Finite-Time Behavior of k-mer Frequencies and Waiting Times in Noisy-Duplication Systems

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

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

***Abstract*—Mutations play a significant role in evolution since they lead to genomic diversity. Among different types of mutations, duplication is thought to be one of the most important. Motivated by the theory of evolution by duplication, we consider a stochastic model for the evolution of sequences under noisy tandem duplication, where segments of the sequences are replicated and approximate copies are added to the sequence. Our goal is to study the statistical properties of the sequence after a given number of mutations. To do so, we study the k-mer frequencies of the evolving sequence. We first bound the expected frequencies of different k-mers after n mutations and relate the convergence rate of the expected trajectories to the parameters of the model (probabilities of different mutations). Then we extend our analysis to second moments of the k-mer trajectories, which allow us to better characterize their evolution. Finally, we will demonstrate the application of the proposed methods to bounding waiting times, the first such results for complex mutation systems.**

I. Introduction

The vast amount of biological diversity is, for the most part, the result of genomic mutations. One of the most influential theories about the generation of new genetic material, which is also well-supported by data [1], is Ohno's evolution by duplication [2]. Ohno hypothesizes that the origins of new genes are copies of existing sequences, which have diverged from the original to acquire new functions. Since the copies are not essential for the organism's survival, they are under less evolutionary pressure to be conserved and are more free to mutate. Ohno's theory is supported by the abundance of repetitive elements; for example the majority of the human genome consists of repeated sequences [3].

In order to gain a better understanding of the generation processes of new genetic sequences, which is dominated by duplications, we consider a simple stochastic model for the generation of new sequences through noisy tandem duplications. In our model, in each step, a substring is randomly chosen and its copy is inserted in tandem. The copy however is not always exact and may differ from the original. We study how the frequencies of different k-mers (strings of length k) change as more mutations occur. k-mer frequencies are of interest since they allow us to determine the substrings that are likely to be generated under different model parameters. Thus their analysis can provide a way to test hypothesis regarding models of evolution. Furthermore, they can provide bounds on the entropy, and thus the compressibility, of sequences [4].

Stochastic tandem duplication models have been studied, for example, to analyze microsatellites [5], to estimate mutation probabilities [6], and to estimate entropies [4], [7]. While the analyses in these works are concerned with limit sets, in this paper, we are interested in finite-time behavior. We first present a linear recurrence for the expected k-mer trajectories. While the recurrence cannot be solved explicitly, we establish bounds on the expected trajectories and provide the rate of convergence. We then extend our analysis to the second-moment of the k-mer trajectories, characterizing the variation around expected paths.

We then turn our attention to estimating waiting times. The waiting time for a given string $\boldsymbol{u}$ in an evolutionary system is the first time index in which $\boldsymbol{u}$ appears as a substring of the evolving sequence. Waiting time problems are of interest since appearances of new patterns in DNA sequences lead to new biological functions and changes in physical attributes [1]. Furthermore, accumulated alterations in certain types of genes, including oncogenes, tumor suppressor genes and genetic instability genes, are known to be responsible for tumorigenesis [8]. Thus, understanding the time scales in which such events take place is of importance in explaining evolutionary trends and the study of diseases such as cancer [9].

A variety of waiting time problems have been studied in the literature, e.g., in [9], [10]. In these works, however, the sequence evolution model consists of independent mutations at the nucleotide or gene level. For example [9] considers a segment of L nucleotides and allows each to mutate independently and uniformly. The type of sequence evolution model assumed in these works simplifies the analysis but ignores the possibility of complex mutations, such as duplications, that lead to dependence among sequence positions. Here, we provide upper and lower bounds on waiting time CDFs, which, to the best of our knowledge, are the first such results for any type of mutation other that iid substitutions.

After presenting the preliminaries in Section II, we study the expected behavior of k-mer frequencies in Section III and present upper bounds on the CDF of waiting times in Section IV. The simplicity of the first-order analysis also allows us to analytically study the effect of mutation probabilities

This work was supported in part by NSF grants under grant nos. CCF-1816409, CCF-1755773, CCF-1908544.

on waiting times. We study the second-moment behavior in Section V, provide lower bounds on the CDF of waiting times in Section VI, and conclude the paper in Section VII.

II. Preliminaries and Notation

For a positive integer n, we use $[n]$ to denote the set $\{1,\dots,n\}$. We denote the alphabet and all finite strings over the alphabet by Σ and Σ^*, respectively, while Σ^k denotes the set of all strings of length k, i.e., k-mers, over Σ. For $\boldsymbol{w},\boldsymbol{w}'\in\Sigma^*$, the concatenation of $\boldsymbol{w}$ and $\boldsymbol{w}'$ is denoted $\boldsymbol{w}\boldsymbol{w}'$. The set of strings at Hamming distance d from $\boldsymbol{w}$ is denoted $\mathcal{B}_d(\boldsymbol{w})$. Moreover, we use $|\boldsymbol{w}|$ to denote the length of $\boldsymbol{w}$. Vectors and strings are denoted by boldface letters such as $\boldsymbol{w}$, while scalars and symbols by normal letters, such as w.

We now describe a stochastic string system representing the evolution of a string under random mutations. Consider an initial string $\boldsymbol{s}_0$ and a process where in each step n a random transform, or "mutation", M_n, is applied to $\boldsymbol{s}_n$, resulting in $\boldsymbol{s}_{n+1}$. Let $\mathcal{M}$ be the set of all possible mutations. To each $m\in\mathcal{M}$ we assign a certain probability. We are, in particular, interested in noisy duplication mutations. For integers $d\geq 0$ and $\ell\geq 1$, the *noisy duplication* $\mathcal{T}_\ell^d:\Sigma^*\to\Sigma^*$ is defined as

$$\forall \boldsymbol{w}\in\Sigma^*,\quad \mathcal{T}_\ell^d(\boldsymbol{w})=\boldsymbol{u}\boldsymbol{a}\boldsymbol{a}'\boldsymbol{v},$$

where $\boldsymbol{a}$ is a substring of $\boldsymbol{w}$ of length ℓ chosen uniformly at random, $\boldsymbol{u}$ and $\boldsymbol{v}$ are strings such that $\boldsymbol{w}=\boldsymbol{u}\boldsymbol{a}\boldsymbol{v}$, and $\boldsymbol{a}'\in\mathcal{B}_d(\boldsymbol{a})$, chosen uniformly at random.

In a noisy duplication system, the set of permitted mutations is $\mathcal{M}=\{\mathcal{T}_\ell^d:\ell'\leq\ell\leq\ell'',0\leq d\leq\ell\}$, where $\ell',\ell''\in\mathbb{Z}_{>0}$. The mutation in step n is chosen to be $\mathcal{T}_\ell^d$ with probability q_ℓ^d, independently of other steps. That is, for each n, $\mathbb{P}(M_n=\mathcal{T}_\ell^d)=q_\ell^d$. Recall that conditioned on $M_n=\mathcal{T}_\ell^d$, a substring $\boldsymbol{a}$ of length ℓ is randomly chosen and an approximate copy $\boldsymbol{a}'\in\mathcal{B}_d(\boldsymbol{a})$ of $\boldsymbol{a}$ is inserted into the string. The distribution over mutations, conditioned on a mutation occurring, is denoted by $\boldsymbol{q}=\left(q_\ell^d\right)_{\ell,d}$.

We denote the length of $\boldsymbol{s}_n$ by L_n and let $\ell_n=L_n-L_{n-1}$. Generally, since a noisy duplication $\mathcal{T}_\ell^d$ adds length ℓ to the evolving string, $\{\ell_n\}$ is a sequence of iid random variables whose distribution is determined by $\boldsymbol{q}$. In this paper, however, we only study systems with a fixed duplication length ℓ, i.e., systems which permit mutations $\mathcal{M}=\{\mathcal{T}_\ell^d:0\leq d\leq\ell\}$, for some $\ell\in\mathbb{Z}_{>0}$, and leave more complex systems to future work.

For a string $\boldsymbol{u}\in\Sigma^*$, denote the number of appearances of $\boldsymbol{u}$ in $\boldsymbol{s}_n$ as $\mu_n^{\boldsymbol{u}}$, and its frequency as $x_n^{\boldsymbol{u}}=\mu_n^{\boldsymbol{u}}/L_n$, where $\boldsymbol{s}_n$ is interpreted as a circular string to avoid complications arising from boundaries. For example, if $\boldsymbol{s}_n=\mathsf{ACGAC}$, then $\mu_n^{\mathsf{AC}}=2, x_n^{\mathsf{AC}}=\frac{2}{5}$. For any ordered set $U\subseteq\Sigma^*$, 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 $\boldsymbol{\mu}_n$.

Let $\tau_{\boldsymbol{u}}(m)$ be the smallest n such that the sequence $\boldsymbol{s}_n$ contains m occurrences of $\boldsymbol{u}$ and, as shorthand, let $\tau_{\boldsymbol{u}}=\tau_{\boldsymbol{u}}(1)$.

III. Expected Behavior in Noisy Duplication String Systems

In this section, we study the expected trajectory of $\boldsymbol{x}_n=(x_n^{\boldsymbol{u}})_{\boldsymbol{u}\in\Sigma^k}$. Our first step is expressing $\mathbb{E}[\boldsymbol{x}_n]$ as a recurrence.

Theorem 1. *In a noisy duplication string system with distribution $\boldsymbol{q}$ and with $\boldsymbol{x}_n=(x_n^{\boldsymbol{u}})_{\boldsymbol{u}\in\Sigma^k}$, we have*

$$\mathbb{E}[\boldsymbol{x}_{n+1}]-\mathbb{E}[\boldsymbol{x}_n]=A\mathbb{E}[\frac{\boldsymbol{x}_n}{L_{n+1}}] \tag{1}$$

for some square matrix A whose elements are determined by $\boldsymbol{q}$ and k, and whose eigenvalues have non-positive real parts.

The matrix A is called the *characteristic matrix* of the system for k-mers. In [4], [6] this theorem is proved for similar systems and thus its proof is omitted here. There, the theorem is used as part of a stochastic approximation framework to find almost-sure limit sets/points for $\boldsymbol{x}_n$ as $n\to\infty$. The matrix A can be computed through a method similar to [4].

In the next theorem, we consider the case in which A is diagonalizable. Let $m=|\Sigma|^k$ so that $A\in\mathbb{R}^{m\times m}$. We pick a set V consisting of m linearly independent eigenvectors of A as a basis for $\mathbb{R}^m$. Having chosen V as a basis, we study the expected trajectory of $\boldsymbol{x}_n$ by representing $\boldsymbol{x}_n$ in this basis and then by bounding the coefficients of this representation. Our analysis in this paper is limited to cases in which A has only real eigenvalues. In all examples that we have studied, the eigenvalues of A are indeed real. We conjecture that this holds for all noisy duplication systems.

Let $V=\{\boldsymbol{v}_s:1\leq s\leq m\}$ and let λ_s be the corresponding eigenvalue of $\boldsymbol{v}_s$, $1\leq s\leq m$. For each $n\geq 0$, we write $\boldsymbol{x}_n=\sum_{s=1}^m\alpha_n^s\boldsymbol{v}_s$ for some $\alpha_n^s\in\mathbb{R}$. We provide upper and lower bounds on each α_n^s.

Theorem 2. *Consider a noisy tandem duplication system with mutations $\mathcal{M}=\{\mathcal{T}_\ell^d:0\leq d\leq\ell\}$, where $\ell\in\mathbb{Z}_{>0}$, and characteristic matrix A. If A is diagonalizable with real eigenvalues, such that $\lambda_s\geq\frac{-L_0}{2}$ for all $1\leq s\leq m$, then:*

1) *For $1\leq s\leq m$ such that $\lambda_s=0$ or $\alpha_0^s=0$,*

$$\mathbb{E}[\alpha_n^s]=\alpha_0^s\quad\forall n\in\mathbb{N}.$$

2) *For $1\leq s\leq m$ such that $\lambda_s\neq 0$ and $\alpha_0^s\neq 0$,*

$$T_n^s<\frac{\mathbb{E}[\alpha_n^s]}{\alpha_0^s}<U_n^s, \tag{2}$$

where

$$U_n^s=(\frac{\lambda_s+L_n}{\lambda_s+L_1})^{\frac{\lambda_s}{\ell}}e^{\lambda_s^2/(L_1\ell)}(1+\frac{\lambda_s}{L_n}),$$
$$T_n^s=(\frac{\lambda_s+L_n}{\lambda_s+L_1})^{\frac{\lambda_s}{\ell}}e^{\lambda_s^2/(L_n\ell)}(1+\frac{\lambda_s}{L_1}),$$

and $L_n=L_0+n\ell$.

Proof: From (1), we have

$$\sum_{s=1}^m\mathbb{E}[\alpha_n^s]\boldsymbol{v}_s=\sum_{s=1}^m\mathbb{E}[\alpha_{n-1}^s]\boldsymbol{v}_s+\sum_{s=1}^m\mathbb{E}[\alpha_{n-1}^s]\frac{A}{L_n}\boldsymbol{v}_s.$$

For a given s, we thus find

$$\mathbb{E}[\alpha_n^s] = \left(1 + \frac{\lambda_s}{L_n}\right)\mathbb{E}[\alpha_{n-1}^s] = \alpha_0^s e^{\sum_{i=1}^n f(i)},$$

where

$$f(i) = \log\left(1 + \frac{\lambda_s}{L_0 + i\ell}\right).$$

It is clear that if $\lambda_s = 0$ or $\alpha_0^s = 0$, then $\mathbb{E}[\alpha_n^s] = \alpha_0^s$. It remains to prove the bounds on $\mathbb{E}[\alpha_n^s]$ when $\alpha_0^s \neq 0$ and $\lambda_s < 0$. We note that $f'(i) = \ell(\frac{1}{L_0+i\ell+\lambda_s} - \frac{1}{L_0+i\ell})$, which is positive and continuous for $i \in [1, \infty)$ since $\lambda_s < 0$. It can be shown by applying Euler's summation formula [11] that

$$\sum_{i=1}^n f(i) < \int_1^n f(x)dx + f(n).$$

For the integral, we have

$$\int_1^n f(x)dx = \frac{1}{\ell}\left[\lambda_s \log\left(\frac{\lambda_s + L_n}{\lambda_s + L_1}\right) + L_n \log\left(1 + \frac{\lambda_s}{L_n}\right) - L_1 \log\left(1 + \frac{\lambda_s}{L_1}\right)\right].$$

Since $x - x^2 \leq \log(1+x) \leq x$ for $-1/2 \leq x \leq 0$, we obtain

$$\log\left(1 + \frac{\lambda_s}{L_n}\right) \leq \frac{\lambda_s}{L_n}, -\log\left(1 + \frac{\lambda_s}{L_1}\right) \leq \left(\frac{\lambda_s}{L_1}\right)^2 - \frac{\lambda_s}{L_1},$$

and therefore

$$\int_1^n f(x)dx \leq \frac{\lambda_s}{\ell}\left[\log\left(\frac{\lambda_s + L_n}{\lambda_s + L_1}\right) + \frac{\lambda_s}{L_1}\right].$$

The desired result follows from $f(n) = \log(1 + \lambda_s/L_n)$. The proof for the lower bound is similar. ∎

Recall that $\lambda_s \leq 0$. From the theorem, the behaviors of both U_n^s and T_n^s are dominated by $(\frac{L_n+\lambda_s}{L_1+\lambda_s})^{\frac{\lambda_s}{\ell}}$, which is $\Theta(n^{\frac{\lambda_s}{\ell}})$ as $n \to \infty$. This implies that when $\mathbb{E}[\boldsymbol{x}_n]$ is represented in an eigenbasis, for an eigenvector whose corresponding eigenvalue λ is not 0, its component in $\mathbb{E}[\boldsymbol{x}_n]$ converge to 0 at the rate of $n^{\lambda/\ell}$, while for eigenvectors whose corresponding eigenvalues are 0, their components remains unchanged and determine the expected value of the limit of $\boldsymbol{x}_n$. Among all nonzero eigenvalues, the largest one determines the convergence rate.

Similar, but more involved, results can be found when A is not diagonalizable. However, in all systems that we have investigated, the characteristic matrices are diagonalizable. We conjecture that this is always the case, i.e., the characteristic matrix of a noisy duplication systems is always diagonalizable. In the rest of the paper, we are only concerned with diagonalizable characteristic matrices.

Example 1. We demonstrate the bounds given in the theorem for a simple noisy duplication system over the alphabet $\{0, 1\}$ with following parameters

$$\mathcal{M} = \{\mathcal{T}_1^0, \mathcal{T}_1^1\},\ \boldsymbol{q} = (q_1^0, q_1^1) = (1 - \delta, \delta). \tag{3}$$

Specifically, in each step, a random symbol a is chosen uniformly from the evolving string, and a symbol b is inserted immediately after a, where $q_1^0 = \mathbb{P}(b{=}a) = 1 - \delta$ and $q_1^1 = \mathbb{P}(b{\neq}a) = \delta$. The vector $\boldsymbol{x}_n = (x_n^{00}, x_n^{01}, x_n^{10}, x_n^{11})^T$ denotes the frequencies of 2-mers. The characteristic matrix of this system for 2-mers can be shown to be

$$A = \begin{bmatrix} -2\delta & 1-\delta & \delta & 0 \\ \delta & -1 & 0 & \delta \\ \delta & 0 & -1 & \delta \\ 0 & \delta & 1-\delta & -2\delta \end{bmatrix}. \tag{4}$$

For $\delta < 1/2$, A is diagonalizable with four distinct eigenvalues: $\lambda_1 = 0, \lambda_2 = -2\delta, \lambda_3 = -1, \lambda_4 = -2\delta - 1$. So we pick a basis of $\mathbb{R}^4$ which is composed of 4 eigenvectors $\boldsymbol{v}_i, 1 \leq i \leq 4$, corresponding to $\lambda_i, 1 \leq i \leq 4$, respectively:

$$\begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 & \boldsymbol{v}_4 \end{bmatrix} = \begin{bmatrix} 1/(2+4\delta) & -1 & -1 & 1 \\ \delta/(1+2\delta) & 0 & 1 & -1 \\ \delta/(1+2\delta) & 0 & -1 & -1 \\ 1/(2+4\delta) & 1 & 1 & 1 \end{bmatrix}. \tag{5}$$

Since $\lambda_1 = 0$, the limit of $\mathbb{E}[\boldsymbol{x}_n]$ is $\boldsymbol{v}_1$ (it can be shown that $\boldsymbol{x}_n$ converges to $\boldsymbol{v}_1$ almost surely) and the other components vanish at the corresponding rates. Figure 1 shows this for the coefficient α_n^2 when $q_1^1 = 0.1, q_1^0 = 0.9$, and $\boldsymbol{s}_0 = 1101100111$. The figure illustrates the upper bound, lower bound, and the expected value as n ranges from 0 to 50. The average value for α_n^2 from 5000 independent trials of the process is also given.

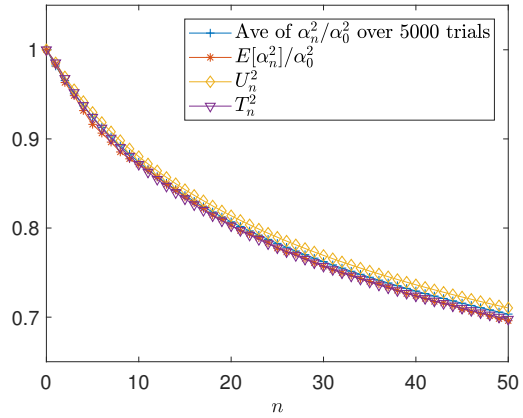


Figure 1: Coefficient of $\boldsymbol{v}_2$ in $\mathbb{E}[\boldsymbol{x}_n]$ vs the number of mutations in a noisy duplication string system with $\Sigma = \{0, 1\}$, $\boldsymbol{s}_0 = 1101100111$, $q_1^1 = 0.9$, $q_1^0 = 0.1$.

IV. Bounding Waiting Times by Mean Trajectories

In this section, we derive bounds on the waiting times for the appearances of k-mers based on the behavior of average trajectories characterized by Theorem 2. This will enable us to quantify the effect of mutation probabilities on waiting times.

Let $\hat{n}_{\boldsymbol{u}}$ be such that $\mathbb{E}[\mu_n^{\boldsymbol{u}}] \simeq 1$ for $n = \hat{n}_{\boldsymbol{u}}$. If $\mathbb{E}[\mu_n^{\boldsymbol{u}}]$ is increasing, for $n \ll \hat{n}_{\boldsymbol{u}}$, the probability of the existence of $\boldsymbol{u}$ in $\boldsymbol{s}_n$ is small and thus we can view $\hat{n}_{\boldsymbol{u}}$ as a rough estimate for $\tau_{\boldsymbol{u}}$. More precisely, as shown below, under certain conditions, the probability that $\tau_{\boldsymbol{u}} \leq \hat{n}/M$ scales as $1/M$.

A comparison of the expected waiting times and $\hat{n}$ for an example system is given in Figure 2, where expected waiting

times are obtained by averaging over simulation trials and $\hat{n}_{11}$ and $\hat{n}_{12}$ are calculated using (1). It can be seen that $\hat{n}_{12}$ has a much smaller relative error. This is due to the higher variance of μ_n^{11}, which is discussed further in Section VI.

Theorem 3. *For $\boldsymbol{u} \in \Sigma^k$, if $\mathbb{E}[x_n^{\boldsymbol{u}}]$ is non-decreasing in n,*

$$\mathbb{P}(\tau_{\boldsymbol{u}} \leq n) \leq (1 + \max(k - \ell, 0))\mathbb{E}[\mu_n^{\boldsymbol{u}}].$$

The theorem provides an upper bound on the CDF of $\tau_{\boldsymbol{u}}$. It can also be extended to provide upper bounds for the CDF of $\tau_{\boldsymbol{u}}(m)$. We omit the proof of the theorem. We demonstrate an analytical application of the theorem via an example in this section and a computational application in Section VI.

Consider a system with parameters given in (3) but over the alphabet $\Sigma = \{0, 1, 2\}$. Let $\boldsymbol{s}_0$ be a string consisting of 0's. In this case, we can actually have a slightly stronger result, namely, $\mathbb{P}(\tau_{\boldsymbol{u}} \leq n) \leq (1 + \delta)\mathbb{E}[\mu_n^{\boldsymbol{u}}]$, with a similar proof. We can study analytically how the waiting time for 12 varies as $\delta \to 0$. Let $\boldsymbol{x}_n$ be the vector of 2-mer frequencies. After finding the characteristic matrix A, from Theorem 2,

$$\mathbb{E}[x_n^{12}] = \frac{1}{2}\delta^2(C_n - 1 + \log\frac{L_n}{L_1} + \frac{1}{L_n}) + O(\delta^3). \quad (6)$$

where

$$\frac{(L_1 - 1)^2}{(L_n - 1)L_1} e^{\frac{1}{L_n}} < C_n < \frac{L_1 - 1}{L_n} e^{\frac{1}{L_1}}.$$

For $\hat{n} = \hat{n}_{12} = \frac{2/\delta^2}{\log(2/\delta^2)}$, we have $\mathbb{E}[\mu_{\hat{n}}^{\boldsymbol{u}}] = 1 + o(1)$ and, furthermore, for a constant $M > 1$, $\mathbb{P}(\tau_{12} \leq \frac{\hat{n}}{M}) \leq \frac{1+o(1)}{M}$. Hence, $\frac{2/\delta^2}{M\log(2/\delta^2)}$ is a lower bound for τ_{12} that holds with probability at least $1/M$.

V. Second-Order Analysis

In this section, we present two theorems for computing the variance of k-mer frequencies $x_n^{\boldsymbol{u}}$ for all $\boldsymbol{u} \in \Sigma^k$.

Theorem 4. *Consider a noisy duplication system as in Theorem 2. For a k-mer $\boldsymbol{v}$, denote its index in $\boldsymbol{x}_n$ by $i_{\boldsymbol{v}}$. For any two k-mers $\boldsymbol{v}, \boldsymbol{w} \in \Sigma^k$ (not necessarily distinct), we have*

$$\mathbb{E}[x_{n+1}^{\boldsymbol{v}} x_{n+1}^{\boldsymbol{w}}] - \left(\frac{L_n}{L_{n+1}}\right)^2 \mathbb{E}[x_n^{\boldsymbol{v}} x_n^{\boldsymbol{w}}] = \frac{\boldsymbol{d}_{\boldsymbol{v},\boldsymbol{w}}^T}{(L_{n+1})^2}\mathbb{E}[\boldsymbol{y}_n] + \left(\frac{L_n}{L_{n+1}}\right)^2 \left(\frac{1}{L_n}\mathbb{E}[x_n^{\boldsymbol{w}} B_{i_{\boldsymbol{v}}}^T \boldsymbol{x}_n] + \frac{1}{L_n}\mathbb{E}[x_n^{\boldsymbol{v}} B_{i_{\boldsymbol{w}}}^T \boldsymbol{x}_n]\right) \quad (7)$$

where $B = A + \ell I$, $B_{i_{\boldsymbol{v}}}$ and $B_{i_{\boldsymbol{w}}}$ are the $i_{\boldsymbol{v}}$-th and the $i_{\boldsymbol{w}}$-th row of B, respectively, $\boldsymbol{y}_n$ is the vector of the frequencies of the $(2k-2)$-mers, and $\boldsymbol{d}_{\boldsymbol{v},\boldsymbol{w}}$ is a constant vector independent of n.

Note that both $B_{i_{\boldsymbol{v}}}^T \boldsymbol{x}_n$ and $B_{i_{\boldsymbol{w}}}^T \boldsymbol{x}_n$ in (7) are linear functions of k-mer frequencies. Based on this fact, we have the following theorem.

Theorem 5. *In any given noisy duplication system with k-mer frequencies $\boldsymbol{x}_n$, let $\boldsymbol{r}_n = (\mathbb{E}[x_n^{\boldsymbol{v}} x_n^{\boldsymbol{w}}])_{\boldsymbol{v},\boldsymbol{w} \in \Sigma^k}$. We have*

$$\begin{pmatrix} \boldsymbol{r}_{n+1} \\ \mathbb{E}[\boldsymbol{y}_{n+1}] \end{pmatrix} = \begin{bmatrix} \left(\frac{L_n}{L_{n+1}}\right)^2 \left(I + \frac{G}{L_n}\right) & \frac{D}{(L_{n+1})^2} \\ \mathbf{0} & I + \frac{A_{\boldsymbol{y}}}{L_{n+1}} \end{bmatrix} \begin{pmatrix} \boldsymbol{r}_n \\ \mathbb{E}[\boldsymbol{y}_n] \end{pmatrix}, \quad (8)$$

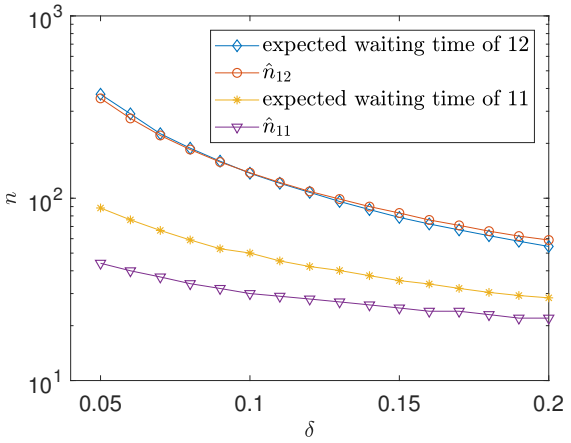


Figure 2: Expected waiting times for 11 and 12 and $\hat{n}_{11}, \hat{n}_{12}$ vs δ for $\Sigma = \{0, 1, 2\}$, $\boldsymbol{s}_0 = 0000000000, q_1^0 = 1 - \delta, q_1^1 = \delta$.

where G, L are both constant matrices uniquely determined by the system, and $A_{\boldsymbol{y}}$ is the characteristic matrix of this system for $(2k-2)$-mers.

We omit the proof of Theorems 4 and 5 and the explicit forms of matrices $G, D, \boldsymbol{d}_{\boldsymbol{v},\boldsymbol{w}}$ due to space limitation. Note that using (8), $\boldsymbol{r}_n$ can be computationally determined. Indeed, $\boldsymbol{r}_n$ can also be written as a product of matrices and the same approach as in Theorem 2 can be taken to derive analytical bounds on $\boldsymbol{r}_n$. However, we leave this to future work and provide examples demonstrating the computational application of the theorem in the next section.

Figure 3a compares the variance of x_n^{12} computed using (8) with the sample variance of 10000 independent trials in the system with parameters given by (3) with $\Sigma = \{0, 1, 2\}$, $\boldsymbol{s}_0 = 0000000000$, and $\delta = 0.2$. Note that since x_n^{12} is bounded, the variance cannot increase unbounded. Indeed, if $x_n^{\boldsymbol{u}}$ converges to a single point, the variance vanishes.

VI. Bounding Waiting Times by Second Moments

In this section, we derive bounds on the waiting time using the first- and second-order analyses of Sections III and V. We first use Chebyshev's inequality to find the range where most of the trajectories lie and then infer bounds on the waiting time.

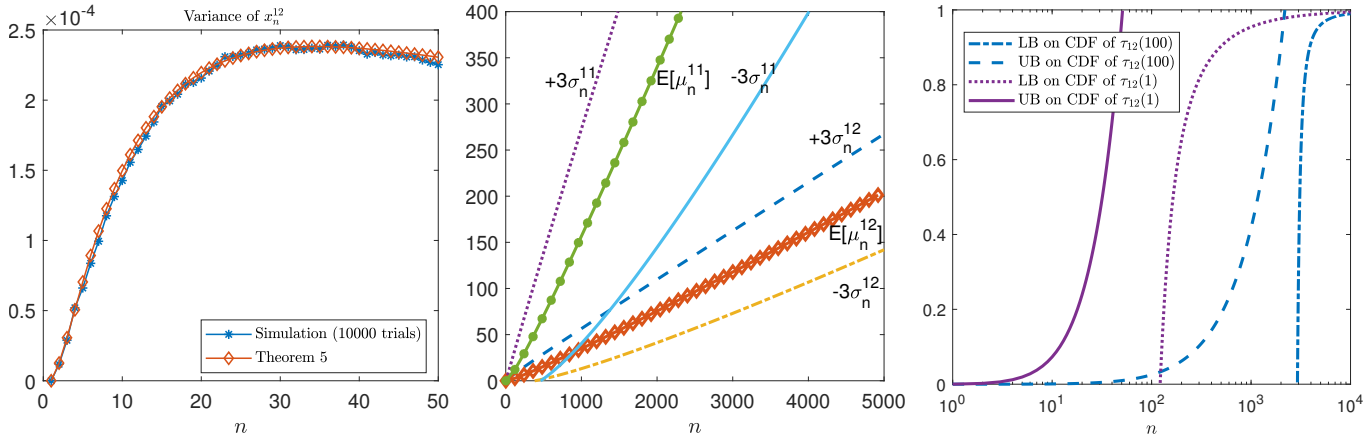
Let $\boldsymbol{u} \in \Sigma^k$ be a k-mer and $\sigma_n^{\boldsymbol{u}}$ be the standard deviation of $\mu_n^{\boldsymbol{u}}$. By Chebyshev's inequality,

$$\mathbb{P}(\mathbb{E}[\mu_n^{\boldsymbol{u}}] - \gamma\sigma_n^{\boldsymbol{u}} \leq \mu_n^{\boldsymbol{u}} \leq \mathbb{E}[\mu_n^{\boldsymbol{u}}] + \gamma\sigma_n^{\boldsymbol{u}}) > 1 - \frac{1}{\gamma^2}, \quad (9)$$

for any $\gamma > 0$. Therefore, by computing the expected value and variance of $\mu_n^{\boldsymbol{u}}$ using Theorems 1, 4 and 5, we can bound $\mu_n^{\boldsymbol{u}}$ in a range that contains most of the probability mass.

Note that for any positive integer n, $\mu_n^{\boldsymbol{u}} \geq m$ is a sufficient (but not necessary) condition for $\tau_{\boldsymbol{u}}(m) \leq n$. Hence, from (9),

$$\mathbb{P}(\tau_{\boldsymbol{u}}(m) \leq n) \geq \mathbb{P}(\mu_n^{\boldsymbol{u}} \geq m) > 1 - \frac{1}{\gamma^2}, \quad (10)$$



(a) Variance of x_n^{12} vs the number of mutations.

(b) Expected values and $\pm 3\sigma$ range for μ_n^{12} and μ_n^{11}, which contain $8/9$ of the probability.

(c) Bounds on $\mathbb{P}(\tau_{12}(m) \leq n)$ for $m = 1, 100$. LB and UB stand for lower and upper bounds.

Figure 3: Second-order analysis for a noisy duplication string system with $\Sigma = \{0, 1, 2\}$, $\boldsymbol{s}_0 = 0000000000$, $q_1^1 = 0.8$, $q_1^0 = 0.2$.

where $\gamma = (\mathbb{E}[\mu_n^{\boldsymbol{u}}] - m)/\sigma_n^{\boldsymbol{u}}$. This tells us that we are likely to see m occurrences of $\boldsymbol{u}$ in the sequence not long after the expected number of occurrences of $\boldsymbol{u}$ hits m, thus providing a lower bound on the CDF $\mathbb{P}(\tau_{\boldsymbol{u}}(m) \leq n)$ and a probabilistic upper bound on $\tau_{\boldsymbol{u}}(m)$.

Example 2. Consider the noisy duplication system with parameters given by (3) and $\Sigma = \{0, 1, 2\}$. Let $\boldsymbol{s}_0 = 0000000000, \gamma = 3, \delta = 0.2$. Figure 3b provides an interval for μ_n^{11} and μ_n^{12} that has probability at least $8/9$. We can observe that the variance of μ_n^{11} is much larger than that of μ_n^{12}. This is in agreement with Figure 2, where it is observed that $\hat{n}_{\boldsymbol{u}}$, obtained based on average trajectories, better matches the waiting time for $\boldsymbol{u} = 12$ compared to $\boldsymbol{u} = 11$. The higher variance of μ_n^{11} is likely due to its high autocorrelation, which means that as soon as an instance of 1 is created, many instances of 11 can be produced by duplicating it. When variance is high, (9) will lead to loose or trivial bounds.

Figure 3c illustrates lower and upper bounds on the CDF of $\tau_{12}(m)$ for $m = 1, 100$, where the upper bounds are based on Section IV. The sharpness of the curves in the figure implies that in fact, most of the probability of $\tau_{12}(m)$ is concentrated in a small interval. In particular, the bounds provide the order of magnitude of the waiting times.

VII. Conclusion

We studied the finite-time behavior of noisy duplication string systems by representing the average trajectories of the frequencies of k-mers in an eigenbasis of the characteristic matrix of the system. We showed the coordinate corresponding to eigenvalue $\lambda \neq 0$ converges to 0 with rate approximately $n^{\lambda/\ell}$. We also provided a method for computing the second moment of k-mer frequencies, as well as bounds on the CDFs of waiting times, which are the first such bounds for any type of mutation other than independent substitution.

References

[1] D. Tautz and T. Domazet-Lošo, "The evolutionary origin of orphan genes", *Nature Reviews Genetics*, vol. 12, no. 10, Oct. 2011.

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

[3] E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome", *Nature*, vol. 409, no. 6822, 2001.

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

[5] S. Kruglyak *et al.*, "Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations", *Proc. Nat. Academy of Sciences*, vol. 95, no. 18, Sep. 1998.

[6] F. Farnoud, M. Schwartz, and J. Bruck, "Estimation of duplication history under a stochastic model for tandem repeats", *BMC Bioinformatics*, vol. 20, no. 1, 2019.

[7] F. Farnoud, M. Schwartz, and J. Bruck, "A Stochastic Model for Genomic Interspersed Duplication", in *Proc. IEEE Int. Symp. Information Theory*, Hong Kong, China, Jun. 2015.

[8] B. Vogelstein and K. W. Kinzler, "Cancer genes and the pathways they control", *Nature medicine*, vol. 10, no. 8, 2004.

[9] R. Durrett and D. Schmidt, "Waiting for Two Mutations: With Applications to Regulatory Sequence Evolution and the Limits of Darwinian Evolution", *Genetics*, vol. 180, no. 3, Nov. 2008.

[10] M. Gerstung and N. Beerenwinkel, "Waiting time models of cancer progression", *Mathematical Population Studies*, vol. 17, no. 3, 2010.

[11] T. M. Apostol, *Introduction to analytic number theory*. Springer Science & Business Media, 2013.