Evolution of N-gram Frequencies under Duplication and Substitution Mutations

Hao Lou
Electrical and Computer Engineering
University of Virginia
Charlottesville, VA 22904, USA
Email: hl2nu@virginia.edu

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

Farzad Farnoud (Hassanzadeh)
Electrical and Computer Engineering
University of Virginia
Charlottesville, VA 22904, USA
Email: farzad@virginia.edu

***Abstract*—The driving force behind the generation of biological sequences are genomic mutations that shape these sequences throughout their evolutionary history. An understanding of the statistical properties that result from 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 types of mutations, tandem duplication and substitution. These play a critical role in forming tandem repeat regions, which are common features of the genome of many organisms. We provide a stochastic model and, via stochastic approximation, study the behavior of the frequencies of N-grams in resulting sequences. Specifically, we show that N-gram frequencies converge almost surely to a set which we identify as a function of model parameters. From these frequencies, other statistics can be derived. In particular, we present a method for finding upper bounds on entropy.**

I. Introduction

Genomic sequences are formed over billions of years by biological mutation processes, including insertions, deletions, duplications, and substitutions. These processes can be viewed as stochastic string editing operations that shape the statistical properties of sequence data. For any given set of such processes and the associated probabilities, however, it is difficult to characterize the statistical properties that will arise. At the same time, understanding these properties is beneficial in both analysis and storage of the vast amount of biological data that is available nowadays.

In this paper, our goal is to provide a better understanding of the behavior of tandem duplication and substitution mutations by studying the evolution of N-gram frequencies (a.k.a. k-mer frequencies) as substrings of a sequence undergoing these mutations. N-gram frequencies are of interest since they allow us to determine the substrings that are likely to be generated under different modeling assumptions. In addition, they can be used to learn other properties of the sequence, e.g., bounds on entropy, which provides a limit of compression. Given sequence data, they can also be used to estimate model parameters, i.e., mutation rates.

Duplication refers to copying a segment of the sequence (called the *template*) and inserting it into the sequence. The two main types of duplication are *interspersed duplication* and *tandem duplication*. In the former, there is generally no relationship between where the template is located and where the copy is inserted. In the latter, which is the type of duplication studied here, the copy is inserted immediately after the template. This process is generally thought to be caused by *slipped-strand mispairings* [1], where during DNA synthesis, one strand in a DNA duplex becomes misaligned with the other. A substitution refers to changing a symbol in the sequence. 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. Depending on the length of the pattern, these repeats are referred to as microsatellites or minisatellites. Tandem repeats are known to cause important phenomena such as chromosome fragility [2].

In our model, a sequence evolves only through tandem duplications of different lengths and substitutions. In reality other mutations, such as deletions, are also present in tandem repeat regions. However, for the sake of simplicity they are not included in our model. The analysis of more complete models is left to future work. Furthermore, the term evolution refers to changes resulting from random mutations. The significantly more complex effect of natural selection is not considered.

Our study starts by considering how N-gram frequencies (number of occurrences divided by the length of the evolving string) change as a result of different mutations. To analyze such frequencies, we use the stochastic approximation method, which enables modeling a discrete dynamic system by a corresponding continuous model described by an ordinary differential equation (ODE). We show that the resulting ODE is stable and prove that N-gram frequencies converge almost surely to a set determined by the ODE. Our approach allows us to compute the limit for the frequency of any N-gram as a function of model parameters. We will then use these results to provide bounds on the entropy of sequences generated by the aforementioned mutation processes.

In previous work, the related problem of finding the combinatorial capacity of tandem duplication systems has been studied [3], [4]. Systems with both tandem duplication and substitution, again from a combinatorial point of view, were studied in [5]. The stochastic approximation framework has been used for studying interspersed duplications [6] and estimation of model parameters in tandem duplication systems [7]. Estimating the entropy of DNA sequences has been studied in [8], [9]. However neither of these are based on stochastic se-

978-1-5386-4780-6/18/$31.00©2018 IEEE

quence evolution models nor are they capable of characterizing the asymptotic frequencies of N-grams.

The rest of the paper is organized as follows. Notation and preliminaries are given in the next section. In Section III, we derive the expected behavior of the N-gram frequencies under tandem duplication and substitution. The proof of convergence and the limits are given in Section IV. Bounds on entropy are presented in Sections V. Section VI provides the concluding remarks. Due to space limitations, some of the proofs are omitted.

II. Preliminaries and Notation

For a positive integer m, let $[m] = \{1, \ldots, m\}$ and $Z_m = \{0, \ldots, m-1\}$. For an alphabet Σ, the set of all finite strings over Σ is denoted Σ^*. The elements in strings are indexed starting from 0, e.g., $\boldsymbol{s} = s_0 \cdots s_{m-1}$, where $|\boldsymbol{s}| = m$ is the length of $\boldsymbol{s}$. For $0 \le i, j \le m-1$, $\boldsymbol{s}_i^j$ denotes $s_i s_{i+1} \cdots s_j$. A j-(sub)string is a (sub)string of length j. For $\boldsymbol{u}, \boldsymbol{v} \in \Sigma^*$, their concatenation is denoted by $\boldsymbol{uv}$. For a positive integer j, $\boldsymbol{u}^j$ is a concatenation of j copies of $\boldsymbol{u}$, where $\boldsymbol{u}$ is a string or a single symbol. Note that the superscript j in $\boldsymbol{u}_i^j$ and $\boldsymbol{u}^j$ has different meanings.

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. The indexing is modulo the length $|\boldsymbol{s}_n|$ of $\boldsymbol{s}_n$. Our attention is focused on tandem duplication and substitution mutations. As an example of a tandem duplication, from $\boldsymbol{s} = 012345$ we may obtain $\boldsymbol{s}' = 01\overline{234}\underline{2345}$, where the template, is overlined and the copy is underlined. The length of a duplication is the length of the template (3 in the preceding example). A substitution changes one of the elements of $\boldsymbol{s}_n$ to a different symbol from the alphabet, e.g., $012345 \to 01\underline{5}345$. We assign probability q_0 to substitutions, where a position is chosen uniformly at random and the current symbol is changed with equal probability to one of the other alphabet symbols. Furthermore, to a duplication of length k we assign probability q_k. The position of the template is chosen at random among the $|\boldsymbol{s}_n|$ possible options. We assume there exists K such that $q_k = 0$ for $k > K$. Hence, $\sum_{k=0}^{K} q_k = 1$.

Stochastic Approximation: Let U denote the set of N-grams, that is, $U = \Sigma^N$, and whenever an ordering of those strings is required we shall assume a lexicographic ordering. For $\boldsymbol{u} \in U$, let $\mu_n^{\boldsymbol{u}}$ denote the number of occurrences of substring $\boldsymbol{u}$ in $\boldsymbol{s}_n$, and $\boldsymbol{\mu}_n = (\mu_n^{\boldsymbol{u}})_{\boldsymbol{u} \in U}$. Let $x_n^{\boldsymbol{u}} = \frac{\mu_n^{\boldsymbol{u}}}{|\boldsymbol{s}_n|}$, we are interested in the asymptotic behavior of $\boldsymbol{x}_n = \frac{\boldsymbol{\mu}_n}{|\boldsymbol{s}_n|} = (x_n^{\boldsymbol{u}})_{\boldsymbol{u} \in U}$, the vector of the frequencies of the substrings.

Let $l_{n+1} = |\boldsymbol{s}_{n+1}| - |\boldsymbol{s}_n|$ and define $E_k[\cdot]$ to be the expected value conditioned on $l_{n+1} = k$. We also let $\{\mathcal{F}_n\}$ be the filtration generated by the random variables $\{\boldsymbol{x}_n, |\boldsymbol{s}_n|\}$. Define

$$\begin{aligned} \boldsymbol{\delta}_k(\boldsymbol{x}) &= \boldsymbol{\delta}_k(\boldsymbol{x}_n) = E_k[\boldsymbol{\mu}_{n+1} - \boldsymbol{\mu}_n | \mathcal{F}_n] \\ \boldsymbol{h}_k(\boldsymbol{x}) &= \boldsymbol{\delta}_k(\boldsymbol{x}) - k\boldsymbol{x} \\ \boldsymbol{h}(\boldsymbol{x}) &= \sum_k q_k \boldsymbol{h}_k(\boldsymbol{x}). \end{aligned} \tag{1}$$

The vector $\boldsymbol{\delta}_k$ represents the expected change in the vector of the frequencies of N-grams assuming a substitution ($k = 0$) or a duplication of length k has occurred. Here, we have assumed that $E_k[\boldsymbol{\mu}_{n+1} - \boldsymbol{\mu}_n | \mathcal{F}_n]$ depends only on $\boldsymbol{x}_n$, and given $\boldsymbol{x}_n$, it is independent from $\boldsymbol{\mu}_n$ and n. The correctness of this assumption will be evident from Theorem 2. Because of independence from n, we write $\boldsymbol{\delta}_k(\boldsymbol{x}) = \boldsymbol{\delta}_k(\boldsymbol{x}_n)$. Furthermore, the element of $\boldsymbol{\delta}_k$ that corresponds to $\boldsymbol{u}$ is denoted by $\delta_k^{\boldsymbol{u}}(\boldsymbol{x})$. More precisely, $\delta_k^{\boldsymbol{u}}(\boldsymbol{x}_n) = E_k[\mu_{n+1}^{\boldsymbol{u}} - \mu_n^{\boldsymbol{u}} | \mathcal{F}_n]$. This notation also extends to $\boldsymbol{h}$. The vector $\boldsymbol{h}$ determines the overall expected behavior of the system.

If a certain set of conditions are satisfied, then Theorem 1 below can be used to relate the discrete system whose expected behavior is described by $\boldsymbol{h}(\boldsymbol{x}_n)$ to a continuous system described by an ordinary differential equation (ODE). The required conditions are similar to those in [6] and hold in our setup. 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 if $q_0 < 1$.

Theorem 1. *(See [10, Theorem 2]) The sequence $\{\boldsymbol{x}_n\}$ converges almost surely to a compact connected internally chain transitive invariant set of the ODE $d\boldsymbol{x}_t/dt = \boldsymbol{h}(\boldsymbol{x}_t)$.*

To find the aforementioned differential equation, we need to find $\delta_k^{\boldsymbol{u}}(\boldsymbol{x})$ for all k with $q_k > 0$ and $\boldsymbol{u} \in U$. In finding $\delta_k^{\boldsymbol{u}}(\boldsymbol{x})$, the following definitions will be useful.

For $\boldsymbol{u} \in \Sigma^*$ and $k > 0$, define $\phi_k(\boldsymbol{u})$ to be a vector of length $|\boldsymbol{u}|$ whose ith element determines if the symbol in position i of $\boldsymbol{u}$ equals the symbol in position $i - k$. More specifically, the ith element of $\phi_k(\boldsymbol{u})$ is

$$\phi_k(\boldsymbol{u})_i = \begin{cases} X_i, & i = 0, 1, 2, \ldots k-1 \\ 0, & i \ge k, u_i = u_{i-k} \\ B, & i \ge k, u_i \ne u_{i-k}, \end{cases}$$

where the X_i and B are dummy variables. Only the positions of '0's in $\phi_k(\boldsymbol{u})$ are of importance to us. Let the lengths of the maximal runs of '0's immediately after X_{k-1} and at the end of $\phi_k(\boldsymbol{u})$ be denoted by $l_{\boldsymbol{u}}$ and $r_{\boldsymbol{u}}$, respectively. Note that either of these may be equal to 0. If $\phi_k(\boldsymbol{u}) = X_0^{k-1} 0^{N-k}$, then $l_{\boldsymbol{u}} = r_{\boldsymbol{u}} = N - k$. Otherwise, we have $\phi_k(\boldsymbol{u}) = X_0^{k-1} 0^{l_{\boldsymbol{u}}} Y 0^{r_{\boldsymbol{u}}}$, for some Y that starts and ends with B. For example, for $\boldsymbol{u} = 0100110110$, we have $\phi_3(\boldsymbol{u}) = X_0^2 00B0000$, $l_{\boldsymbol{u}} = 2$, and $r_{\boldsymbol{u}} = 4$.

To enable us to succinctly represent the results, we define several functions. These functions relate $\boldsymbol{u}$ to the frequencies of other substrings that can generate $\boldsymbol{u}$ via appropriate duplication events. First, for $N \ge k$, let

$$G_k^{\boldsymbol{u}}(\boldsymbol{x}) = \sum_z x^{\boldsymbol{u}_0^{z-1} \boldsymbol{u}_{z+k}^{N-1}},$$

where the sum is over all z such that $(\phi_k(\boldsymbol{u}))_z^{z+k-1} = 0^k$. For $\boldsymbol{u} = 0100110110$, $G_3^{\boldsymbol{u}}(\boldsymbol{x}) = 2x^{0100110}$.

Furthermore, for $k > 0$ and $N \ge k+1$, let

$$F_{k,l}^{\boldsymbol{u}}(\boldsymbol{x}) = \sum_{i=1}^{\min(l_{\boldsymbol{u}}, k-1)} x^{\boldsymbol{u}_i^{N-1}}, \; F_{k,r}^{\boldsymbol{u}}(\boldsymbol{x}) = \sum_{i=1}^{\min(r_{\boldsymbol{u}}, k-1)} x^{\boldsymbol{u}_0^{N-1-i}},$$

$$M_k^{\boldsymbol{u}}(\boldsymbol{x}) = \begin{cases} \sum_{b=N-k+1}^{k-1} x^{\boldsymbol{u}_b^{k-1}\boldsymbol{u}_0^{b-1}}, & \text{if } \phi_k(\boldsymbol{u}) = X_0^{k-1}0^{N-k} \\ 0, & \text{else} \end{cases}$$

We use $\mathcal{B}_1(\boldsymbol{u})$ to denote set of strings at Hamming distance 1 from $\boldsymbol{u}$. Also for $\boldsymbol{u}, \boldsymbol{v} \in \Sigma^*$, the indicator function $I(\boldsymbol{u}, \boldsymbol{v})$ equals 1 if $\boldsymbol{u} = \boldsymbol{v}$ and equals 0 otherwise.

III. Evolution of Substring Frequencies

In this section, we first find $\boldsymbol{\delta}_k(\boldsymbol{x}) = (\delta_k^{\boldsymbol{u}}(\boldsymbol{x}))_{\boldsymbol{u}\in U}$ for $k > 0$ (duplication) and then for $k = 0$ (substitution). This will enable us to show that the substring frequencies converge almost surely and find the limit set. We only consider substrings $\boldsymbol{u}$ of length $N > k$ for each k since the frequencies of shorter substrings can be obtained from these. So for the whole model, we can only consider $\boldsymbol{u}$ of length $N > K$.

Theorem 2. *For an integer $k > 0$ and a string $\boldsymbol{u} = u_0u_1\cdots u_{N-1}$, if $k+1 \le N < 2k$, then*

$$\delta_k^{\boldsymbol{u}}(\boldsymbol{x}) = F_{k,l}^{\boldsymbol{u}}(\boldsymbol{x}) + F_{k,r}^{\boldsymbol{u}}(\boldsymbol{x}) + M_k^{\boldsymbol{u}}(\boldsymbol{x}) - (N-1-k)x^{\boldsymbol{u}},$$

and if $N \ge 2k$,

$$\delta_k^{\boldsymbol{u}}(\boldsymbol{x}) = F_{k,l}^{\boldsymbol{u}}(\boldsymbol{x}) + F_{k,r}^{\boldsymbol{u}}(\boldsymbol{x}) + G_k^{\boldsymbol{u}}(\boldsymbol{x}) - (N-1-k)x^{\boldsymbol{u}}.$$

Before proving the theorem, we present two examples for $k = 3$ and $\Sigma = [3]$:

$$\begin{aligned} \delta_3^{1231}(\boldsymbol{x}) &= x^{123} + x^{231} + x^{312} \\ \delta_3^{12312312}(\boldsymbol{x}) &= 3x^{12312} + x^{123123} + x^{1231231} + \\ &\qquad x^{312312} + x^{2312312} - 4x^{12312312} \end{aligned}$$

Proof: Suppose a duplication of length k 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 N-substrings of $\boldsymbol{s}_n$ that are eliminated (do not exist in $\boldsymbol{s}_{n+1}$) and the N-substrings of $\boldsymbol{s}_{n+1}$ that are new (do not exist in $\boldsymbol{s}_n$). Any new N-substring must intersects with both the template and the copy in $\boldsymbol{s}_{n+1}$. Likewise, an eliminated N-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 N-substring that includes neither its leftmost symbol nor its rightmost symbol.

As an example, suppose

$$\boldsymbol{s}_n = \boldsymbol{v}12345678\boldsymbol{w}, \qquad \boldsymbol{s}_{n+1} = \boldsymbol{v}123456\underline{45}678\boldsymbol{w},$$

where $k = 3$, the (new) copy is underlined, and $\boldsymbol{v}, \boldsymbol{w} \in \sum^*$. Let $N = 5$. The new 5-substrings are 34564, 45645, 56456, 64567 and the eliminated substring is 34567. Formally, let

$$\begin{aligned} \boldsymbol{s}_n &= a_0\cdots a_i a_{i+1}\cdots a_{i+k}a_{i+k+1}\cdots a_{|s_n|-1}, \\ \boldsymbol{s}_{n+1} &= a_0\cdots a_i a_{i+1}\ldots a_{i+k}a_{i+1}\ldots a_{i+k}a_{i+k+1}\ldots a_{|s_n|-1}, \end{aligned}$$

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

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

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

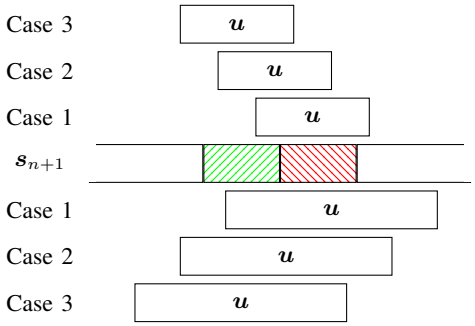

Figure 1. Possible cases for new occurrences of $\boldsymbol{u}$ in $\boldsymbol{s}_{n+1}$. Cases above and below $\boldsymbol{s}_{n+1}$ correspond to $k+1 \le N < 2k$ and $N \ge 2k$, respectively. The hatched boxes, from left to right, are the template and the copy.

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. In the example above, if $\boldsymbol{u} = 45645$, then $\boldsymbol{y}_3 = \boldsymbol{u}$ and thus $Y_3 = 1$.

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

$$\begin{aligned} \delta_k^{\boldsymbol{u}}(\boldsymbol{x}) &= \sum_{b=1}^{N-1} E_k[Y_b|\mathcal{F}_n] - E_k[W|\mathcal{F}_n] \\ &= \sum_{b=1}^{N-1} E_k[Y_b|\mathcal{F}_n] - (N-k-1)x^{\boldsymbol{u}}, \end{aligned}$$

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

To find $\delta_k^{\boldsymbol{u}}$, it suffices to find $E_k[Y_b|\mathcal{F}_n]$, or equivalently, $\Pr(Y_b = 1|\mathcal{F}_n, l = k)$. 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 3–5. Summing over the expressions provided by these lemmas provides the desired result. We omit the details, as well as the proofs of Lemmas 4 and 5 due to similarity to that of Lemma 3. ∎

Lemma 3 (Case 1)**.** *For $1 \le b < \min(k, N-k+1)$,*

$$E_k[Y_b|\mathcal{F}_n] = x^{\boldsymbol{u}_b^{N-1}} I(\boldsymbol{u}_0^{b-1}, \boldsymbol{u}_k^{k+b-1}).$$

Proof: For $1 \le b < \min(k, N-k+1)$ (regardless of $N \ge 2k$ or $N < 2k$), the new occurrences of $\boldsymbol{u}$ always contains 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 $\boldsymbol{u}_0^{b-1} = \boldsymbol{u}_k^{k+b-1}$.

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 $\boldsymbol{u}_b^{N-1}$, which has probability $x^{\boldsymbol{u}_b^{N-1}}$. ∎

Lemma 4 (Case 2)**.** *Suppose* $\min(k, N-k+1) \le b < \max(N-k+1, k)$*. If* $N \ge 2k$*, then*

$$E_k[Y_b|\mathcal{F}_n] = x^{\boldsymbol{u}_0^{b-k-1}\boldsymbol{u}_b^{N-1}} I(\boldsymbol{u}_{b-k}^{b-1}, \boldsymbol{u}_b^{b+k-1}),$$

and if $k+1 \le N \le 2k-2$*, then*

$$E_k[Y_b|\mathcal{F}_n] = x^{\boldsymbol{u}_b^{k-1}\boldsymbol{u}_0^{b-1}} I(\boldsymbol{u}_0^{N-1-k}, \boldsymbol{u}_k^{N-1}).$$

Note that this case cannot occur if $N = 2k-1$.

Lemma 5 (Case 3)**.** *For* $\max(N-k+1, k) \le b \le N-1$*,*

$$E_k[Y_b|\mathcal{F}_n] = x^{\boldsymbol{u}_0^{b-1}} I(\boldsymbol{u}_{b-k}^{N-k-1}, \boldsymbol{u}_b^{N-1}).$$

We now turn our attention to substitutions ($k = 0$).

Theorem 6. *For a string* $\boldsymbol{u}$ *of length* N*, we have*

$$\delta_0^{\boldsymbol{u}} = \frac{1}{|\Sigma|-1} \sum_{\boldsymbol{v} \in \mathcal{B}_1(\boldsymbol{u})} x^{\boldsymbol{v}} - N x^{\boldsymbol{u}}.$$

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

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

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 $x^{\boldsymbol{v}}/(|\Sigma|-1)$. On the other hand, an occurrence of $\boldsymbol{u}$ is eliminated if a substitution occurs in any of its N positions. So the expected number occurrences that vanish is $Nx^{\boldsymbol{u}}$. ∎

IV. ODE and the Limits of Substring Frequencies

Theorems 2 and 6 provide expressions for $\boldsymbol{\delta}_k$ for $0 \le k \le K$. 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, $\boldsymbol{x} = (x^{\boldsymbol{u}})_{\boldsymbol{u} \in U}$, where U is the set of strings of length $N \ge k+1$. The ODE is of the form $\frac{d\boldsymbol{x}_t}{dt} = A\boldsymbol{x}_t$, where A is determined using Theorems 2 and 6 as described in (1). On the right side in expressions for $\boldsymbol{\delta}_k$, terms of the form $x^{\boldsymbol{v}}$ appear where $\boldsymbol{v} \notin U$. However, we have $|\boldsymbol{v}| < N$, so we replace $x^{\boldsymbol{v}}$ with $\sum_{\boldsymbol{w}} x^{\boldsymbol{w}}$, where the summation is over all strings $\boldsymbol{w}$ of length N such that $\boldsymbol{v}$ is a prefix of $\boldsymbol{w}$. For example, consider $q_0 = \alpha$, $q_1 = 1-\alpha$, and $\Sigma = \{0,1\}$. From Theorems 2 and 6, for $\boldsymbol{x} = (x^{00}, x^{01}, x^{10}, x^{11})$, we have $d\boldsymbol{x}/dt = A\boldsymbol{x}$, where

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

Theorem 7. *Consider a tandem duplication and substitution system with distribution* $\boldsymbol{q} = (q_k)$ *over these mutations such that* $q_k = 0$ *for* $k > K$ *and* $q_0 < 1$*. The frequencies of substrings* $\boldsymbol{u}$ *of length* $N \ge K+1$ *converges almost surely to the null space of the matrix* A*, described above.*

Proof: We first show that the resulting ODE is stable. This is done by applying the Gershgorin circle theorem to the columns of A (see e.g., (2)). In each column, the diagonal element is the only element that can be negative. We show that each column of A sums to 0, which 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. Define A_k to be the matrix satisfying $\boldsymbol{h}_k(\boldsymbol{x}) = A_k\boldsymbol{x}$ so that $\boldsymbol{h}(\boldsymbol{x}) = A\boldsymbol{x} = \sum_k q_k \boldsymbol{h}_k(\boldsymbol{x}) = \sum_k q_k A_k \boldsymbol{x}$. For the example of (2), the matrices A_0 and A_1 are

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

We show that each column of A_k sums to zero for each k, which implies the desired result. Fix $\boldsymbol{v} \in U$ and consider the column in A_k that corresponds to $x^{\boldsymbol{v}}$. To identify the elements in this column, we must consider expressions for $h_k^{\boldsymbol{u}}(\boldsymbol{x}) = \delta_k^{\boldsymbol{u}}(\boldsymbol{x}) - k\boldsymbol{x}$ and check if $x^{\boldsymbol{v}}$ appears on the right side. For $k > 0$, the only negative term corresponds to $h_k^{\boldsymbol{v}}$, where the coefficient is $-(N-1)$. Inspecting the proofs of Lemmas 3–5 shows that for each value of $b \in [N-1]$, there is only one $\boldsymbol{u}$ such that $x^{\boldsymbol{v}}$ appears in $h_k^{\boldsymbol{u}}$ with a nonnegative coefficient, and the coefficient is 1. For example, for $b = 1$, from Lemma 3, this $\boldsymbol{u}$ is equal to $v_k v_1^{N-1}$. Since there are $N-1$ possible choices for b, the sum of every column in A_k is 0, as desired. For $k = 0$, we have $h_k^{\boldsymbol{u}}(\boldsymbol{x}) = \delta_k^{\boldsymbol{u}}(\boldsymbol{x})$, where $\delta_k^{\boldsymbol{u}}(\boldsymbol{x})$ is given in Theorem 6. The column corresponding to $x^{\boldsymbol{v}}$ has a negative term equal to $-N$ and $N(|\Sigma|-1)$ positive terms, where each of the positive terms is equal to $\frac{1}{|\Sigma|-1}$, 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. 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.

From the Jordan canonical form theorem, we can write $A = PJP^{-1}$, for an invertible matrix of generalized eigenvectors P. Let $\boldsymbol{y}_t = P^{-1}\boldsymbol{x}_t$, let C be any compact internally chain transitive set of the ODE $\dot{\boldsymbol{y}}_t = J\boldsymbol{y}_t$. From the boundedness of $\boldsymbol{x}_t$, and thus $\boldsymbol{y}_t$, and the fact that all eigenvalues of A except for 0 have negative real parts, we can prove that 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 example, for the matrix A of (2), the vector in the null space whose elements sum to 1, and thus the limit of $\boldsymbol{x}_n$, is

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

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 $\{a\boldsymbol{z}_1 + (1-a)\boldsymbol{z}_2 : 0 \le a \le 1\}$.

V. Bounds on Entropy

The entropy of this process may be upper bounded using techniques from semiconstrained systems [11]–[13]. We first formally define the entropy, and then argue that it is upper bounded by an appropriately defined semiconstrained system.

Consider the string $\boldsymbol{s}_n$, obtained from $\boldsymbol{s}_0$ by n rounds of mutations, as described previously. Its length is $|\boldsymbol{s}_n| = |\boldsymbol{s}_0| + \sum_{i=1}^{n} l_i$, and its expected length is $E[|\boldsymbol{s}_n|] = |\boldsymbol{s}_0| + n\sum_{i=1}^{K} i q_i$. We define the entropy after n rounds as

$$\mathcal{H}_n = -\frac{1}{E[|\boldsymbol{s}_n|]} \sum_{\boldsymbol{w}\in\Sigma^*} \Pr(\boldsymbol{s}_n = \boldsymbol{w}) \log_{|\Sigma|} \Pr(\boldsymbol{s}_n = \boldsymbol{w}),$$

as well as $\mathcal{H}_\infty = \limsup_{n\to\infty} \mathcal{H}_n$.

Let us recall some definitions concerning semiconstrained systems (see [13]). Let $\mathcal{P}(\Sigma^N)$ denote the set of all probability measures on Σ^N. A *semiconstrained system* is defined by $\Gamma \subseteq \mathcal{P}(\Sigma^N)$. The set of admissible words of the semiconstrained system, denoted $\mathcal{B}(\Gamma)$, contains exactly all finite words over the alphabet Σ, whose N-gram distribution is in Γ, and $\mathcal{B}_\ell(\Gamma) = \mathcal{B}(\Gamma) \cap \Sigma^\ell$. An expansion of Γ by $\epsilon > 0$ is defined as

$$\mathbb{B}_\epsilon(\Gamma) = \left\{ \boldsymbol{\xi} \in \mathcal{P}(\Sigma^N) \ : \ \inf_{\boldsymbol{\nu}\in\Gamma} \|\boldsymbol{\nu} - \boldsymbol{\xi}\|_{\mathrm{TV}} \le \epsilon \right\},$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total-variation norm. The capacity of Γ is then defined as

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

which intuitively measures the information per symbol in strings whose N-gram distribution is "almost" in Γ.

Theorem 8. *For the mutation process described above, $\mathcal{H}_\infty \le \mathsf{cap}(\Gamma)$, where Γ is the compact connected internally chain transitive invariant set that $\boldsymbol{x}_n$ converges almost surely to by Theorem 1.*

Remark 9. We comment that if $\Gamma = \{\boldsymbol{\xi}\}$, i.e., Γ contains a single shift-invariant measure[1], then $\mathsf{cap}(\Gamma)$ has a nice form (see [11], [13]):

$$\mathsf{cap}(\Gamma) = -\sum_{a_1\ldots a_N \in \Sigma^N} \boldsymbol{\xi}^{a_1\ldots a_N} \log_{|\Sigma|} \frac{\boldsymbol{\xi}^{a_1\ldots a_N}}{\bar{\boldsymbol{\xi}}^{a_1\ldots a_{N-1}}},$$

where $\bar{\boldsymbol{\xi}}$ is the marginal of $\boldsymbol{\xi}$ on the first $N-1$ coordinates, i.e., $\bar{\boldsymbol{\xi}}^{a_1\ldots a_{N-1}} = \sum_{b\in\Sigma} \boldsymbol{\xi}^{a_1\ldots a_{N-1}b}$.

We use the preceding remark to find an upper bound on the system whose limit is given by (3). We have $\bar{\boldsymbol{\xi}}^0 = \bar{\boldsymbol{\xi}}^1 = 1/2$. It then follows that for this system $\mathcal{H}_\infty \le H_2\left(\frac{2\alpha}{1+3\alpha}\right)$, where H_2 is the binary entropy function.

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

VI. Conclusion

In this paper, we provided a method for determining the limits of N-gram frequencies (as substrings of the evolving sequence) for tandem duplications and substitutions. We also presented a method for finding upper bounds on the entropy of these systems. One direction for extending the current work is including other mutation types that are present in tandem repeat regions such as deletions and insertions. Open problems include quantifying the finite-time behavior in these systems, determining the rate of convergence to the limits, and lower bounds on entropy.

References

[1] 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.

[2] 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.

[3] F. Farnoud, M. Schwartz, and J. Bruck, "The Capacity of String-Duplication Systems", *IEEE Trans. Information Theory*, vol. 62, no. 2, pp. 811–824, Feb. 2016.

[4] S. Jain, F. Farnoud, and J. Bruck, "Capacity and Expressiveness of Genomic Tandem Duplication", *IEEE Trans. Information Theory*, vol. 63, no. 10, Oct. 2017.

[5] 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.

[6] 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.

[7] ——, "Estimating Mutation Rates under a Stochastic Model for Tandem Duplication and Substitution", in *In Preparation,* http://www.people.virginia.edu/~ffh8x/d/smtd.pdf.

[8] A. O. SCHMITT and H. HERZEL, "Estimating the Entropy of DNA Sequence", *Journal of Theoretical Biology*, vol. 188, no. 3, pp. 369–377, Oct. 1997.

[9] 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.

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

[11] O. Elishco, T. Meyerovitch, and M. Schwartz, "Semiconstrained systems", *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 1688–1702, 2016.

[12] ——, "On encoding semiconstrained systems", *IEEE Transactions on Information Theory*, 2017, To appear.

[13] ——, "On independence and capacity of multidimensional semiconstrained systems", *arXiv preprint arXiv:1709.05105*, 2017.