$$A = \sum_{\ell=0}^{M-1} q_\ell A_\ell. \tag{14}$$

Note that $\boldsymbol{h}_\ell(\boldsymbol{x}) = A_\ell\boldsymbol{x}$ and $\boldsymbol{h}(\boldsymbol{x}) = \sum_\ell q_\ell \boldsymbol{h}_\ell(\boldsymbol{x}) = A\boldsymbol{x}$.

For example, consider $q_0 = \alpha$, $q_1 = 1-\alpha$, $\mathcal{A} = \{0,1\}$, and $\boldsymbol{x} = (x^{00}, x^{01}, x^{10}, x^{11})$. From Theorems 8 and 9, it can be shown that

$$A_0 = \begin{pmatrix} -2 & 1 & 1 & 0\\ 1 & -2 & 0 & 1\\ 1 & 0 & -2 & 1\\ 0 & 1 & 1 & -2\end{pmatrix}, \quad A_1 = \begin{pmatrix} 0 & 1 & 0 & 0\\ 0 & -1 & 0 & 0\\ 0 & 0 & -1 & 0\\ 0 & 0 & 1 & 0\end{pmatrix}.$$

and

$$A = \begin{pmatrix} -2\alpha & 1 & \alpha & 0\\ \alpha & -(1+\alpha) & 0 & \alpha\\ \alpha & 0 & -(1+\alpha) & \alpha\\ 0 & \alpha & 1 & -2\alpha\end{pmatrix}. \tag{15}$$

Theorem 10. *Consider a tandem duplication and substitution system with distribution $q = (q_\ell)_{0\leqslant\ell<M}$ over these mutations, with $q_0 < 1$, and let A be the matrix defined for this system by* (14). *The frequencies of substrings $\boldsymbol{u}$ of length $k \geqslant M$, $(x^{\boldsymbol{u}})_{\boldsymbol{u}\in\mathcal{A}^k}$, converge almost surely to the null space of the matrix A.*

Proof: We first show that the resulting ODE is stable by showing that every eigenvalue of matrix A is either 0 or has a negative real part. This is done by applying the Gershgorin circle theorem [51] to the columns of A (see e.g., (15)). According to the Gershgorin circle theorem, every eigenvalue of A lies within at least one of the closed discs $D_1, \ldots, D_{|\mathcal{A}|^k}$ in the complex plain, where the i-th disc centers at the i-th diagonal entry of A with radius equal to the sum of the absolute values of the non-diagonal entries in the i-th column. Since in each column, the diagonal element is the only element that can be negative, it suffices to show that each column of A sums to 0, which then implies that the rightmost point of each circle is the origin. Thus, each eigenvalue of A is either 0 or has a negative real part.

We now show that each column of A_ℓ sums to zero for any ℓ. Fix $\boldsymbol{v} \in U$ and consider the column in A_ℓ that corresponds to $x^{\boldsymbol{v}}$. We denote this column by $A_\ell^{\boldsymbol{v}}$ for simplicity. To identify the element in $A_\ell^{\boldsymbol{v}}$ that corresponds to $\boldsymbol{u}$ (i.e., the element in A in the column corresponding to $\boldsymbol{v}$ and the row corresponding to $\boldsymbol{u}$), we must consider expressions for $h_\ell^{\boldsymbol{u}}(\boldsymbol{x}) = \delta_\ell^{\boldsymbol{u}}(\boldsymbol{x}) - \ell x^{\boldsymbol{u}}$ and check if $x^{\boldsymbol{v}}$ appears on the right side. The coefficient of $x^{\boldsymbol{v}}$ in $\delta_\ell^{\boldsymbol{u}}(\boldsymbol{x}) - \ell x^{\boldsymbol{u}}$ is exactly the entry of $A_\ell^{\boldsymbol{v}}$ in the row corresponding to $x^{\boldsymbol{u}}$. For $\ell > 0$, from (5) and Lemmas 5–7, we can see that the only term with a negative coefficient is $-(k-1)x^{\boldsymbol{u}}$, and the terms with nonnegative coefficients are $\sum_{b=1}^{k-1}\mathbb{E}_\ell[Y_b|\mathcal{F}_n]$. Therefore the case in which $x^{\boldsymbol{v}}$ appears in $\delta_\ell^{\boldsymbol{u}}(\boldsymbol{x}) - \ell x^{\boldsymbol{u}}$ with negative coefficient happens only when $\boldsymbol{u} = \boldsymbol{v}$, which implies that $A_\ell^{\boldsymbol{v}}$ has exactly one negative entry, which equals $-(k-1)$. Then we study the case in which $x^{\boldsymbol{v}}$ appears in $\delta_\ell^{\boldsymbol{u}}(\boldsymbol{x}) - \ell x^{\boldsymbol{u}}$ with a nonnegative coefficient. By Lemmas 5–7, this happens if and only if $x^{\boldsymbol{v}} = \mathbb{E}_\ell[Y_b|\mathcal{F}_n]$ for some $1 \leqslant b \leqslant k-1$. Note that $\mathbb{E}_\ell[Y_b|\mathcal{F}_n]$ has different forms when b has different values. Inspecting the proofs of Lemmas 5–7 shows that for each value of $b \in [k-1]$, there is precisely one $\boldsymbol{u}$ such that $x^{\boldsymbol{v}} = \mathbb{E}_\ell[Y_b|\mathcal{F}_n]$. Hence, for each $b \in [k-1]$, $x^{\boldsymbol{v}}$ appears in $h_\ell^{\boldsymbol{u}}$ with a nonnegative coefficient, and the coefficient is 1. For example, for $b = 1$, from Lemma 5, this $\boldsymbol{u}$ is equal to $\boldsymbol{v}_\ell\boldsymbol{v}_{1,k-1}$. Since there are $k-1$ possible choices for b, the sum of all nonnegative coefficients is $k-1$, which is also the sum of all nonnegative entries in $A_\ell^{\boldsymbol{v}}$. Therefore the sum of all entries in $A_\ell^{\boldsymbol{v}}$, and thus every column in A_ℓ, is 0, as desired. For $\ell = 0$, we have $h_\ell^{\boldsymbol{u}}(\boldsymbol{x}) = \delta_\ell^{\boldsymbol{u}}(\boldsymbol{x})$, where $\delta_\ell^{\boldsymbol{u}}(\boldsymbol{x})$ is given in Theorem 9. The column corresponding to $x^{\boldsymbol{v}}$ has a negative term equal to $-k$ and $k(|\mathcal{A}|-1)$ positive terms, where each of the positive terms is equal to $\frac{1}{|\mathcal{A}|-1}$, so the sum is again 0.

We have shown that all eigenvalues are either 0 or have negative real parts. For any valid initial point $\boldsymbol{x}_0$, the sum of the elements must be 1. Furthermore, each element must be nonnegative. The fact that the columns of A sum to 0 shows that the sum of the elements of any solution $\boldsymbol{x}_t$ also equals 1 as $d\boldsymbol{x}_t/dt = A\boldsymbol{x}_t$. Furthermore, since only diagonal terms in A can be negative, each element of $\boldsymbol{x}_t$ is also nonnegative. Thus $\boldsymbol{x}_t$ is bounded.

By the Jordan canonical from theorem [38], any square matrix over $\mathbb{C}$ can be decomposed into the form QBQ^{-1} for some invertible matrix Q. Here

$$B = \begin{pmatrix} B_1 & & & \\ & B_2 & & \\ & & \ddots & \\ & & & B_m \end{pmatrix}$$

is a block diagonal matrix consisting of Jordan blocks, and the Jordan blocks have the form

$$B_i = \begin{pmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{pmatrix}, \text{ for all } i,$$

where λ_i is one of the eigenvalues of the original matrix. So we can write $A = PJP^{-1}$ for some invertible matrix P, where $J = \begin{pmatrix} J' & 0 \\ 0 & J'' \end{pmatrix}$ and J' and J'' are square matrices corresponding to the eigenvalue $\lambda = 0$ and other eigenvalues respectively. Let $\boldsymbol{y}_t = P^{-1}\boldsymbol{x}_t$, so that $\dot{\boldsymbol{y}}_t = J\boldsymbol{y}_t$, which we can write in the form $\dot{\boldsymbol{u}}_t = J'\boldsymbol{u}_t$ and $\dot{\boldsymbol{w}}_t = J''\boldsymbol{w}_t$ with $\boldsymbol{y}_t = (\boldsymbol{u}_t, \boldsymbol{w}_t)^T$. Let C be any compact internally chain transitive set of the ODE $\dot{\boldsymbol{y}}_t = J\boldsymbol{y}_t$. We first show that if $\boldsymbol{y} = (\boldsymbol{u}, \boldsymbol{w}) \in C$, then $\boldsymbol{w} = 0$. Consider the flow starting from $\boldsymbol{y}_0 = (\boldsymbol{u}_0, \boldsymbol{w}_0)^T \in C$ with $\boldsymbol{w}_0 \neq \boldsymbol{0}$. We have $\boldsymbol{w}_t = e^{J''}\boldsymbol{w}_0$. Since J'' has only eigenvalues with negative real parts, $\|\boldsymbol{w}_t\| \leqslant c_0 e^{-c_1 t}\|\boldsymbol{w}_0\|$ for $t \geqslant 0$ and some constants $c_0, c_1 > 0$. If $\boldsymbol{y} = (\boldsymbol{u}, \boldsymbol{w}) \in C$, then $\boldsymbol{w}$ is also in an internally chain transitive set of lower dimension. For $T, \epsilon > 0$, let $\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(n)} = \boldsymbol{w}^{(1)}$ be a chain of points such that the flow of $\dot{\boldsymbol{w}}_t = J''\boldsymbol{w}_t$ starting at $\boldsymbol{w}^{(i)}$ meets the ϵ-neighborhood of $\boldsymbol{w}^{(i+1)}$ after a time $\geqslant T$. We thus have

$$\|\boldsymbol{w}^{(i+1)}\| \leqslant c_0 e^{-c_1 T}\|\boldsymbol{w}^{(i)}\| + \epsilon. \tag{16}$$

Since T, ϵ are arbitary, we choose them such that $c_0 e^{-c_1 T} < 1/2$ and $c_0 e^{-c_1 T}\|\boldsymbol{w}^{(1)}\| < \epsilon < \|\boldsymbol{w}^{(1)}\|/2$ if $\|\boldsymbol{w}^{(1)}\| > 0$. Hence, $\|\boldsymbol{w}^{(2)}\| \leqslant c_0 e^{-c_1 T}\|\boldsymbol{w}^{(i)}\| + \epsilon < 2\epsilon$ and by induction $\|\boldsymbol{w}^{(i+1)}\| \leqslant c_0 e^{-c_1 T}\|\boldsymbol{w}^{(i)}\| + \epsilon < 2\epsilon$ for $i > 1$. This leads to a contraction since it implies that $\|\boldsymbol{w}^{(n)}\| = \|\boldsymbol{w}^{(1)}\| < 2\epsilon$. Thus $\|\boldsymbol{w}^{(1)}\| = 0$ and for any $\boldsymbol{y} = (\boldsymbol{u}, \boldsymbol{w})^T \in C$ we must have $\boldsymbol{w} = \boldsymbol{0}$.

Next, note that since $\boldsymbol{x}_t$ is bounded, so is $\boldsymbol{y}_t$. Hence for $\boldsymbol{y} = (\boldsymbol{u}, \boldsymbol{0})^T \in C$, $e^{J't}\boldsymbol{u}$ must be a constant since it contains no exponential terms ($\lambda = 0$) and cannot contain a polynomial term in t with degree $\geqslant 1$ (because of boundedness). So all flows initiated in C are constant. The same must hold for all flows in D, for any D that is an internally chain transitive invariant set of the ODE $\dot{\boldsymbol{x}}_t = A\boldsymbol{x}_t$. Hence, any point in $\boldsymbol{x} \in D$ must be in the null space of A, that is, $A\boldsymbol{x} = 0$. ■

For the matrix A of (15), for $0 < \alpha < 1$, the vector in the null space whose elements sum to 1, and thus the limit of $\boldsymbol{x}_n$, is

$$\frac{1}{2(1+3\alpha)}(\alpha+1, 2\alpha, 2\alpha, \alpha+1)^T. \tag{17}$$

If we let $\alpha = \frac{1}{4}$ as an example, the limit of $\boldsymbol{x}_n$ then is

$$\lim_{n\to\infty}(x_n^{00}, x_n^{01}, x_n^{10}, x_n^{11})^T = (\frac{5}{14}, \frac{1}{7}, \frac{1}{7}, \frac{5}{14})^T. \tag{18}$$

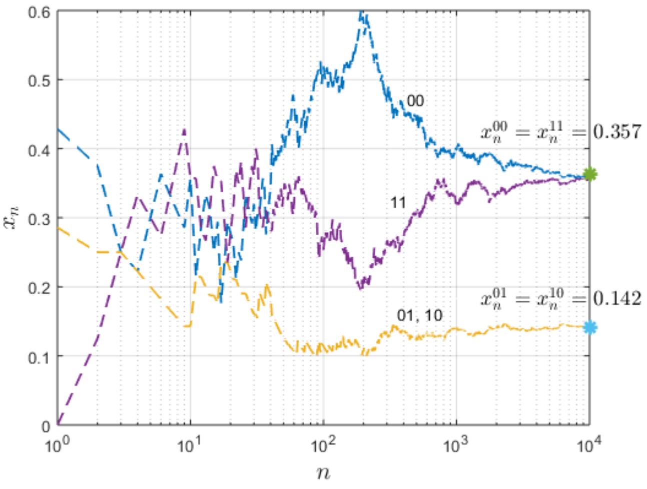

Figure 2. 2-mer frequencies vs the number of mutations in a tandem duplication and substitution system, with $\mathcal{A} = \{0, 1\}, \boldsymbol{s}_0 = 0100010, q_0 = \frac{1}{4}$, and $q_1 = \frac{3}{4}$.

Figure 2 shows the result of simulation of the above TDS system, where $\mathcal{A} = \{0, 1\}$, $\boldsymbol{s}_0 = 0100010, q_0 = \frac{1}{4}$ and $q_1 = \frac{3}{4}$. As the number n of mutations increases, the frequency vector $\boldsymbol{x}_n$ converges to the analytical result (18). Note that the limits do not depend on the initial sequence $\boldsymbol{s}_0$.

Let us consider the two extreme cases. As $\alpha \to 1$, all four 2-substrings become equally likely, each with probability $1/4$. Note however that our analysis is not applicable to $q_0 = \alpha = 1$ since the condition $\sum_n 1/|\boldsymbol{s}_n|^2 < \infty$ is not satisfied. On the other hand, for a small probability of substitution, $0 < \alpha \ll 1$, almost all 2-substrings are either 00 or 11, as expected. For $\alpha = 0$, the null space is spanned by $\boldsymbol{z}_1 = (1, 0, 0, 0)^T$ and $\boldsymbol{z}_2 = (0, 0, 0, 1)^T$ and the limit set is $\{a\boldsymbol{z}_1 + (1-a)\boldsymbol{z}_2 : 0 \leqslant a \leqslant 1\}$. In this case, the asymptotic behavior of k-mer frequencies will depend on the initial sequence $\boldsymbol{s}_0$.

D. Bounds on Entropy

We now turn to provide upper bounds on the entropy. We first formally define the entropy, and then argue that the entropy is upper bounded by the capacity of an appropriately defined semiconstrained system [13]–[15].

Consider the string $\boldsymbol{s}_n$, obtained from $\boldsymbol{s}_0$ by n rounds of mutations, as described previously. Its expected length is $\mathbb{E}[|\boldsymbol{s}_n|] = |\boldsymbol{s}_0| + n\sum_{\ell=1}^{M-1}\ell q_\ell$. We define the entropy after n rounds as

$$\begin{aligned}\mathcal{H}_n &= \frac{1}{\mathbb{E}[|\boldsymbol{s}_n|]} \cdot H(\boldsymbol{s}_n) \\ &= -\frac{1}{\mathbb{E}[|\boldsymbol{s}_n|]}\sum_{\boldsymbol{w}\in\mathcal{A}^*}\Pr(\boldsymbol{s}_n = \boldsymbol{w})\log_{|\mathcal{A}|}\Pr(\boldsymbol{s}_n = \boldsymbol{w}), \end{aligned} \tag{19}$$

and the entropy $\mathcal{H}_\infty = \limsup_{n\to\infty}\mathcal{H}_n$. We note that $H(\boldsymbol{s}_n)$ is the usual entropy of $\boldsymbol{s}_n$ (except for the fact that we use base-$|\mathcal{A}|$ logarithms instead of the usual base-2 logarithms).

It is common to define the entropy of DNA sequences based on the limit of block entropies [23], [34], [47]. Specifically, let $h_k = -\sum_{\boldsymbol{u}\in\mathcal{A}^k} p_{\boldsymbol{u}}\log p_{\boldsymbol{u}}$, where $p_{\boldsymbol{u}}$ is the probability of observing $\boldsymbol{u}$. Entropy is then obtained as $h_{k+1} - h_k$ for $k \to \infty$. This definition may lead to misleading results. For example, consider a string system in which $\boldsymbol{s}_n$ is the De Bruijn sequence of order n (which contains all strings of length n precisely once), obtained according to some deterministic algorithm. Based on block entropies, the entropy of the system can be shown to equal $\log |\mathcal{A}|$, while the system is in fact deterministic. The definition in (19) gives the correct entropy, i.e., 0, since there is only one possibility for $\boldsymbol{s}_n$ for each n.

Let us recall some definitions concerning semiconstrained systems (see [14]). Fix k and let $\mathcal{P}(\mathcal{A}^k)$ denote the set of all probability measures on $\mathcal{A}^k$. A *semiconstrained system* is defined by $\Gamma_k \subseteq \mathcal{P}(\mathcal{A}^k)$. The set of the admissible words of the semiconstrained system, denoted $\mathcal{B}(\Gamma_k)$, contains exactly all finite words over the alphabet $\mathcal{A}$ whose k-mer distribution is in Γ_k. Let $\mathcal{B}_n(\Gamma_k) = \mathcal{B}(\Gamma_k) \cap \mathcal{A}^n$. An expansion of Γ_k by $\epsilon > 0$ is defined as

$$\mathbb{B}_\epsilon(\Gamma_k) = \left\{ \boldsymbol{\xi} \in \mathcal{P}(\mathcal{A}^k) \ : \ \inf_{\boldsymbol{\nu} \in \Gamma_k} \|\boldsymbol{\nu} - \boldsymbol{\xi}\|_{\mathrm{TV}} \leqslant \epsilon \right\},$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total-variation norm. Thus, $\mathbb{B}_\epsilon(\Gamma_k)$ contains all the measures in Γ_k as well as those which are ϵ-close to some measure in Γ_k. The capacity of Γ_k is then defined as

$$\mathsf{cap}(\Gamma_k) = \lim_{\epsilon \to 0^+} \limsup_{n \to \infty} \frac{1}{n} \log_{|\mathcal{A}|} |\mathcal{B}_n(\mathbb{B}_\epsilon(\Gamma_k))|,$$

which intuitively measures the information per symbol in strings whose k-mer distribution is in (or "almost" in) Γ_k.

Theorem 11. *For the mutation process described above, for $k \in \mathbb{N}^+$, if the vector of the frequencies $\boldsymbol{x}$ of strings of length k converges almost surely to a set Γ_k, then $\mathcal{H}_\infty \leqslant \mathsf{cap}(\Gamma_k)$.*

Proof: Fix some positive real number $\epsilon > 0$. Denote by X the indicator random variable defined by

$$X = \begin{cases} 0 & ||\boldsymbol{s}_n| - \mathbb{E}[|\boldsymbol{s}_n|]| \geqslant \epsilon n, \\ 1 & \text{otherwise.} \end{cases}$$

By Hoeffding's inequality,

$$\Pr(X = 0) \leqslant 2\exp\left(-\frac{2\epsilon^2}{(M-1)^2} n\right).$$

We also note that $|\boldsymbol{s}_n| \leqslant |\boldsymbol{s}_0| + (M-1)n$ for all n.

Now, let Y be the indicator random variable defined by

$$Y = \begin{cases} 0 & \boldsymbol{x}_n \notin \mathbb{B}_\epsilon(\Gamma_k), \\ 1 & \text{otherwise.} \end{cases}$$

We know that $\boldsymbol{x}_n$ converges almost surely to some point in Γ_k as $n \to \infty$, and thus, there exists $N(\epsilon)$ such that for all $n \geqslant N(\epsilon)$,

$$\Pr(Y = 0) \leqslant \epsilon.$$

We combine X and Y by defining the indicator random variable,

$$Z = X \cdot Y.$$

By the union bound,

$$\Pr(Z = 0) \leqslant \epsilon + 2\exp\left(-\frac{2\epsilon^2}{(M-1)^2} n\right). \qquad (20)$$

Using standard bounds on the joint entropy and conditional entropy,

$$H(\boldsymbol{s}_n) \leqslant H(\boldsymbol{s}_n, Z) = H(\boldsymbol{s}_n | Z) + H(Z).$$

By (20), for large enough n, we have

$$H(Z) \leqslant H_2\left(\epsilon + 2\exp\left(-\frac{2\epsilon^2}{(M-1)^2} n\right)\right) \log_{|\mathcal{A}|} 2,$$

where $H_2(x) = -x\log_2 x - (1-x)\log_2(1-x)$ is the binary entropy function.

We also have

$$H(\boldsymbol{s}_n | Z) = H(\boldsymbol{s}_n | Z = 0) + H(\boldsymbol{s}_n | Z = 1).$$

For the first summand, by the definition of conditional entropy, and after replacing the unknown distribution with a uniform one to obtain an upper bound, we get

$$\begin{aligned} &H(\boldsymbol{s}_n | Z = 0) \\ &\leqslant \left(\epsilon + 2\exp\left(-\frac{2\epsilon^2}{(M-1)^2} n\right)\right) \log_{|\mathcal{A}|} \left| \bigcup_{i=1}^{|\boldsymbol{s}_0| + (M-1)n} \mathcal{A}^i \right| \\ &\leqslant \left(\epsilon + 2\exp\left(-\frac{2\epsilon^2}{(M-1)^2} n\right)\right) \\ &\qquad \cdot \log_{|\mathcal{A}|}(|\boldsymbol{s}_0| + (M-1)n + 1). \end{aligned}$$

Similarly, for the second summand,

$$\begin{aligned} H(\boldsymbol{s}_n | Z = 1) &\leqslant \left(1 - \epsilon - 2\exp\left(-\frac{2\epsilon^2}{(M-1)^2} n\right)\right) \\ &\qquad \cdot \log_{|\mathcal{A}|} \left(\sum_{i = \mathbb{E}(|\boldsymbol{s}_n|) - \epsilon n}^{\mathbb{E}(|\boldsymbol{s}_n|) + \epsilon n} |\mathcal{B}_i(\mathbb{B}_\epsilon(\Gamma_k))| \right) \\ &\leqslant \log_{|\mathcal{A}|} \left(\sum_{i = \mathbb{E}(|\boldsymbol{s}_n|) - \epsilon n}^{\mathbb{E}(|\boldsymbol{s}_n|) + \epsilon n} |\mathcal{B}_i(\mathbb{B}_\epsilon(\Gamma_k))| \right). \end{aligned}$$

However, by the definition of the capacity of semiconstrained systems, for all large enough n,

$$|\mathcal{B}_i(\mathbb{B}_\epsilon(\Gamma_k))| \leqslant |\mathcal{A}|^{i \cdot \mathsf{cap}(\mathbb{B}_\epsilon(\Gamma_k)) + \epsilon}.$$

It follows that

$$H(\boldsymbol{s}_n | Z = 1) \leqslant \log_{|\mathcal{A}|} \left(2\epsilon n |\mathcal{A}|^{(\mathbb{E}(|\boldsymbol{s}_n|) + \epsilon n)(\mathsf{cap}(\mathbb{B}_\epsilon(\Gamma_k)) + \epsilon)} \right).$$

Combining all of these together,

$$\begin{aligned} \mathcal{H}_n &= \frac{1}{\mathbb{E}(|\boldsymbol{s}_n|)} \cdot H(\boldsymbol{s}_n) \\ &\leqslant \frac{1}{\mathbb{E}(|\boldsymbol{s}_n|)} \log_{|\mathcal{A}|} \left(2\epsilon n |\mathcal{A}|^{(\mathbb{E}(|\boldsymbol{s}_n|) + \epsilon n)(\mathsf{cap}(\mathbb{B}_\epsilon(\Gamma_k)) + \epsilon)} \right) \\ &\quad + \frac{\epsilon + 2\exp\left(-\frac{2\epsilon^2}{(M-1)^2} n\right)}{\mathbb{E}(|\boldsymbol{s}_n|)} (|\boldsymbol{s}_0| + (M-1)n + 1) \\ &\quad + \frac{1}{\mathbb{E}(|\boldsymbol{s}_n|)} H_2\left(\epsilon + 2\exp\left(-\frac{2\epsilon^2}{(M-1)^2} n\right)\right) \log_{|\mathcal{A}|} 2. \end{aligned}$$

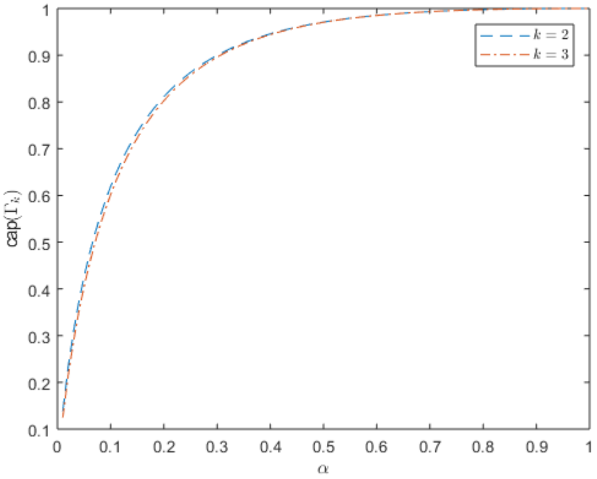

Figure 3. Entropy bound vs the probability of substitution, with $\mathcal{A} = \{0, 1\}$.

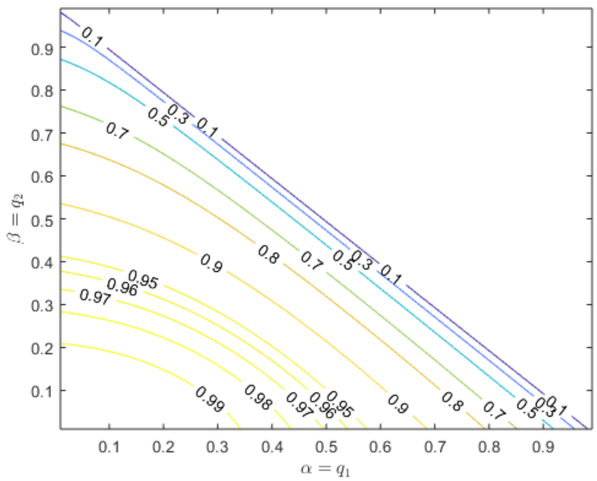

Figure 4. Contour plot of entropy bounds, with $\mathcal{A} = \{0, 1\}$, $k = 3$, $q_0 = 1 - \alpha - \beta$, $q_1 = \alpha$, $q_2 = \beta$.

Taking $\limsup_{n\to\infty}$ of both sides we obtain

$$\mathcal{H}_\infty \leqslant \left(1 + \frac{\epsilon}{\sum_{i=1}^{M-1} i q_i}\right) \cdot (\mathsf{cap}(\mathbb{B}_\epsilon(\Gamma_k)) + \epsilon) + \frac{\epsilon(M-1)}{\sum_{i=1}^{M-1} i q_i} + H_2(\epsilon)\log_{|\mathcal{A}|} 2.$$

Finally, taking $\lim_{\epsilon\to 0^+}$ of both sides, we obtain the claim. ■

Remark. We comment that if $\Gamma_k = \{\boldsymbol{\xi}_k\}$, i.e., Γ_k contains a single shift-invariant measure[1], then $\mathsf{cap}(\Gamma_k)$ has a nice form for all $k \in \mathbb{N}^+$ (see [14], [15]):

$$\mathsf{cap}(\Gamma_k) = -\sum_{a_1\dots a_k \in \mathcal{A}^k} \boldsymbol{\xi}_k^{a_1\dots a_k} \log_{|\mathcal{A}|} \frac{\boldsymbol{\xi}_k^{a_1\dots a_k}}{\bar{\boldsymbol{\xi}}_k^{a_1\dots a_{k-1}}},$$

where $\bar{\boldsymbol{\xi}}_k$ is the marginal of $\boldsymbol{\xi}_k$ on the first $k-1$ coordinates, i.e., $\bar{\boldsymbol{\xi}}_k^{a_1\dots a_{k-1}} = \sum_{b\in\mathcal{A}} \boldsymbol{\xi}_k^{a_1\dots a_{k-1}b}$. Furthermore, $\forall k \in \mathbb{N}^+$,

$$\mathsf{cap}(\Gamma_k) \geqslant \mathsf{cap}(\Gamma_{k+1}),$$

which follows from the fact that $\mathsf{cap}(\Gamma_k)$ can be viewed as the conditional entropy of a symbol given the $k-1$ previous symbols in a stationary process.

Using the preceding remark and Theorem 11, we can find a series of upper bounds on a given system:

$$\mathsf{cap}(\Gamma_1) \geqslant \mathsf{cap}(\Gamma_2) \geqslant \dots \geqslant \mathsf{cap}(\Gamma_k) \geqslant \dots \geqslant \mathcal{H}_\infty,$$

with Γ_k being the limit of $(x^{\boldsymbol{u}})_{\boldsymbol{u}\in\mathcal{A}^k}$.

In particular, for the system whose limit is given by (17), we have $\boldsymbol{\xi}^0 = \boldsymbol{\xi}^1 = {}^1\!/_2, \boldsymbol{\xi}^{00} = \boldsymbol{\xi}^{11} = {}^{(\alpha+1)}\!/_{2(1+3\alpha)}, \boldsymbol{\xi}^{01} = \boldsymbol{\xi}^{10} = {}^{\alpha}\!/_7$. It then follows that for this system $\mathcal{H}_\infty \leqslant H_2\left(\frac{2\alpha}{1+3\alpha}\right) = \mathsf{cap}(\Gamma_2)$. We can also compute $\mathsf{cap}(\Gamma_k)$ for $k = 3, 4, \dots$. Figure 3 shows the entropy bound we find using 2-mer and 3-mer frequencies. The two bounds are close, which suggests that we may be close to the exact entropy values. However, in the absence of a lower bound, this conjecture cannot be verified. The figure shows that when there is only one possible duplication length, the source of diversity is substitution, as may be expected. As $\alpha \to 1$, the relative number of substitutions increases, causing Γ_k to be close to the uniform distribution, and the entropy tends to 1. On the other hand, as $\alpha \to 0$, only duplications occur. This leads to the generation of low complexity sequences that consists of long runs of 0s and 1s, and thus entropy that is close to 0.

Figure 4 shows the entropy bound computed using 3-mer frequencies for the case in which $\mathcal{A} = \{0, 1\}$, $q_1 = \alpha, q_2 = \beta$ and $q_0 = 1 - \alpha - \beta$. So in this system, duplications of lengths 1 and 2 are both possible. It can be seen that similar to Figure 3, even a small probability of substitution leads to relatively high values of entropy. Furthermore, we note that, as may be expected, longer duplications lead to a smaller value of entropy.

V. Interspersed Duplication

In this section, we study the evolution of k-mer frequencies of the interspersed-duplication system, also using the stochastic-approximation technique.

Let $U = \bigcup_{i=1}^k \mathcal{A}^i$, i.e., the set consisting of all non-empty strings of length at most k. Also, let the vectors $\boldsymbol{x}_n$ and $\boldsymbol{\mu}_n$ be defined as before using U.

Theorem 12. *Consider $\boldsymbol{u} \in U$. In an interspersed-duplication system, for $\ell < |\boldsymbol{u}|$, we have*

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

[1] A shift-invariant measure $\boldsymbol{\xi}_k \in \mathcal{P}(\mathcal{A}^k)$ is a measure that satisfies $\sum_{a\in\mathcal{A}} \boldsymbol{\xi}_k^{a\boldsymbol{w}} = \sum_{a\in\mathcal{A}} \boldsymbol{\xi}_k^{\boldsymbol{w}a}$ for all $\boldsymbol{w} \in \mathcal{A}^{k-1}$. The k-mer distributions of cyclic strings are always shift invariant, and thus a converging sequence of such measures also converges to a shift-invariant measure.

Proof: The term $-(|\boldsymbol{u}|-1)x_n^{\boldsymbol{u}}$ accounts for the expected number of lost occurrences of $\boldsymbol{u}$ in $\boldsymbol{s}_n$ as a result of inserting the duplicate substring. To illustrate, assume $\mathcal{A}=\{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}$, $\boldsymbol{u}=\mathsf{ACT}$ and $\ell=1$. An occurrence of $\boldsymbol{u}=\mathsf{ACT}$ will be lost if for example an occurrence of the symbol G is duplicated and inserted after A in this occurrence of $\boldsymbol{u}$, since it becomes AGCT. The probability that a certain occurrence is lost equals $\frac{|\boldsymbol{u}|-1}{L_n}$. Since there are $\mu_n^{\boldsymbol{u}}$ such occurrences, the expected number of lost occurrences of $\boldsymbol{u}$ equals $\mu_n^{\boldsymbol{u}}\frac{|\boldsymbol{u}|-1}{L_n}=x_n^{\boldsymbol{u}}(|\boldsymbol{u}|-1)$. 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 $\boldsymbol{u}$ where the first i symbols come from the duplicate and the next $|\boldsymbol{u}|-i$ are from the substring that starts after the point of insertion of the duplicate. There are $\mu_n^{\boldsymbol{u}_{1,i}}$ occurrences of $\boldsymbol{u}_{1,i}$. The duplicate ends with one of these with probability $\frac{\mu_n^{\boldsymbol{u}_{1,i}}}{L_n}=x_n^{\boldsymbol{u}_{1,i}}$. Furthermore, the duplicate is inserted before an occurrence of $u_{i+1,|\boldsymbol{u}|-i}$ with probability $x_n^{\boldsymbol{u}_{i+1,|\boldsymbol{u}|-i}}$. Hence, the probability of a new occurrence created in this way is $x_n^{\boldsymbol{u}_{1,i}}x_n^{\boldsymbol{u}_{i+1,|\boldsymbol{u}|-i}}$, 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 $\boldsymbol{u}$. The last summation accounts for new occurrences of $\boldsymbol{u}$ in which the duplicate substring forms a middle part of $\boldsymbol{u}$ of length ℓ and previously existing substrings contribute a prefix of length i and a suffix of length $|\boldsymbol{u}|-\ell-i$. In terms of our running example with $\boldsymbol{u}=\mathsf{ACT}$ 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 $x_n^{\boldsymbol{u}_{1,i}\boldsymbol{u}_{i+\ell+1,|\boldsymbol{u}|-\ell-i}}x_n^{\boldsymbol{u}_{i+1,\ell}}=x_n^{\mathsf{AT}}x_n^{\mathsf{C}}$, where $i=1$. ■

Theorem 13. *For $\ell\geqslant|\boldsymbol{u}|$, we have*

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

Proof: The first two summations are similar to the first two summations for the case of $\ell<|\boldsymbol{u}|$, but a term corresponding to the third summation is not present. The term $(\ell-|\boldsymbol{u}|+1)x_n^{\boldsymbol{u}}$ corresponds to the cases in which a new occurrence of $\boldsymbol{u}$ is created as a substring of the duplicate substring. ■

Note that $\delta_\ell^{\boldsymbol{u}}$ depends only on $\boldsymbol{x}_n$ and is Lipschitz since $\boldsymbol{x}_n\in[0,1]^{|U|}$. Thus, (A.4) and (A.5) hold.

Since $h_\ell^{\boldsymbol{u}}(\boldsymbol{x}_n)=\delta_\ell^{\boldsymbol{u}}(\boldsymbol{x}_n)-\ell x_n^{\boldsymbol{u}}$, we have for $\ell<|\boldsymbol{u}|$ and $\ell\geqslant|\boldsymbol{u}|$, respectively,

$$\begin{aligned}h_\ell^{\boldsymbol{u}}(\boldsymbol{x}_n)=&-(\ell+|\boldsymbol{u}|-1)x_n^{\boldsymbol{u}}+\sum_{i=1}^{\ell}x_n^{\boldsymbol{u}_{1,i}}x_n^{\boldsymbol{u}_{i+1,|\boldsymbol{u}|-i}}\\&+\sum_{i=1}^{\ell}x_n^{\boldsymbol{u}_{1,|\boldsymbol{u}|-i}}x_n^{\boldsymbol{u}_{|\boldsymbol{u}|-i+1,i}}\\&+\sum_{i=1}^{|\boldsymbol{u}|-\ell-1}x_n^{\boldsymbol{u}_{1,i}\boldsymbol{u}_{i+\ell+1,|\boldsymbol{u}|-\ell-i}}x_n^{\boldsymbol{u}_{i+1,\ell}}\end{aligned}\tag{21}$$

$$h_\ell^{\boldsymbol{u}}(\boldsymbol{x}_n)=-2(|\boldsymbol{u}|-1)x_n^{\boldsymbol{u}}+2\sum_{i=1}^{|\boldsymbol{u}|-1}x_n^{\boldsymbol{u}_{1,i}}x_n^{\boldsymbol{u}_{i+1,|\boldsymbol{u}|-i}}\tag{22}$$

Recall that $\boldsymbol{h}_\ell(\boldsymbol{x})=\left(h_\ell^{\boldsymbol{u}}(\boldsymbol{x})\right)_{\boldsymbol{u}\in U}$. So from (21) and (22), we can find the ODE $d\boldsymbol{x}_t/dt=\boldsymbol{h}(\boldsymbol{x}_t)=\sum_{\ell=1}^{M-1}q_\ell\boldsymbol{h}_\ell(\boldsymbol{x}_t)$. As an example, if $k=3$ and $\mathcal{A}=\{\mathsf{A},\mathsf{C}\}$, then $U=(\mathsf{A},\mathsf{C},\mathsf{AA},\mathsf{AC},\mathsf{CA},\mathsf{CC},\mathsf{AAA},\dots,\mathsf{CCC})$ and some of the equations of the ODE system are

$$\begin{aligned}\frac{d}{dt}x_t^{\mathsf{A}}&=\frac{d}{dt}x_t^{\mathsf{C}}=0,\\\frac{d}{dt}x_t^{\mathsf{AA}}&=-2x_t^{\mathsf{AA}}+2\left(x_t^{\mathsf{A}}\right)^2,\\\frac{d}{dt}x_t^{\mathsf{AC}}&=-2x_t^{\mathsf{AC}}+2x_t^{\mathsf{A}}x_t^{\mathsf{C}},\\\frac{d}{dt}x_t^{\mathsf{AAC}}&=-(4-q_1)x_t^{\mathsf{AAC}}+2x_t^{\mathsf{A}}x_t^{\mathsf{AC}}+(2-q_1)x_t^{\mathsf{C}}x_t^{\mathsf{AA}}.\end{aligned}\tag{23}$$

For a vector $\boldsymbol{x}$ that contains the elements $\left(x^a\right)_{a\in\mathcal{A}}$ and for $\boldsymbol{v}\in\mathcal{A}^*$, define $p(\boldsymbol{v},\boldsymbol{x})=\prod_{a\in\mathcal{A}}\left(x^a\right)^{n_{\boldsymbol{v}}(a)}$, where $n_{\boldsymbol{v}}(a)$ is the number of occurrences of a in $\boldsymbol{v}$, and note that $p(\boldsymbol{v}\boldsymbol{w},\boldsymbol{x})=p(\boldsymbol{v},\boldsymbol{x})p(\boldsymbol{w},\boldsymbol{x})$. We now turn to find the solutions to the ODE $d\boldsymbol{x}_t/dt=\boldsymbol{h}(\boldsymbol{x}_t)$.

Lemma 14. *Consider the ODE $d\boldsymbol{x}_t/dt=\boldsymbol{h}(\boldsymbol{x}_t)$ where $\boldsymbol{h}(\boldsymbol{x})=\sum_{\ell=1}^{M-1}q_\ell\boldsymbol{h}_\ell(\boldsymbol{x})$ and the elements of $\boldsymbol{h}_\ell(\boldsymbol{x})$ are given by (21) and (22). The solution to this ODE is*

$$x_t^{\boldsymbol{v}}=p(\boldsymbol{v},\boldsymbol{x}_0)+\sum_i b_i^{\boldsymbol{v}}e^{-d_i^{\boldsymbol{v}}t},\qquad\boldsymbol{v}\in U,\tag{24}$$

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

Proof: We prove the lemma by induction. The claim (24) holds for $\boldsymbol{v}\in\mathcal{A}$, 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 $\boldsymbol{u}\in U$ such that $|\boldsymbol{u}|>1$, and assume that (24) holds for all $\boldsymbol{v}\in U$ such that $|\boldsymbol{v}|<|\boldsymbol{u}|$. We show that it also holds for $\boldsymbol{u}$, i.e., $x_t^{\boldsymbol{u}}=p(\boldsymbol{u},\boldsymbol{x}_0)+\sum_i b_i^{\boldsymbol{u}}e^{-d_i^{\boldsymbol{u}}t}$. Using the assumption, we rewrite (21) and (22) as

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

for $\ell<|\boldsymbol{u}|$, and

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

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

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

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

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

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

For example, the solutions to (23) with $q_1 = 0$ are

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

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 14 to characterize the limits of the frequencies of substrings in interspersed-duplication systems.

Theorem 15. *Let $U = \bigcup_{i=1}^k \mathcal{A}^i$, and let $\boldsymbol{x}_n = (x_n^{\boldsymbol{u}})_{\boldsymbol{u} \in U}$ 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*

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

Note that the existence of the limits x_∞^a of x_n^a, for $a \in \mathcal{A}$, was also shown in Theorem 1.

Proof: From Theorem 4, we know that the limit set of $\boldsymbol{x}_n$ is an internally chain transitive invariant set of the ODE described by (21) and (22). Let this set, which consists of points of the form $\boldsymbol{y} = (y^{\boldsymbol{v}})_{\boldsymbol{v} \in U}$, be denoted by H. Since for each $\boldsymbol{u} \in U$, $x_n^{\boldsymbol{u}} \in [0, 1]$, we can assume that $H \subseteq [0, 1]^{|U|}$ without any loss of generality. We now use these facts to show that $y^{\boldsymbol{u}} = p(\boldsymbol{u}, \boldsymbol{y})$ for each $\boldsymbol{y} \in H$ and $\boldsymbol{u} \in U$.

Suppose to the contrary that there exist $\boldsymbol{y} \in H$ and $\boldsymbol{u} \in U$ such that $y^{\boldsymbol{u}} \neq p(\boldsymbol{u}, \boldsymbol{y})$. Among all possible choices for such $\boldsymbol{y}$ and $\boldsymbol{u}$, choose the ones where the length $|\boldsymbol{u}|$ of $\boldsymbol{u}$ is minimum. We know that the length of $\boldsymbol{u}$ will be at least 2 since $y^a = p(a, \boldsymbol{y})$ for all $a \in \mathcal{A}$. Hence, for all $\boldsymbol{v} \in \mathcal{A}^*$ with $|\boldsymbol{v}| < |\boldsymbol{u}|$, and all $\boldsymbol{z} \in H$, we have $z^{\boldsymbol{v}} = p(\boldsymbol{v}, \boldsymbol{z})$. Using this fact, one can show that if $\boldsymbol{x}_0 \in H$, the solution to the ODE described by (21) and (22) will have the form $x_t^{\boldsymbol{u}} = p(\boldsymbol{u}, \boldsymbol{x}_0) + b e^{-c^{\boldsymbol{u}} t}$, where $b = x_0^{\boldsymbol{u}} - p(\boldsymbol{u}, \boldsymbol{x}_0)$ and $c^{\boldsymbol{u}} \geqslant |\boldsymbol{u}|$ by a similar proof as Lemma 14.

By the definition of internal chain transitivity, for any $\epsilon > 0$ and $T > 0$, there exist $N \geqslant 1$ and a sequence $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_N$ with $\boldsymbol{y}_i \in H$, $\boldsymbol{y}_0 = \boldsymbol{y}_N = \boldsymbol{y}$ such that for $0 \leqslant i < N$, if $\boldsymbol{x}_0 = \boldsymbol{y}_i$, then there exists $t \geqslant T$ 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 $t' \geqslant T$, $\boldsymbol{x}_{t'}$ is in the ϵ-neighborhood of $\boldsymbol{y}_{i+1}$. Since H is invariant, we know that $\boldsymbol{y}_i \in H$ and therefore $x_{t'}^{\boldsymbol{u}} = p(\boldsymbol{u}, \boldsymbol{y}_i) + (y_i^{\boldsymbol{u}} - p(\boldsymbol{u}, \boldsymbol{y}_i)) e^{-c^{\boldsymbol{u}} t'}$. So we have

$$\left| y_{i+1}^{\boldsymbol{u}} - x_{t'}^{\boldsymbol{u}} \right| = \left| y_{i+1}^{\boldsymbol{u}} - p(\boldsymbol{u}, \boldsymbol{y}_i) - (y_i^{\boldsymbol{u}} - p(\boldsymbol{u}, \boldsymbol{y}_i)) e^{-c^{\boldsymbol{u}} t'} \right| \leqslant \epsilon. \tag{25}$$

Furthermore, since for $a \in \mathcal{A}$, $x_{t'}^a = p(a, \boldsymbol{y}_i) = y_i^a$, we also have

$$\left| y_{i+1}^a - y_i^a \right| \leqslant \epsilon. \tag{26}$$

So if $p(\boldsymbol{u}, \boldsymbol{y}_{i+1}) > 0$, we have,

$$\begin{aligned} & p(\boldsymbol{u}, \boldsymbol{y}_i) - p(\boldsymbol{u}, \boldsymbol{y}_{i+1}) \\ & \leqslant \prod_{a \in \mathcal{A}} (y_{i+1}^a)^{n_{\boldsymbol{u}}(a)} \left(\prod_{a \in \mathcal{A}} \left(\frac{y_{i+1}^a + \epsilon}{y_{i+1}^a} \right)^{n_{\boldsymbol{u}}(a)} - 1 \right) \\ & \leqslant \prod_{a \in \mathcal{A}} \left(1 + \frac{\epsilon}{y_{i+1}^a} \right)^{n_{\boldsymbol{u}}(a)} - 1 \\ & \leqslant \left(1 + \frac{\epsilon}{\min\limits_{a \in \mathcal{A}} y_{i+1}^a} \right)^{|\boldsymbol{u}|} - 1, \end{aligned} \tag{27}$$

and if $p(\boldsymbol{u}, \boldsymbol{y}_{i+1}) = 0$,

$$p(\boldsymbol{u}, \boldsymbol{y}_i) - p(\boldsymbol{u}, \boldsymbol{y}_{i+1}) \leqslant \epsilon^{n_{\boldsymbol{u}}(a)}. \tag{28}$$

Thus, from (25), (27) and (28), it follows that

$$y_{i+1}^{\boldsymbol{u}} - p\big(\boldsymbol{u}, \boldsymbol{y}_{i+1}\big) \leqslant e^{-c^{\boldsymbol{u}} T} + \epsilon + O(\epsilon). \tag{29}$$

In particular, (29) holds for $i = N - 1$, i.e.,

$$y^{\boldsymbol{u}} - p(\boldsymbol{u}, \boldsymbol{y}) \leqslant e^{-c^{\boldsymbol{u}} T} + O(\epsilon).$$

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

The theorem shows that for $\boldsymbol{u} \in \mathcal{A}^*$, the frequency of $\boldsymbol{u}$ converges to the frequency of same in an iid sequence where the probability of $a \in \mathcal{A}$ equals x_∞^a. Figure 5 illustrates an example, obtained via simulation, where the system starts with $\boldsymbol{s}_0 = \mathsf{AGCGTATGCG}$ 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 $x_n^{\mathsf{GGG}} = 0.0992 \simeq \big(x_n^{\mathsf{G}}\big)^3 = 0.1084$.

The limit set for $(x^{\boldsymbol{u}})_{\boldsymbol{u} \in \mathcal{A}^k}$ implied by Theorem 15 includes the uniform distribution. As a result, the application of Theorem 11 leads to the trivial upper bound of $|\mathcal{A}|$. It thus appears that determining the entropy of ID systems requires determining not only the limit set for the k-mer frequencies but also their limiting distribution, as well as results that can relate this distribution to entropy. We leave pursuing this

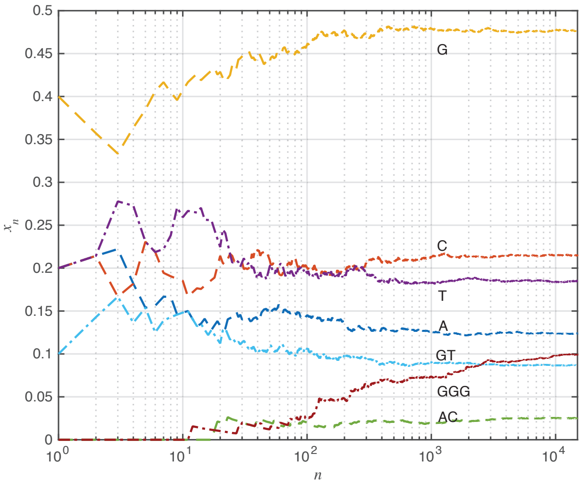

Figure 5. Symbol frequencies vs the number of duplications in an interspersed-duplication system, with $\boldsymbol{s}_0 = \mathsf{AGCGTATGCG}$, and $q_4 = q_6 = 1/2$.

direction to future work. Nevertheless, the fact that the k-mer frequencies are similar to those in iid sequences, suggest that interspersed duplication leads to high entropy. At least for certain special cases, this is indeed the case. For binary ($\mathcal{A} = \{0, 1\}$) interspersed duplications of length 1, the entropy is found in [12] as

$$\frac{\log_2 e}{t_0 + t_1}((t_0 + t_1)\mathsf{H}_{t_0+t_1} - t_0\mathsf{H}_{t_0} - t_1\mathsf{H}_{t_1}), \tag{30}$$

where t_0 and t_1 are the numbers of 0s and 1s in $\boldsymbol{s}_0$, respectively, and H_t is the t-th Harmonic number. For $t_0 = t_1 \to \infty$, the entropy can be shown to equal 1.

VI. Conclusion

We studied the limiting behavior of two stochastic duplication systems, tandem duplication with substitution and interspersed duplication. We used stochastic approximation to compute the limits of k-mer frequencies for tandem duplications and substitutions. We also provided a method for determining upper bounds on the entropy of these systems. For interspersed duplication system, we established that k-mer frequencies tend to the corresponding probabilities in sequences generated by iid sources. This suggests that these systems have high entropy, and the structure of the limit set for k-mers prevents us from obtaining non-trivial upper bounds. Many problems are left open. First, for tandem duplication and substitution systems, other mutations, such as deletions were not studied; and for interspersed duplication, substitutions, deletions and other mutations were not considered. Moreover, for interspersed duplication, providing nontrivial upper bounds on the entropy requires further research. While we conjecture that the upper bounds presented here are close to actual values, lower bounds on the entropy are needed to verify this claim. Since this work was limited to the asymptotic analysis of these systems, more research is required to quantify their finite-time behavior.

Appendix

We give the proofs of Lemmas 5–7 in this appendix. Recall that the lemmas give us the expected number of new occurrences of a k-mer $\boldsymbol{u}$ in three different cases according to the relative position of the newly created substring and the template sequence, as illustrated in Figure 1.

A. *Lemma 5 (Case 1)*

In this case we have $1 \leqslant b < \min(\ell, k-\ell+1)$ (regardless of whether $k \geqslant 2\ell$ or $k < 2\ell$), the new occurrences of $\boldsymbol{u}$ always contain some (but not all) of the template and all of the new copy. This scenario is labeled as Case 1 in Figure 1.

Suppose $Y_b = 1$. Since the copy and the template are identical, elements of $\boldsymbol{u}$ that coincide with the same positions in these two substrings must also be identical. So a necessary condition for $Y_b = 1$ is

$$\boldsymbol{u}_{1,b} = \boldsymbol{u}_{1+\ell,b}.$$

Assume this condition is satisfied. Then $Y_b = 1$ if and only if the sequence starting at the beginning of the template in $\boldsymbol{s}_n$ is equal to $\boldsymbol{u}_{b+1,k-b}$, which has probability $x^{\boldsymbol{u}_{b+1,k-b}}$.

As an example for $k \geqslant 2\ell$, consider

$$\begin{aligned}\boldsymbol{s}_n &= \boldsymbol{v}1234567\boldsymbol{w},\\ \boldsymbol{s}_{n+1} &= \boldsymbol{v}12\overline{3\underline{123}456}7\boldsymbol{w}, \text{ where } Y_b = 1 \text{ for } b = 1,\\ u_1 &= u_4 = 3,\end{aligned}$$

where $\boldsymbol{v}, \boldsymbol{w} \in \mathcal{A}^*$, $\boldsymbol{u}$ is overlined, and the copy is underlined. Note that $\boldsymbol{s}_n$ contains $\boldsymbol{u}_{b+1,k-b} = 123456$. For $k < 2\ell$, consider

$$\begin{aligned}\boldsymbol{s}_n &= \boldsymbol{v}1234\boldsymbol{w},\\ \boldsymbol{s}_{n+1} &= \boldsymbol{v}12\overline{3\underline{123}4}\boldsymbol{w}, \text{ where } Y_b = 1 \text{ for } b = 1,\\ u_1 &= u_4 = 3.\end{aligned}$$

B. *Lemma 6 (Case 2)*

In Case 2, $\boldsymbol{u}$ either i) contains both the template and the copy completely, or ii) intersects with both but contains neither. Note that this case cannot occur if $k = 2\ell - 1$.

First, assume $k \geqslant 2\ell$. The condition on b translates to $\ell \leqslant b < k - \ell + 1$ and the new occurrence of $\boldsymbol{u}$ contains both the template and the copy. This is labeled as Case 2 in Figure 1 (below $\boldsymbol{s}_{n+1}$). With the same logic as in Case 1, it is clear that we need

$$\boldsymbol{u}_{b-\ell+1,\ell} = \boldsymbol{u}_{b+1,\ell},$$

Assuming this condition is satisfied, we have $Y_b = 1$ if and only if the substring $\boldsymbol{u}_{1,b-\ell}\boldsymbol{u}_{b+1,k-b}$ occurs in $\boldsymbol{s}_n$ at a certain position, which occurs with probability $x^{\boldsymbol{u}_{1,b-\ell}\boldsymbol{u}_{b+1,k-b}}$.

For example, consider

$$\begin{aligned}\boldsymbol{s}_n &= \boldsymbol{v}412356\boldsymbol{w},\\ \boldsymbol{s}_{n+1} &= \boldsymbol{v}\overline{4123\underline{123}}56\boldsymbol{w}, \text{ where } Y_b = 1 \text{ for } b = 4,\\ \boldsymbol{u}_{2,3} &= \boldsymbol{u}_{5,3} = 123.\end{aligned}$$

Now suppose $\ell + 1 \leqslant k \leqslant 2\ell - 2$. The condition on b from the statement of the lemma is $k - \ell + 1 \leqslant b < \ell$. The new occurrence of $\boldsymbol{u}$ contains some (but not all) of the elements of the template and some (but not all) of the elements of the copy, as illustrated in Figure 1, Case 2, above $\boldsymbol{s}_{n+1}$. The following constraint on $\boldsymbol{u}$ must hold

$$\boldsymbol{u}_{1,k-\ell} = \boldsymbol{u}_{\ell+1,k-\ell},$$

implying that $\phi_\ell(\boldsymbol{u}) = \mathsf{X}^\ell 0^{k-\ell}$. For example, consider

$$\begin{aligned} \boldsymbol{s}_n &= \boldsymbol{v}123\boldsymbol{w}, \\ \boldsymbol{s}_{n+1} &= \boldsymbol{v}1\overline{23\underline{12}}\underline{3}\boldsymbol{w}, \text{ where } Y_b = 1 \text{ for } b = 2, \\ u_1 &= u_4 = 2. \end{aligned}$$

We have $Y_b = 1$ iff the sequence starting at the beginning of the template in $\boldsymbol{s}_n$ is equal to $\boldsymbol{u}_{b+1,\ell-b}\boldsymbol{u}_{1,b}$, which has probability $x^{\boldsymbol{u}_{b+1,\ell-b}\boldsymbol{u}_{1,b}}$.

C. Lemma 7 (Case 3)

In this case, we have $\max(k-\ell+1,\ell) \leqslant b \leqslant k-1$ (regardless of whether $k \geqslant 2\ell$ or $k < 2\ell$), the new occurrence of $\boldsymbol{u}$ contains the template and some (but not all) of the elements of the copy. This is labeled as Case 3 in Figure 1. The constraint on $\boldsymbol{u}$ is

$$\boldsymbol{u}_{b-\ell+1,k-b} = \boldsymbol{u}_{b+1,k-b}.$$

As examples, consider

$$\begin{aligned} \boldsymbol{s}_n &= \boldsymbol{v}456123\boldsymbol{w}, \\ \boldsymbol{s}_{n+1} &= \boldsymbol{v}\overline{456123\underline{1}}\underline{23}\boldsymbol{w}, \text{ where } Y_b = 1 \text{ for } b = 6, \\ u_4 &= u_7 = 1, \end{aligned}$$

for $k \geqslant 2\ell$, and

$$\begin{aligned} \boldsymbol{s}_n &= \boldsymbol{v}4123\boldsymbol{w}, \\ \boldsymbol{s}_{n+1} &= \boldsymbol{v}\overline{4123\underline{1}}\underline{23}\boldsymbol{w}, \text{ where } Y_b = 1 \text{ for } b = 4, \\ u_2 &= u_5 = 1, \end{aligned}$$

for $\ell < k < 2\ell$.

We have $Y_b = 1$ if and only if $\boldsymbol{u}_{1,b}$ occurs in $\boldsymbol{s}_n$ at a certain position, which has probability $x^{\boldsymbol{u}_{1,b}}$.

References

[1] C. Adami, "Information theory in molecular biology", *Physics of Life Reviews*, vol. 1, no. 1, pp. 3–22, 2004.

[2] G. Battail, "Biology needs information theory", *Biosemiotics*, vol. 6, no. 1, pp. 77–103, 2013.

[3] V. S. Borkar, *Stochastic Approximation*. Cambridge University Press, 2008.

[4] M. D. Cao, T. I. Dix, L. Allison, and C. Mears, "A simple statistical algorithm for biological sequence compression", in *2007 Data Compression Conference (DCC'07)*, IEEE, 2007, pp. 43–52.

[5] S. Chandak, K. Tatwawadi, I. Ochoa, M. Hernaez, and T. Weissman, "Spring: A next-generation compressor for fastq data", *Bioinformatics*, 2018.

[6] Y. M. Chee, J. Chrisnata, H. M. Kiah, and T. T. Nguyen, "Deciding the confusability of words under tandem repeats", *arXiv preprint arXiv:1707.03956*, 2017.

[7] X. Chen, S. Kwong, and M. Li, "A compression algorithm for DNA sequences and its applications in genome comparison", *Genome informatics*, vol. 10, pp. 51–61, 1999.

[8] B. Chern, I. Ochoa, A. Manolakos, A. No, K. Venkat, and T. Weissman, "Reference based genome compression", in *2012 IEEE Information Theory Workshop*, IEEE, 2012, pp. 427–431.

[9] S. Dancoff and H. Quastler, "The information content and error rate of living things", *Essays on the Use of Information Theory in Biology*, 1953.

[10] David Williams, *Probability with Martingales*. Cambridge: Cambridge University Press, 1991.

[11] L. Dolecek and V. Anantharam, "Repetition error correcting sets: Explicit constructions and prefixing methods", *SIAM Journal on Discrete Mathematics*, vol. 23, no. 4, pp. 2120–2146, 2010.

[12] O. Elishco, F. Farnoud, M. Schwartz, and J. Bruck, "The entropy rate of some Pólya string models", *IEEE Trans. Information Theory*, 2019, to appear.

[13] O. Elishco, T. Meyerovitch, and M. Schwartz, "On encoding semiconstrained systems", *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2474–2484, 2018.

[14] ——, "On independence and capacity of multidimensional semiconstrained systems", *IEEE Transactions on Information Theory*, vol. 64, no. 10, pp. 6461–6483, 2018.

[15] ——, "Semiconstrained systems", *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 1688–1702, 2016.

[16] M. Farach, M. Noordewier, S. Savari, L. Shepp, A. Wyner, and J. Ziv, "On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence", in *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '95, Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1995, pp. 48–57.

[17] F. Farnoud, M. Schwartz, and J. Bruck, "A stochastic model for genomic interspersed duplication", in *IEEE Int. Symp. Information Theory (ISIT)*, Jun. 2015, pp. 904–908.

[18] ——, "Estimation of duplication history under a stochastic model for tandem repeats", *BMC Bioinformatics*, vol. 20, no. 1, p. 64, 2019.

[19] ——, "The capacity of string-duplication systems", *IEEE Trans. Information Theory*, vol. 62, no. 2, pp. 811–824, Feb. 2016.

[20] B. Hajek, *Random processes for engineers*. Cambridge university press, 2015.

[21] P. Hanus, B. Goebel, J. Dingel, J. Weindl, J. Zech, Z. Dawy, J. Hagenauer, and J. C. Mueller, "Information and communication theory in molecular biology", *Electrical Engineering*, vol. 90, no. 2, pp. 161–173, 2007.

[22] R. Heckel, I. Shomorony, K. Ramchandran, and N. David, "Fundamental limits of DNA storage systems", in *2017 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2017, pp. 3130–3134.

[23] H. Herzel, W. Ebeling, and A. O. Schmitt, "Entropies of biosequences: The role of repeats", *Physical Review E*, vol. 50, no. 6, p. 5061, 1994.

[24] S. Jain, F. Farnoud, M. Schwartz, and J. Bruck, "Noise and uncertainty in string-duplication systems", in *IEEE Int. Symp. Information Theory (ISIT)*, Aachen, Germany, Jun. 2017.

[25] S. Jain, F. Farnoud, and J. Bruck, "Capacity and expressiveness of genomic tandem duplication", *IEEE Trans. Information Theory*, vol. 63, no. 10, Oct. 2017.

[26] S. Jain, F. F. Hassanzadeh, M. Schwartz, and J. Bruck, "Duplication-correcting codes for data storage in the DNA of living organisms", *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 4996–5010, 2017.

[27] H. M. Kiah, G. J. Puleo, and O. Milenkovic, "Codes for DNA storage channels", in *2015 IEEE Information Theory Workshop (ITW)*, IEEE, 2015, pp. 1–5.

[28] G. Korodi and I. Tabus, "Normalized maximum likelihood model of order-1 for the compression of DNA sequences", in *2007 Data Compression Conference (DCC'07)*, IEEE, 2007, pp. 33–42.

[29] 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.

[30] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding over sets for DNA storage", in *2018 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2018, pp. 2411–2415.

[31] A. Lenz, A. Wachter-Zeh, and E. Yaakobi, "Bounds on codes correcting tandem and palindromic duplications", *arXiv preprint arXiv:1707.00052*, 2017.

[32] M. Levy and E. Yaakobi, "Mutually uncorrelated codes for DNA storage", *IEEE Transactions on Information Theory*, 2018.

[33] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, "An information-based sequence distance and its application to whole mitochondrial genome phylogeny", *Bioinformatics*, vol. 17, no. 2, pp. 149–154, 2001.

[34] P. Liò, A. Politi, M. Buiatti, and S. Ruffo, "High statistics block entropy measures of DNA sequences", *Journal of Theoretical Biology*, vol. 180, no. 2, pp. 151–160, May 1996.

[35] D. Loewenstern and P. N. Yianilos, "Significantly lower entropy estimates for natural DNA sequences", *Journal of Computational Biology*, vol. 6, no. 1, pp. 125–142, 1999.

[36] H. Lou, F. Farnoud, and M. Schwartz, "Evolution of N-gram frequencies under duplication and substitution mutations", in *IEEE Int. Symp. Information Theory (ISIT)*, Jun. 2018.

[37] H. Mahmoud, *Pólya urn models*. Chapman and Hall/CRC, 2008.

[38] C. D. Meyer, *Matrix analysis and applied linear algebra*. Siam, 2000, vol. 71.

[39] O. Milenkovic, G. Alterovitz, G. Battail, T. P. Coleman, J. Hagenauer, S. P. Meyn, N. Price, M. F. Ramoni, I. Shmulevich, and W. Szpankowski, "Introduction to the special issue on information theory in molecular biology and neuroscience", Institute of Electrical and Electronics Engineers, 2010.

[40] A. S. Motahari, G. Bresler, and N. David, "Information theory of DNA shotgun sequencing", *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6273–6289, 2013.

[41] N. Mundy and A. J. Helbig, "Origin and evolution of tandem repeats in the mitochondrial DNA control region of shrikes (lanius spp.)", *Journal of Molecular Evolution*, vol. 59, no. 2, pp. 250–257, 2004.

[42] S. Ohno, *Evolution by Gene Duplication*. Springer-Verlag, 1970.

[43] Y. L. Orlov and V. N. Potapov, "Complexity: An internet resource for analysis of DNA sequence complexity", *Nucleic Acids Research*, vol. 32, no. Web Server issue, W628–W633, Jul. 2004.

[44] D. S. Pavlichin, T. Weissman, and G. Yona, "The human genome contracts again", *Bioinformatics*, vol. 29, no. 17, pp. 2199–2202, 2013.

[45] N. Raviv, M. Schwartz, and E. Yaakobi, "Rank-modulation codes for DNA storage with shotgun sequencing", *IEEE Transactions on Information Theory*, vol. 65, no. 1, pp. 50–64, 2018.

[46] W. Rudin *et al.*, *Principles of mathematical analysis*. McGraw-hill New York, 1964, vol. 3.

[47] A. O. Schmitt and H. Herzel, "Estimating the entropy of DNA sequences", *Journal of Theoretical Biology*, vol. 188, no. 3, pp. 369–377, Oct. 1997.

[48] A. Sievers, K. Bosiek, M. Bisch, C. Dreessen, J. Riedel, P. Froß, M. Hausmann, and G. Hildenbrand, "K-mer content, correlation, and position analysis of genome DNA sequences for the identification of function and evolutionary features", *Genes*, vol. 8, no. 4, Apr. 2017.

[49] Y. Tang, Y. Yehezkeally, M. Schwartz, and F. Farnoud, "Single-error detection and correction for duplication and substitution channels", in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, 2019.

[50] K. Usdin, "The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases", *Genome Research*, vol. 18, no. 7, pp. 1011–1019, Jul. 2008.

[51] R. S. Varga, *Geršgorin and his circles*. Springer Science & Business Media, 2010, vol. 36.

[52] S. Vinga, "Information theory applications for biological sequence analysis", *Briefings in bioinformatics*, vol. 15, no. 3, pp. 376–389, 2013.

[53] H. Wan, L. Li, S. Federhen, and J. C. Wootton, "Discovering simple regions in biological sequences associated with scoring schemes", eng, *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, vol. 10, no. 2, pp. 171–185, 2003.

[54] A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski, "Alignment-free sequence comparison: Bene-

fits, applications, and tools", *Genome Biology*, vol. 18, no. 1, p. 186, Oct. 2017.

Hao Lou (S'18) is a PhD candidate in the Department of Electrical and Computer Engineering at the University of Virginia. His research interests include data deduplication, stochastic and information-theoretic modeling of DNA mutations, compression of metagenomic sequencing data and computational biology. He received his Bachelors degree from Xi'an Jiaotong University, China in 2017.

Moshe Schwartz (M'03–SM'10) is an associate professor at the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Israel. His research interests include algebraic coding, combinatorial structures, and digital sequences.

Prof. Schwartz received the B.A. (*summa cum laude*), M.Sc., and Ph.D. degrees from the Technion – Israel Institute of Technology, Haifa, Israel, in 1997, 1998, and 2004 respectively, all from the Computer Science Department. He was a Fulbright post-doctoral researcher in the Department of Electrical and Computer Engineering, University of California San Diego, and a post-doctoral researcher in the Department of Electrical Engineering, California Institute of Technology. While on sabbatical 2012–2014, he was a visiting scientist at the Massachusetts Institute of Technology (MIT).

Prof. Schwartz received the 2009 IEEE Communications Society Best Paper Award in Signal Processing and Coding for Data Storage. He has also been serving as an Associate Editor for Coding Techniques for the IEEE Transactions on Information Theory since 2014.

Jehoshua Bruck (S'86–M'89–SM'93–F'01) is the Gordon and Betty Moore Professor of computation and neural systems and electrical engineering at the California Institute of Technology (Caltech). His current research interests include information theory and systems and the theory of computation in nature.

Dr. Bruck received the B.Sc. and M.Sc. degrees in electrical engineering from the Technion-Israel Institute of Technology, in 1982 and 1985, respectively, and the Ph.D. degree in electrical engineering from Stanford University, in 1989. His industrial and entrepreneurial experiences include working with IBM Research where he participated in the design and implementation of the first IBM parallel computer; cofounding and serving as Chairman of Rainfinity (acquired in 2005 by EMC), a spin-off company from Caltech that created the first virtualization solution for Network Attached Storage; as well as cofounding and serving as Chairman of XtremIO (acquired in 2012 by EMC), a start-up company that created the first scalable all-flash enterprise storage system.

Dr. Bruck is a recipient of the Feynman Prize for Excellence in Teaching, the Sloan Research Fellowship, the National Science Foundation Young Investigator Award, the IBM Outstanding Innovation Award and the IBM Outstanding Technical Achievement Award.

Farzad Farnoud (Hassanzadeh) (M'13) is an Assistant Professor in the Department of Electrical and Computer Engineering and the Department of Computer Science at the University of Virginia. Previously, he was a postdoctoral scholar at the California Institute of Technology.

He received his MS degree in Electrical and Computer Engineering from the University of Toronto in 2008. From the University of Illinois at Urbana-Champaign, he received his MS degree in mathematics and his Ph.D. in Electrical and Computer Engineering in 2012 and 2013, respectively. His research interests include the information-theoretic and probabilistic analysis of genomic evolutionary processes; rank aggregation and gene prioritization; and coding for flash memory and DNA storage. He is the recipient of the 2013 Robert T. Chien Memorial Award from the University of Illinois for demonstrating excellence in research in electrical engineering and the recipient of the 2014 IEEE Data Storage Best Student Paper Award.