Estimation of the Duplication History under a Stochastic Model for Tandem Repeats

Farzad Farnoud (Hassanzadeh), Moshe Schwartz, Jehoshua Bruck

***Abstract*—We present a stochastic model for tandem duplication and substitution mutations that can be used to estimate relative mutation rates and the total number of mutations in a tandem repeat sequence. Parameters of the model include the probability of a substitution mutation and the probabilities of tandem duplications of various lengths. Our model indicates that if the probability of substitution mutations is insignificant, little information can be obtained from a sequence (which has undergone only tandem duplication). On the other hand, in the presence of both substitution and tandem duplication, one can estimate the parameters of the model, using which we can estimate the number of mutations of each type. We validate our estimation method via Monte Carlo simulation and show that it outperforms the state-of-the-art algorithm for discovering the duplication history. We also apply our method to tandem repeat sequences in the human genome, where it demonstrates the different behaviors of micro- and mini-satellites and can be used to compare mutation rates across chromosomes.**

I. Introduction

Tandem repeats, which form about 3% of the human genome [1], are segments of DNA that primarily consist of repeats of a certain pattern. The number of copies in tandem repeats is highly variable and is prone to change due to tandem duplication mutations. Furthermore, tandem repeats are subject to point mutations [2]. The variability of tandem repeats enables them to be used in population genetics [3] as well as human identity testing [4]. Tandem repeats may cause expansion diseases, gene silencing [5], and rapid morphological variation [6].

A mechanisms suggested for the formation of tandem repeat sequences, especially those of shorter lengths, is slipped-strand mispairing [7], also known as replication slippage [8]. This mechanism refers to the misalignment of the template and the nascent strand during DNA replication. It is thought that the presence of similar or identical sequences, which may occur by chance in the first place, increases the probability of misalignment [7].

In this work, we present and analyze a model of the evolution of tandem repeat sequences via tandem duplication and substitution mutations. The starting point is a short sequence which we refer to as the *seed*. At each mutation step, either a tandem duplication or a substitution mutation occurs, each with a given probability. Furthermore, tandem duplications of different lengths do not necessarily have the same probability. We show analytically that certain statistical features of the sequence converge as the number of mutations increases. This in turn allows us to i) predict the behavior of the sequence after a large number of mutations if we have the parameters of the model, or ii) estimate the parameters of the model given the sequence after a large number of mutations. In other words, given a sequence that is the result of the aforementioned process, we can estimate conditional mutation probabilities without any other information or comparison with homologous sequences from other organisms.

We study two cases in the evolution of tandem repeats. First, we consider the case in which substitution mutations do not occur and the only type of mutation is tandem duplication. We show that in this case, while the prediction of evolutionary behavior is easy, estimation of model parameters, including the probabilities of tandem duplications of given lengths, is difficult. This is because as the number of mutations increases, the sequence demonstrates periodic behavior, lacking features that can be leveraged for estimation. Perhaps surprisingly, the period of this sequence is not necessarily the most common or the shortest possible tandem duplication length.

We then consider the more interesting case in which both tandem duplication and substitution mutations occur. In this case, substitutions disrupt the periodic pattern that would arise from tandem duplications. As a result, after a large number of mutations, the resulting sequence is more complex and informative, allowing us to estimate the model parameters. Specifically, from such a sequence, we can estimate the probability of a substitution in each step, as well as the probabilities of tandem duplications of different lengths. Furthermore, we can estimate the total number of mutations that gave rise to the sequence under study. We apply this method to the tandem repeats in the human genome, which enables us to investigate the prevalence of substitutions in repeats of different lengths and to compare the average number of mutations among chromosomes. We show that two classes of tandem repeats are observed based on their mutation profiles and that this classification is compatible with the mini- and micro-satellite classification based on pattern length. Furthermore, our analysis illustrates that the average numbers of mutations in some chromosomes are higher than others. Interestingly, this agrees with another measure of mutation activity, i.e., comparison with the chimpanzee genome: The chromosomes with higher mutation counts in repeated regions are the same as the ones that have diverged most from chimpanzee chromosomes. Our results demonstrate that the proposed estimation method can be used to study various aspects of tandem repeat sequences, such as the effects of different factors on mutation rates, at a large scale.

In previous work on modeling tandem duplication and substitution mutations, it is often assumed that in each step, the length of the sequence grows by at most one repeat unit, which simplifies the analysis; see, e.g., [9] and references therein. Our model however allows duplications of lengths longer than one repeat unit at a time. Note that models

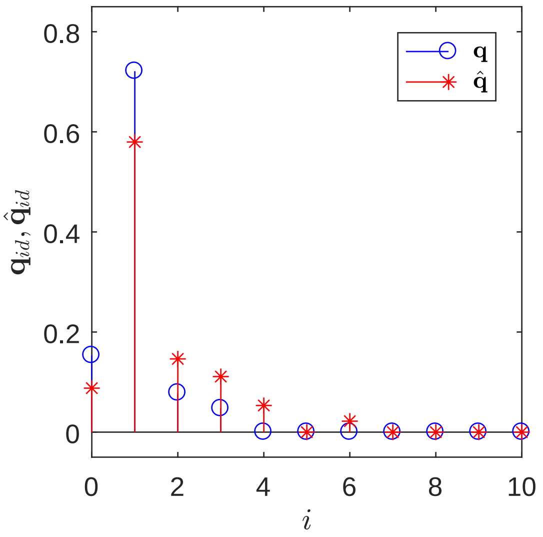


Figure 1: An example of estimating $\mathbf{q}$ as $\hat{\mathbf{q}}$.

that do not allow longer duplications may underestimate the probability of substitution and overestimate the probability of tandem duplication since more duplication events are needed to account for the observed copy number. Models proposed in [10] and [11] include duplications of lengths longer than one repeat unit. But these works only consider perfect tandem repeats, in which all copies are identical. Imperfect tandem repeats, however, are common in genomic data. Furthermore, unlike [10] and [9] that use Markov chains and branching processes for modeling, our analysis is based on stochastic approximation, which enables the description of new aspects of the problem. In particular, we see that the observed period in a tandem repeat sequence is not necessarily the most common duplication length (Theorem 1) and that the presence of substitutions allows the estimation of mutation probabilities ((14)). Finally, these papers are not concerned with recovering the duplication history, which is a focus of the current paper.

Recovering the duplication history has been studied in [12], [13], [14], which take a combinatorial approach to solving the problem. Via simulation, we show that the method proposed in this paper outperforms the state-of-the-art method, called DTSCORE [14]. Estimation of the duplication history using a stochastic model, to the best of our knowledge, has not appeared in the literature before.

The Model and Examples of Estimation: Before describing the methods in more detail, we present the proposed model and an example of the estimation of conditional mutation rates. Consider a sequence over some alphabet $\mathcal{A}$ that evolves over time through tandem duplication and substitution mutations. The starting sequence, called the *seed*, is denoted by $s^{(0)}$. In each step i, a mutation converts the sequence $s^{(i-1)}$ to $s^{(i)}$.

If the mutation occurring at time i is a substitution, its position is chosen at random among all symbols of $s^{(i)}$. That symbol is then changed randomly to one of the other symbols of $\mathcal{A}$. If the mutation is a tandem duplication of length ℓ, a substring of length ℓ is chosen uniformly at random, duplicated, and inserted in tandem. We use q_0 to denote the probability that the mutation in any given step is a substitution and q_ℓ, $\ell > 0$, to denote that it is a tandem duplication of length ℓ. We assume that there exists K such that $q_\ell = 0$ for all $\ell > K$. Finally, we let $\mathbf{q} = (q_0, q_1, \dots, q_K)$ where $\sum_{i=0}^{K} q_i = 1$. Note that $\mathbf{q}$ represents conditional mutation probabilities given that a mutation occurs and not the mutation probabilities per generation.

Let $\mathcal{A} = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$. As an example, suppose that $s^{(i-1)} = \mathsf{TC\underline{ACA}GT}$, and that in step i the underlined substring of length $\ell = 3$ is duplicated in tandem. The result equals $s^{(i)} = \mathsf{TC\underline{ACA}\overline{ACA}GT}$, where the inserted copy of ACA is over-lined. If in the next step a substitution occurs, the result may be $s^{(i+1)} = \mathsf{T\underline{G}ACAACAGT}$.

In the estimation task, one is given the sequence $s^{(n)}$ and from this sequence only, she must recover the values of $\mathbf{q}$ and n. Postponing the details to Section III.III-A, we present two example of estimating these parameters to make the task clearer. In the first example, the sequence $s^{(n)}$ is obtained via simulation by letting $s^{(0)} = \mathsf{ACAAGATGC}$, $q_0 = 0.1533$, $q_d = 0.7212$, $q_{2d} = 0.0784$, $q_{3d} = 0.0471$ for $d = 9$ (these values are chosen randomly). We randomly apply $n = 100$ mutations to $s^{(0)}$. The result $s^{(100)} = \mathsf{ACAAGATGCACAAGA}\cdots$ along with d is then passed to the estimator (described is Section II.II-C and referred to as SMTR). This provides us with the estimate $\hat{n}$ of n as 64.53 and the estimate $\hat{\mathbf{q}}$ of $\mathbf{q}$ as in Figure 1. Note that the only information given to the estimator is the sequence $s^{(100)}$ and d. It can be observed that $\hat{n}$ and $\hat{\mathbf{q}}$ are reasonable estimates of n and q. Furthermore, as we show later, the accuracy of the estimates increases as the number n of mutations grows.

The next example illustrates the estimation of parameters for a segment of the human genome, namely Chromosome 1: 933,911–935,015 (Ensembl), given in Figure 2. For this sequence, SMTR gives the probability of substitution as $\hat{q}_0 = 0.3687$, and the probabilities of duplications of lengths i as $(\hat{q}_{28i})_{i=1}^{5} = (0.1499, 0.1257, 0.0232, 0.2656, 0.0669)$. All other $\hat{q}_i$ are estimated to be 0. Furthermore, the number of mutations is estimated as $\hat{n} = 17.67$ out of which 11.15 are tandem duplications and 6.51 are substitutions.

The general analysis of the model is presented in the next section. This analysis is specialized to the cases with and without substitutions in Sections II.II-B and II.II-C, respectively. Simulation results of our estimation method, as well as its application to the human genome, are given in Section III. The paper is concluded in Section IV.

II. Modeling and Estimation Method

We will first present an overview of our method. Our approach relies on designing a stochastic model for the evolution of tandem repeats in the presence of tandem duplication and substitution mutations. Assuming the parameters of the model (the conditional probabilities of duplication and substitution mutations) are known, we study the asymptotic behavior of tandem repeat sequences. This analysis is based on the autocorrelation function since this feature well represents the

```
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTCCGTTACAGGTGGGCAGGGGAGGCG
GCTCCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCT GCTCCGTTACAGGTGGGCGGGGGAGGCT
GCTCCGTTACAGGTGGGCGGGGGGGGCG GCTGCGTTACAGGTGGGCGGGGGGGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTCCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCAGGGGAGGCG GCTGCGTTACAGGTGGGCAGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTCCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGAGGCG GCTGCGTTACAGGTGGGCGGGGGAGGCG
GCTGCGTTACAGGTGGGCGGGGGGGGCG GCTGCGTTACAGGTGGGCGGGCGG
```

Figure 2: An example of a tandem repeat sequence from the human genome.

(approximate) periodicity that results from duplication and substitution mutations. We determine the limit set of the autocorrelation function as a function of model parameters. We will then address the inverse problem of estimating the parameters given a sequence, assuming that its autocorrelation is close to the limit. This in turn enables us to estimate the counts of mutations of different types in the history of the sequence.

While this paper is focused on tandem duplication and substitution mutations, the method proposed for the analysis of stochastic models of mutations is more broadly applicable. For example, it is extendable to other types of mutations and features other than autocorrelation. It is thus a versatile method for extracting information from genomic data.

A. General Analysis via Stochastic Approximation

In this section, we first present a general framework for the problem of analyzing the evolution of sequences under duplication and substitution mutation processes using stochastic approximation, which relates the problem to an ordinary differential equation (ODE). We then specialize this method to the analysis of the autocorrelation function in systems with tandem duplication and substitution.

Let s be a circular sequence over some alphabet $\mathcal{A}$ that "evolves" over time. The process starts with $s^{(0)}$ and in each step, $s^{(i)}$ is obtained from $s^{(i-1)}$ through a random mutation. The reason that we choose s to be a circular string, and not a linear one, is to avoid the technical difficulties of dealing with its boundaries. In our notation $s^{(i)}$ is the instance of s at time i. However, if it causes no ambiguity, we may use s instead of $s^{(i)}$. We use L_i to denote the length of $s^{(i)}$.

For an ordered set U, let $\boldsymbol{R}_n = (R_n^u)_{u\in U}$ be a vector representing the number of appearances of objects $u \in U$ in the sequence s at time n and let $\boldsymbol{\rho}_n = \frac{\boldsymbol{R}_n}{L_n}$ be the normalized version of $\boldsymbol{R}_n$. For example, U can be the set of all strings over $\mathcal{A}$ with length at most three. Our goal is to find out how $\boldsymbol{\rho}_n$ changes with n by finding a differential equation whose solution approximates $\boldsymbol{\rho}_n$.

Define $\mathcal{F}_n$ to be the filtration generated by the random variables $\{\boldsymbol{\rho}_n, L_n\}$. Furthermore, let $\mathbb{E}_\ell[\,\cdot\,]$ denote the expected value conditioned on the fact that the length of the duplicated substring is ℓ and let $\boldsymbol{\delta}_\ell = \mathbb{E}_\ell[\boldsymbol{R}_{n+1}|\mathcal{F}_n] - \boldsymbol{R}_n$. Recall that q_0 is the probability of a substitution and $q_i, i > 0$ is the probability of the event that a sequence of length $\ell = i$ is duplicated. The following set of conditions must be satisfied for our analysis. Among them, we assume (A1). It is easy to see that (A2)-(A3) are true, and (A4)-(A5) will be evident after we find $\boldsymbol{\delta}_\ell$.

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

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

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

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

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

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

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

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

$$\begin{aligned}
\mathbb{E}\big[\boldsymbol{\rho}_{n+1} - \boldsymbol{\rho}_n \big| \mathcal{F}_n\big] &= \sum_{\ell=0}^{K} q_\ell \big(\mathbb{E}_\ell\big[\boldsymbol{\rho}_{n+1}\big|\mathcal{F}_n\big] - \boldsymbol{\rho}_n\big)\\
&= \sum_{\ell=0}^{K} q_\ell \left(\frac{\boldsymbol{R}_n + \boldsymbol{\delta}_\ell(\boldsymbol{\rho}_n)}{L_n + \ell} - \frac{\boldsymbol{R}_n}{L_n}\right)\\
&= \frac{1}{L_n}\sum_{\ell=0}^{K} q_\ell \boldsymbol{h}_\ell(\boldsymbol{\rho}_n)\big(1 + O\big(L_n^{-1}\big)\big)\\
&= \frac{1}{L_n}\boldsymbol{h}(\boldsymbol{\rho}_n)\big(1 + O\big(L_n^{-1}\big)\big),
\end{aligned} \tag{2}$$

where $\boldsymbol{h}_\ell(\boldsymbol{\rho}) = \boldsymbol{\delta}_\ell(\boldsymbol{\rho}) - \ell\boldsymbol{\rho}$, $\boldsymbol{h}(\boldsymbol{\rho}) = \sum_{\ell=0}^{K} q_\ell \boldsymbol{h}_\ell(\boldsymbol{\rho})$, and where we have used $1/(L_n + \ell) = \big(1 + O\big(L_n^{-1}\big)\big)/L_n$ which follows from the boundedness of ℓ (see (A1)).

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

$$\begin{aligned}
\boldsymbol{\rho}_{n+1} - \mathbb{E}\big[\boldsymbol{\rho}_{n+1}\big|\mathcal{F}_n\big] &= \frac{\boldsymbol{R}_{n+1}}{L_{n+1}} - \mathbb{E}\left[\frac{\boldsymbol{R}_{n+1}}{L_{n+1}}\middle|\mathcal{F}_n\right]\\
&= \frac{1 + O\big(L_n^{-1}\big)}{L_n}\left(\boldsymbol{R}_{n+1} - \mathbb{E}[\boldsymbol{R}_{n+1}|\mathcal{F}_n]\right)\\
&= \frac{1}{L_n}\big(1 + O\big(L_n^{-1}\big)\big)\boldsymbol{M}_{n+1}
\end{aligned} \tag{3}$$

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

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

The aforementioned assumptions and analysis ensure that this system can be analyzed through stochastic approximation [15, Theorem 2], by relating the discrete system describing $\boldsymbol{\rho}_n$ to a continuous system. In particular, the sequence

$\{\boldsymbol{\rho}_n\}$ converges almost surely to a compact connected internally chain transitive invariant set of the ODE

$$\frac{d\boldsymbol{\rho}_t}{dt} = \boldsymbol{h}(\boldsymbol{\rho}_t). \tag{4}$$

While different properties of the sequence can be analyzed via the aforementioned method, for our purpose, the autocorrelation of the sequence is the most suitable. The autocorrelation function R^r of a sequence $s = s_1 \cdots s_{|s|}$, $s_i \in \mathcal{A}$, at lag r, is defined as

$$R^r = \sum_{i=1}^{|s|} \langle s_i, s_{i+r} \rangle,$$

where each index of s should be replaced by its representative (modulo $|s|$) in the residue system $\{1, \dots, |s|\}$, and where $\langle \alpha, \beta \rangle = 1$ iff $\alpha = \beta$ and $\langle \alpha, \beta \rangle = 0$ otherwise.

Let R_n^r denote the autocorrelation of function after n mutations starting from the seed sequence and let $\rho_n^r = \frac{R_n^r}{L_n}$. We study the vectors $\boldsymbol{R}_n = \left(R_n^0, R_n^1, \dots, R_n^{m-1}\right)$ and $\boldsymbol{\rho}_n = \frac{\boldsymbol{R}_n}{L_n}$, for a constant m. Note that $R_n^0 = L_n$ and $\rho_n^0 = 1$.

Let $s = s^{(n)} = s_1 \cdots s_{|s|}$ be the sequence at time n. At time $n+1$, either a substitution or a duplication has happened. In the former case, suppose the symbol at position i is changed to another symbol of the alphabet, and in the latter case, suppose that the substring $s_{i+1} \cdots s_{i+\ell}$ is duplicated in a tandem manner; after duplication the sequence becomes

$$s_1 \cdots s_i s_{i+1} \cdots s_{i+\ell} s_{i+1} \cdots s_{i+\ell} s_{i+\ell+1} \cdots s_{|s|}.$$

Fix the value of i. For $\ell = 0$, i.e., the case of a substitution,

$$R_{n+1}^r = R_n^r - \langle s_i, s_{i+r} \rangle - \langle s_i, s_{i-r} \rangle + \langle s_i', s_{i+r} \rangle + \langle s_i', s_{i-r} \rangle,$$

where s_i' denote the new (mutated) symbol.

Now we consider the case of $\ell > 0$, which corresponds to tandem duplications. For $0 < \ell \le r$, we have

$$R_{n+1}^r = R_n^r - \sum_{j=i+\ell-r+1}^{i} \langle s_j, s_{j+r} \rangle + \sum_{j=i+\ell-r+1}^{i+\ell} \langle s_j, s_{j+r-\ell} \rangle.$$

The conditions resulting in the first summation $j \le i$ and $j + r > i + \ell$ and those resulting in the second summation are $i < j \le i + \ell$ or $i + \ell < j + r \le i + 2\ell$. For $\ell > r$,

$$R_{n+1}^r = R_n^r + \sum_{j=i+1}^{i+\ell-r} \langle s_j, s_{j+r} \rangle + \sum_{j=i+\ell-r+1}^{i+\ell} \langle s_j, s_{j+r-\ell} \rangle.$$

Let $\boldsymbol{h}_\ell(\boldsymbol{\rho}_n) = \left(h_\ell^0(\boldsymbol{\rho}_n), \dots, h_\ell^{m-1}(\boldsymbol{\rho}_n)\right) = \mathbb{E}_\ell[\boldsymbol{R}_{n+1} - \boldsymbol{R}_n | \mathcal{F}_n] - \ell \boldsymbol{\rho}_n$. To compute $\boldsymbol{h}_\ell$, we first find the following expected values, where $r > 0$ and where i is randomly and uniformly distributed among the L_n positions:

$$\begin{aligned}
\mathbb{E}_0[\langle s_i, s_{i+r} \rangle | \mathcal{F}_n] &= \mathbb{E}_0[\langle s_i, s_{i-r} \rangle | \mathcal{F}_n] = \rho_n^r, \\
\mathbb{E}_0[\langle s_i', s_{i+r} \rangle | \mathcal{F}_n] &= \mathbb{E}_0[\langle s_i', s_{i-r} \rangle | \mathcal{F}_n] = \frac{1-\rho_n^r}{3}, \\
\mathbb{E}_\ell\left[\sum_{j=i+\ell-r+1}^{i} \langle s_j, s_{j+r} \rangle \middle| \mathcal{F}_n\right] &= \frac{1}{L_n} \sum_{i=1}^{L_n} \sum_{j=i+\ell-r+1}^{i} \langle s_j, s_{j+r} \rangle \\
&= (r-\ell)\rho_n^r, \\
\mathbb{E}_\ell\left[\sum_{j=i+\ell-r+1}^{i+\ell} \langle s_j, s_{j+r-\ell} \rangle \middle| \mathcal{F}_n\right] &= \frac{1}{L_n} \sum_{i=1}^{L_n} \sum_{j=i+\ell-r+1}^{i+\ell} \langle s_j, s_{j+r-\ell} \rangle \\
&= r\rho_n^{r-\ell}, \\
\mathbb{E}_\ell\left[\sum_{j=i+1}^{i+\ell-r} \langle s_j, s_{j+r} \rangle \middle| \mathcal{F}_n\right] &= \frac{1}{L_n} \sum_{i=1}^{L_n} \sum_{j=i+1}^{i+\ell-r} \langle s_j, s_{j+r} \rangle \\
&= (\ell - r)\rho_n^r.
\end{aligned}$$

From these we find that

$$h_\ell^r(\boldsymbol{\rho}) = \begin{cases} -\frac{8}{3}\rho^r + \frac{2}{3}, & \ell = 0 \\ r\rho^{r-\ell} - r\rho^r, & \ell > 0 \end{cases} \tag{5}$$

and

$$\frac{d}{dt}\rho_t^r = q_0\left(-\frac{8}{3}\rho_t^r + \frac{2}{3}\right) + r\sum_{\ell>0} q_\ell \rho_t^{r-\ell} - (1-q_0) r \rho_t^r \tag{6}$$

for $0 < r < m-1$. We thus see that the set of equations governing $\boldsymbol{\rho}$ are linear.

For $m \ge K$, we can write (6) as

$$\frac{d}{dt}\boldsymbol{\rho}_t = A\boldsymbol{\rho}_t, \tag{7}$$

where A is the $m \times m$ matrix whose rows and columns are indexed by $\{0, 1, \dots, m-1\}$ and its elements are given as

$$A_{rj} = \begin{cases} 2q_0/3 + rq_r, & \text{if } r > j = 0, \\ rq_{r-j} + rq_{r+j}, & \text{if } r > j > 0, \\ q_0\left(r - \frac{8}{3}\right) + rq_{2r} - r, & \text{if } r = j > 0, \\ rq_{r+j}, & \text{if } j > r > 0, \\ 0, & r = 0. \end{cases} \tag{8}$$

For example, if $m = 3, K = 2$, then

$$A = \begin{pmatrix} 0 & 0 & 0 \\ \frac{2q_0}{3} + q_1 & -\frac{5q_0}{3} + q_2 - 1 & 0 \\ \frac{2q_0}{3} + 2q_2 & 2q_1 & -\frac{2q_0}{3} - 2 \end{pmatrix}. \tag{9}$$

To determine the stability of (7) we use the Gershgorin circle theorem. We note that $\sum_j A_{rj} = -2q_0$ and that $A_{rr} = -\frac{8q_0}{3} - r(1 - q_0 - q_{2r})$ for $r > 0$. The circles centered at A_{rr} and with radius $\sum_j A_{rj} - A_{rr} = \frac{2q_0}{3} + r(1 - q_0 - q_{2r})$ in the complex plane either do not intersect the right half of the plane or they intersect it only at 0. Hence, by the Gershgorin circle theorem, the eigenvalues of A are either 0 or have negative real parts. Let $(\lambda_j)_{j=0}^{m-1}$ denote the eigenvalues of A, with $\lambda_0 = 0$ and $\lambda_j = a_j + \imath b_j$ for $j > 0$, where $a_j < 0$ and $\imath$ denotes $\sqrt{-1}$. For $0 \le r \le m-1$, we have

$$\rho_t^r = c_{r0}(t) + \sum_{j=1}^{m-1} c_{jk}(t) e^{a_j t + \imath b_j t},$$

where $c_{rj}(t)$ are polynomials in t of degree at most m. Since $\rho^r(t)$ is bounded between 0 and 1, it is evident that $c_{r0}(t)$ is in fact a constant. Let this constant be denoted by ρ^r_∞. We thus have

$$\rho^r_t = \rho^r_\infty + \sum_{j=1}^{m-1} c_{jk}(t) e^{a_j t + \imath b_j t}, \tag{10}$$

which implies that $\left(\rho^0_t, \dots, \rho^{m-1}_t\right)$ converges to $\boldsymbol{\rho}_\infty = \left(\rho^0_\infty, \dots, \rho^{m-1}_\infty\right)$. Note that $\rho^0_\infty = 1$.

From (10), we have $\lim_{t\to\infty} \frac{d}{dt}\rho^r_t = 0$. By taking the limit of the equation $\frac{d}{dt}\boldsymbol{\rho}_t = A\boldsymbol{\rho}_t$ as $t \to \infty$, it follows that

$$A\boldsymbol{\rho}_\infty = 0,$$

implying that $\boldsymbol{\rho}_\infty$ is in the null space of A. It can then be shown that $\boldsymbol{\rho}_n$ converges almost surely to the null space of A.

In the following sections, we consider the null space of A in two cases. First, we assume $q_0 = 0$, that is, there are no substitutions. Next we study the case with positive probability of substitutions, i.e., $q_0 > 0$.

B. Tandem Duplication

In this section, we consider the case in which the only type of occurring mutations is tandem duplication. We show that in this case the null space of A is simple.

Theorem 1. *Suppose $q_0 = 0$. Let $P = \{i : i > 0, q_i > 0\}$ and $d = \gcd P$. The normalized autocorrelation $\boldsymbol{\rho}_n = \left(\rho^0_n, \dots, \rho^{m-1}_n\right)$ converges almost surely to a vector $\boldsymbol{\rho}_\infty = \left(\rho^0_\infty, \dots, \rho^{m-1}_\infty\right)$, where ρ^j_∞ is periodic in j with period d, $\rho^j_\infty = 1$ if $j \equiv 0 \pmod d$, and $\rho^j_\infty = \rho^{d-j}_\infty$. In particular, every pair of symbols at distance d in $s^{(n)}$ are, with high probability, the same.*

The theorem implies that regardless of the seed, after many duplications, the sequence becomes almost periodic with period d. The periodicity is expected since no substitutions occur. However, the period is not the dominant or the shortest duplication length, but rather it is the gcd of all lengths i for which the probability of duplication q_i is positive. For example, if duplications of lengths 4 and 6 occur, the sequence becomes approximately periodic with period 2. Since given P, d does not depend on the values of the q_i, observing d does not provide enough information for estimating $\mathbf{q}$ and thus, in this case, we are not able to solve the inverse problem. Nevertheless, the study of this case lays the foundation for the more complex case in which substitutions are present and where we are able to solve the inverse problem.

To prove Theorem 1, we first prove the following lemma.

Lemma 2. *Let $q_0 = 0$, $P = \{i > 0 : q_i > 0\}$, and $d = \gcd P$. Furthermore, let $S(t) = \operatorname{Span}\{\boldsymbol{v}_0, \dots, \boldsymbol{v}_{\lfloor t/2 \rfloor}\}$, where $\boldsymbol{v}_i = (v_{i,0}, \dots, v_{i,m-1})^T$, with*

$$v_{i,j} = \begin{cases} 1, & j \equiv \pm i \pmod t, \\ 0, & \textit{otherwise.} \end{cases}$$

We have $\operatorname{Null}(A) = S(d)$.

Proof: Let $B = (B_{rj})$ be an $m \times m$ matrix, with rows and columns indexed by $0, \dots, m-1$, defined by

$$B_{rj} = \begin{cases} q_r, & \text{if } j = 0 \\ q_{r-j} + q_{r+j}, & \text{if } 0 < j < r \\ q_{2r} - 1, & \text{if } j = r \\ q_{r+j}, & \text{if } j > r \end{cases} \tag{11}$$

Since $q_0 = 0$, we have $\operatorname{Null}(B) = \operatorname{Null}(A)$. We further recall that $q_i \in \mathbb{R}$, $q_i \in [0,1]$, and $\sum_i q_i = 1$. Additionally, assume $i_1 < i_2 < \cdots < i_k$ are the only indices for which $q_{i_j} > 0$. Finally, we assume m is large enough to enable us to see all the nonzero q_i's in the matrix, or more formally, we require $m \geq i_k$.

We are interested in finding the null-space of B. Instead of doing this directly, we consider the matrix

$$A' = I + B.$$

The goal now is to find the right eigenspace of A' for the eigenvalue 1.

First we prove $S(d) \subseteq \operatorname{Null}(B)$. We do this by showing that for all $\boldsymbol{v}_i \in S(d)$, $\boldsymbol{v}_i$ is in the right eigenspace of A' corresponding to the eigenvalue 1, i.e., $A'\boldsymbol{v}_i = \boldsymbol{v}_i$. This is immediate when we note that when $i \equiv \pm a \pmod d$, in the ith row of A' (numbering of rows and columns starts from 0), coordinates $j \equiv \pm a \pmod d$ contain all the elements q_{id}, and in particular, $q_{i_1}, q_{i_2}, \dots, q_{i_k}$.

It is obvious that the vectors in $S(d)$ are linearly independent. To complete the proof we need to show that the geometric multiplicity of the eigenvalue 1 of A' is $|S(d)| = \lfloor d/2 \rfloor + 1$.

The matrix A' is stochastic, and therefore, its spectral radius is $\rho(A') = 1$. Let $G_{A'}$ be the (weighted) directed graph whose adjacency matrix is given by A'. By Perron-Frobenius theory, it is well known that the eigenvalues of A' are the union (in the multiset sense) of the eigenvalues of the irreducible components of $G_{A'}$. Additionally, the geometric multiplicity of $\rho(A')$ (also called the Perron-Frobenius (PF) eigenvalue of A') is 1 for each irreducible component.

Combining the above, and remembering the PF eigenvalue of an irreducible graph is a weighted average of the out-weight of its vertices, we obtain that the geometric multiplicity of $\rho(A') = 1$ is exactly the number of irreducible sink components of $G_{A'}$. Thus, as a final step in the proof, we show that the number of irreducible sink components of $G_{A'}$ is exactly $\lfloor d/2 \rfloor + 1$.

Let us denote the vertices of the graph $G_{A'}$ by $w_0, w_1, \dots, w_{m-1}$. From each w_ℓ, $\ell > 0$, we have k out-going edges corresponding to $i_1, i_2, \dots, i_k$. The edge corresponding to i_j is directed from w_ℓ to $w_{\ell - i_j}$ when $\ell \geq i_j$, and otherwise to $w_{i_j - \ell}$. When describing a path we shall refer to this edge as "taking i_j from w_ℓ". Finally, vertex w_0 has a single out-going edge which is also a self-loop.

By construction, all vertices w_ℓ for $\ell \geq i_k$ have incoming edges from vertices $w_{\ell'}$ with $\ell' > \ell$. Thus, they are certainly not part of an irreducible sink component. We therefore concentrate on vertices $w_0, w_1, \dots, w_{i_k - 1}$ only. We now look at the irreducible components over these vertices.

For each i_j, starting from w_ℓ, $0 < \ell < i_k$, we can take a path representing the orbit of $-i_j$ modulo i_k. If $\ell \geq i_j$, then we can move from w_ℓ to $w_{\ell - i_j}$. If $\ell < i_j$ we can also do this but we require two steps: from w_ℓ to $w_{i_j - \ell}$ by taking an i_j step, and then from $w_{i_j-\ell}$ to $w_{i_k-(i_j-\ell)}$ by taking an i_k step. Indeed

$$i_k - (i_j - \ell) \equiv \ell - i_j \pmod{i_k}.$$

Since we can do this for every i_j every node w_ℓ is connected to every node $w_{\ell'}$ with $\ell \equiv \ell' \pmod{d}$. Also, by taking the i_k edge from each node, we can see that from every node w_ℓ we can reach every node $w_{\ell'}$ with $\ell \equiv \pm \ell' \pmod{d}$.

The only exception to the above are nodes w_ℓ with $\ell \equiv 0 \pmod{d}$ since they get "stuck" at w_0, which is an irreducible sink component on its own. We therefore reach the conclusion that there are exactly $\lfloor d/2 \rfloor + 1$ irreducible sink components which are of the form

$$V_a = \{ w_\ell \mid 0 < \ell < i_k, \ell \equiv \pm a \pmod{d} \},$$

for $0 < a \leq d/2$, as well as $V_0 = \{w_0\}$. ■

Proof of Theorem 1: Since $\boldsymbol{\rho}_\infty$ is in the null space of A, where the null space of A is given by Lemma 2, $\boldsymbol{\rho}_\infty$ is a linear combination of the vectors $S(d)$. Furthermore, by definition we know that $\rho^0_\infty = 1$. In the basis of $S(d)$ given in Lemma 2, the only vector that has a nonzero element in the 0th coordinate is

$$v_{0,j} = \begin{cases} 1, & j \equiv 0 \pmod{d}, \\ 0, & \text{otherwise.} \end{cases}$$

So the coefficient of $\boldsymbol{v}_0$ in the linear combination describing $\boldsymbol{\rho}_\infty$ is 1 and thus $\rho^j_\infty = 1$ if $j \equiv 0 \pmod{d}$. We hence have Theorem 1. ■

C. Tandem Duplication and Substitution

We now consider both tandem duplication and substitution mutations and describe how the parameters of the model, as well as the number of mutations of each type, may be estimated. Note that while the parameters of the model are unknown, we have access to the sequence $s^{(n)}$ for some n.

We show that the autocorrelation function converges to a single point when duplication and substitution mutations are present. This fact, proved in the following lemma, will facilitate the design of the estimator.

Lemma 3. *Let* $q_0 > 0$, $P = \{i > 0 : q_i > 0\}$, $d = \gcd P$, *and let* A *be the matrix of* (8). *We have* $\mathrm{Null}(A) = \mathrm{Span}(\boldsymbol{v})$, *where* $\boldsymbol{v} = (v_0, \ldots, v_{m-1})^T$ *is a vector satisfying* $v_0 = 1$ *and* $v_j = \frac{1}{4}$ *for* $j \not\equiv 0 \pmod{d}$.

For example, for $d = 3$, $\boldsymbol{v} = (1, \frac{1}{4}, \frac{1}{4}, v_3, \frac{1}{4}, \frac{1}{4}, v_6, \frac{1}{4}, \ldots)^T$.

Proof: Consider a matrix A' obtained from A by replacing the first all-zero row with the row vector $(1, 0, \ldots, 0)$. By a simple application of the Gershgorin circle theorem,

$$|A'_{rr}| - \sum_{j \neq r} |A'_{rj}| = 2q_0 > 0,$$

for all $r > 0$, and therefore all the eigenvalues of A' are nonzero, i.e., $\mathrm{rank}(A') = m$. Thus, we have $\mathrm{rank}(A) = m - 1$, and therefore $\dim \mathrm{Null}(A) = 1$.

We now show $A\boldsymbol{v} = 0$, which along with $\dim \mathrm{Null}(A) = 1$, implies that $\mathrm{Null}(A) = \mathrm{Span}(\boldsymbol{v})$. Let $(A\boldsymbol{v})_r$ denote the rth element of $A\boldsymbol{v}$ for $r = 0, 1, \ldots, m-1$. Since the 0th row of A is all zero, we have $(A\boldsymbol{v})_0 = 0$. Based on (8), for $(A\boldsymbol{v})_r = 0$ to hold when $r > 0$, we require

$$q_0 \left(\frac{2}{3}(1 - 4v_r) + r v_r \right) + r \sum_{k : 1 \leq kd \leq m} v_{|kd - r|} q_{kd} - r v_r = 0.$$

This holds for $r \not\equiv 0 \pmod{d}$ if we let $v_j = \frac{1}{4}$ for all $j \not\equiv 0 \pmod{d}$. Finally, for $r \equiv 0 \pmod{d}$, $r > 0$, we can choose $v_d, v_{2d}, \ldots$ such that the above equality holds as these are not restricted in the statement of the lemma. ■

From the lemma, it follows that there is only one valid solution to the equation $A\boldsymbol{\rho}_\infty = 0$ which satisfies $\rho^0_\infty = 1$. This unique point is the limit of the autocorrelation function.

We have thus shown that if we know $\mathbf{q}$, we can determine $\boldsymbol{\rho}_\infty$. We now turn to the estimation problem, which is the inverse of determining $\boldsymbol{\rho}_\infty$ using $\mathbf{q}$. In other words, we are given a sequence whose autocorrelation we can compute and our goal is to determine $\mathbf{q}$. Note that we can rewrite the equation $A\boldsymbol{\rho}_\infty = 0$, where A is the matrix given in (8), as

$$C \begin{pmatrix} q_0 \\ q_1 \\ \vdots \\ q_m \end{pmatrix} = \begin{pmatrix} \rho^1_\infty \\ 2\rho^2_\infty \\ \vdots \\ (m-1)\rho^{m-1}_\infty \end{pmatrix}, \tag{12}$$

where $C = (C_{ri})$ is a $(m-1) \times (m+1)$ matrix whose elements are

$$C_{ri} = \begin{cases} \frac{2}{3} + \left(r - \frac{8}{3}\right)\rho^r_\infty, & i = 0 \\ r\rho^{i-r}_\infty, & \text{otherwise,} \end{cases}$$

where $r \in \{1, \ldots, m-1\}$ and $i \in \{0, 1, \ldots, m\}$.

Given $\boldsymbol{\rho}_\infty$, we can solve (12) for $\mathbf{q}$. Since we only know the sequence after a finite time n, we approximate $\boldsymbol{\rho}_\infty$ by $\boldsymbol{\rho}_n = \left(\rho^0_n, \ldots, \rho^{m-1}_n\right)$ computed from $s^{(n)}$. In our model, there exists K such that $q_i = 0$ for $i > K$. However, the value of K is unknown to us. We thus choose some m' and assume that $q_i = 0$ for $i > m'$. The value of m' can be chosen for example based on our knowledge of the underlying biological processes, such as slipped-strand mispairings [7], that lead to tandem repeats. Furthermore, the value of m' should be chosen large enough so that $m' \geq K$ with a high degree of confidence. Note that there are $m' + 1$ unknown quantities, namely, the elements $q_0, \ldots, q_{m'}$ of $\mathbf{q}$. Another parameter is the number of equations used to estimate $\mathbf{q}$, denoted m'', which should be chosen close to m'. Having chosen m', m'', we can write (12) as

$$C' \begin{pmatrix} q_0 \\ q_1 \\ \vdots \\ q_{m'} \end{pmatrix} = \begin{pmatrix} \rho^1_n \\ 2\rho^2_n \\ \vdots \\ m''\rho^{m''}_n \end{pmatrix}, \tag{13}$$

where C' is the matrix containing the first m'' rows and the first $m' + 1$ columns of C. Now to obtain an estimate of $\mathbf{q} = (q_0, q_1, \ldots, q_{m'})$ we can solve the least-square curve fitting problem

$$\begin{aligned} \hat{\mathbf{q}} = \arg\min_{\mathbf{q}}\ & \|C'\mathbf{q} - \boldsymbol{\rho}_n\|_2^2 \\ \text{s.t.}\ \ & \mathbf{1}\mathbf{q}^T = 1 \\ & q_i \geq 0,\ \text{for } 0 \leq i \leq m'. \end{aligned} \tag{14}$$

The solution $\hat{\mathbf{q}}$ of this problem contains an estimate of the substitution probability q_0 and the probabilities q_ℓ of duplications of lengths ℓ. Noting that the expected length added to the sequence by each mutation is $\sum_{i=1}^{m'} i\hat{q}_i$, we estimate the total number n of mutations that have occurred as

$$\hat{n} = \frac{\left|s^{(n)}\right| - \left|s^{(0)}\right|}{\sum_{i=1}^{m'} i\hat{q}_i}, \tag{15}$$

where we assume the length of the seed $s^{(0)}$ is equal to the pattern length. The estimator based on the proposed Stochastic Model of Tandem Repeats and defined by (14) and (15) is referred to as SMTR.

In tandem repeat sequences observed in genomes, such as the one given in Figure 2, it is clear that duplication events have lengths that are multiples of a certain value, leading to a pattern of that length appearing many times. We refer to this length as the *pattern length* and to the number of times that the pattern appears as the *copy number*. While in general SMTR does not need to know the pattern length to provide estimates of the mutation probabilities, if it is known, we can set $q_i = 0$ for i not divisible by the pattern length.

Since our method relies on asymptotic approximation, for short sequences, specifically those with copy number ≤ 3, we provide an alternative estimation algorithm. In such sequences there is ≤ 2 duplication events of length equal to the pattern length and 0 or more substitutions. The number of duplications can be found easily from the length of the sequence. Let a_i be the number of distinct symbols appearing at the ith position (relative to the start of the pattern) of different copies minus 1. For example, for $\underline{\text{A}}\text{CT}\underline{\text{G}}\text{CT}\underline{\text{A}}\text{CT}$, we have $a_1 = 1$, since two symbols, A and G, appear in the first position of different copies, and $a_2 = a_3 = 0$. The a_i can be used to infer the number of substitutions. A substitution will contribute to a_i only if it occurs after the first duplication event. To account for hidden substitutions, we estimate the number of substitutions as $(\sum_i a_i)\frac{(r+1)}{r}$, where r is the number of duplication events. So we have estimates both for the number of substitutions and the number of duplications. Note that in this simple analysis, we have assumed that each substitution results in a new symbol, which is a reasonable assumption for a small number of mutations.

III. Simulation and Data Analysis Results

In this section, we use simulation to evaluate the performance of SMTR by comparing its estimates of the model parameters with the true values. We also compare SMTR to DTSCORE introduced by Elemento and Gascuel [16], which was shown to outperform similar methods [14]. Further, we apply SMTR to tandem repeats in the human genome to study variation across chromosomes and pattern lengths.

In the results that follow, we set the computation parameters as follows. First, we find $\boldsymbol{\rho} = (\rho^r)$ for $r = 0, 1, \ldots, \lfloor\frac{|s|}{2}\rfloor$. This ensures that each value of the autocorrelation function is the average of at least $|s|/2$ values. Furthermore, we let $m' = m'' = \min\Big(\max(10d, 5r^*), \Big\lfloor\frac{|s|}{2}\Big\rfloor\Big)$, where $r^* = \arg\max_r \rho^r$. The $\max$ here is intended to ensure that m' is large enough, while the $\min$ ensures that all needed values of ρ are available. We note that in (15), if $\hat{q}_0$ is close to 1, then the estimate $\hat{n}$ for n may be very large. It is reasonable to expect that $\hat{n}$ is not larger than the length of the sequence. Thus, we heuristically add the constraint $(d, 2d, \ldots, m'd)(q_d, q_{2d}, \ldots, q_{m'd})^T \geq 1$ to (14), where d is the pattern length. This ensures that on average each mutation contributes at least 1 to the length of the sequence. Finally, while the estimation method is geared towards tandem repeats with substitution mutations, our inspection of the results shows that for perfect tandem repeats, the algorithm returns probability near zero for substitution mutations, as expected, and nearly uniform probability for different duplication lengths. Thus, in the results that follow, we apply it to tandem repeats regardless of the apparent presence of substitution mutations.

A. Simulation Results

We now turn to evaluating the performance of SMTR through simulation and also compare it with DTSCORE [16]. We show that SMTR provides more accurate estimates and is significantly faster compared to DTSCORE.

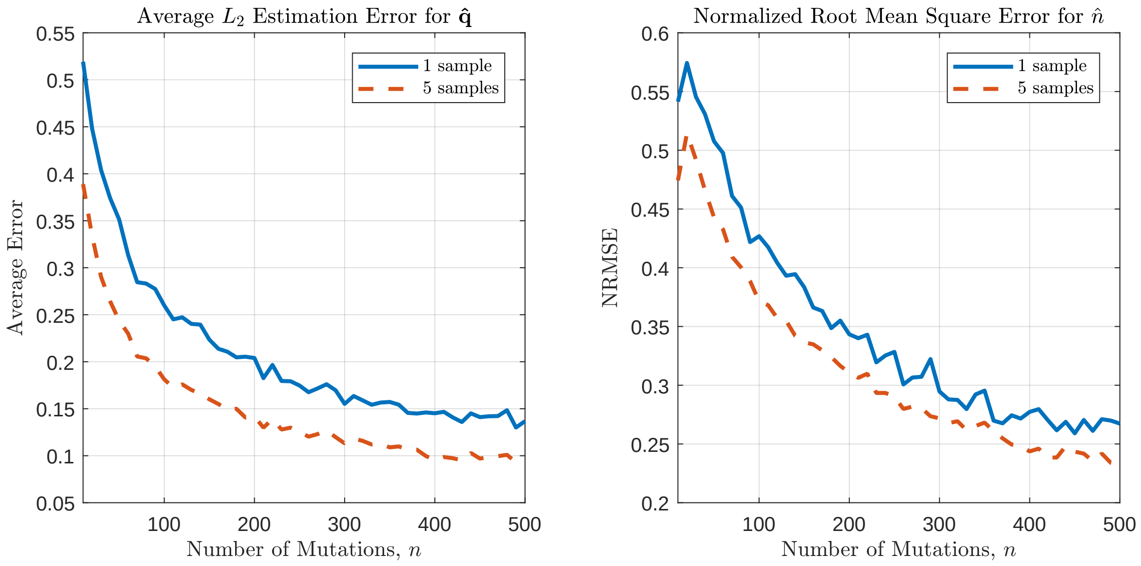
In our simulation set up, we first generate a random seed $s^{(0)}$ of a random length d that then undergoes n random substitutions and tandem duplications, where the probabilities of these events are given by $\mathbf{q}$, itself randomly generated. The resulting sequence $s^{(n)}$ and the pattern length d are then passed to the SMTR estimator, which of course does not know $s^{(0)}$, n, or $\mathbf{q}$. We evaluate the performance by finding the L_2 error in estimating $\hat{\mathbf{q}}$, $\|\hat{\mathbf{q}} - \mathbf{q}\|_2$, averaged across N experiments for each value of n. We also find the normalized root mean square (NRMS) error in estimating n. For a given value of n, NRMS Error is defined as

$$\mathrm{NRMSE}(n, \hat{n}) = \frac{1}{n}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{n}_i - n)^2}\ ,$$

where N is the number of experiments with n mutations and $\hat{n}_i$ is the estimate for n in the ith experiment.

We find the errors for two different cases: for a pair of given values for n and $\mathbf{q}$, we estimate $\hat{n}$ and $\hat{\mathbf{q}}$ based on 1) a single sequence and 2) n_s sequences all generated with parameters $\mathbf{q}$ and n. In the latter case, estimates are obtained for each sequence individually and then averaged. The multiple-sample case is intended to show that performance improves, as expected, with more data. Due to the large number of tandem repeat sequences in many genomes, it is reasonable to expect that for a set of factors affecting duplication probabilities, e.g., GC content and pattern length, a given set of values for these factors is likely to arise multiple times. When studying the effects of such factors on mutation rates, we may expect a similar performance improvement by averaging the estimates among all instances with the same set of values for the factors.

More detail on the simulation setup is given below. The results are given in Figure 3 where n ranges from 10 to



(a)

(b)

Figure 3: Errors of the estimate $\hat{\mathbf{q}}$ of $\mathbf{q}$, (a), and the estimate $\hat{n}$ of n, (b) .

500, with step size equal to 10. For each value of n, the experiment is performed $N = 500$ times, and in each of these N trials, estimates are obtained based on a single sequence and based on $n_s = 5$ sequences drawn for the same seed and $\mathbf{q}$. We observe that as n increases, the errors sharply decrease. For a single sequence and a small number of mutations, the estimation algorithm relies on a very limited amount of data. As the number n of mutations increases, the sequence becomes longer, providing more data in the form of the autocorrelation function and asymptotic approximations become more accurate. It is also observed that with more samples for the same set of parameters, more accurate estimates are obtained.

We now describe the process of sequence generation and estimation in more detail. The seed sequence $s^{(0)}$ is a random sequence over the alphabet $\{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}$ of a random length that is chosen uniformly from the set $\{4, 5, \dots, 10\}$. We set d for the duplication process (all duplication lengths will be multiples of d) equal to the length of the seed. To choose $\mathbf{q}$, we choose q_0, q_d, q_{2d}, and q_{3d} by randomly selecting a point from the simplex

$$\{(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \mid \alpha_1 + \cdots + \alpha_4 = 1,\ \alpha_1, \dots, \alpha_4 \geq 0\}.$$

All other values of $\mathbf{q}$ are set to 0. We then perform n mutation steps, each a substitution with probability q_0 or a tandem duplication of length id with probability q_{id}, for $i \in \{1, 2, 3\}$. If in a tandem duplication step, the length chosen for duplication is larger than the length of the sequence, the whole sequence is duplicated. Note that since the length of the sequence grows, such an event may only happen a few times at the beginning of the process.

We now compare the performance of SMTR with DTSCORE [16]. DTSCORE is a distance-based algorithm designed to find the duplication history in the form of a tree for a given sequence, thus providing estimates for the number of duplications of various lengths. In [14], it was shown that DTSCORE performs better than other algorithms for identifying the duplication tree, including TRHIST [12] and WINDOWS [13]. Due to the slower speed of DTSCORE (the worst-case time complexity is $O(L^4)$, where L is the copy number of the sequence), we restrict the range of the number of mutations n to $\{10, 20, \dots, 120\}$ and also reduce $N = 200$ but maintain $n_s = 5$. The comparison is given in Figure 4. Since from DTSCORE, we can only derive estimates for the counts of duplications but not substitutions, we compare the accuracy of estimating $\mathbf{q}' = (q'_1, q'_2, \dots)$ where q'_i for $i \geq 1$ is defined as

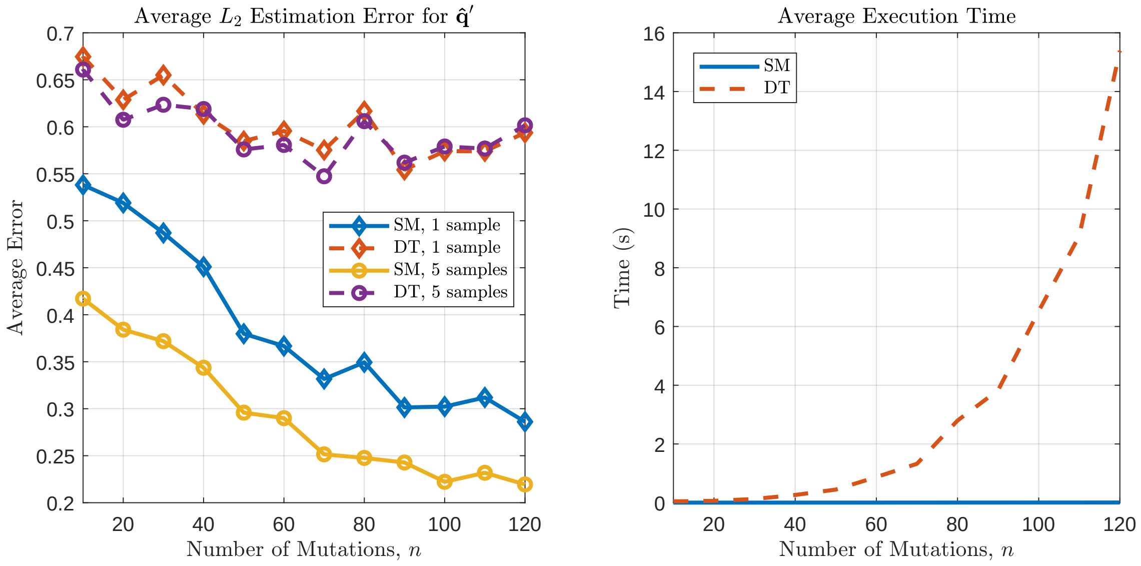
$$q'_i = \frac{q_i}{1 - q_0}\ .$$

From Figure 4a, it is clear that SMTR estimates $\mathbf{q}'$ with significantly higher accuracy than DTSCORE. Furthermore, if multiple samples from the same distribution are available, the improvement for SMTR is larger than for DTSCORE. Finally, the execution time of SMTR is faster than DTSCORE. In particular, for $n = 120$, on average, SMTR needs no more than 0.015 seconds to compute the estimate, while DTSCORE needs 14.11 seconds, almost 3 orders of magnitude longer.

B. Tandem Repeats in the Human Genome

We now apply SMTR to tandem repeats in the human genome to estimate the number of substitution and tandem duplication mutations for each. We use these estimates to explore the variation of mutation rates for minisatellite and microsatellites and across chromosomes.

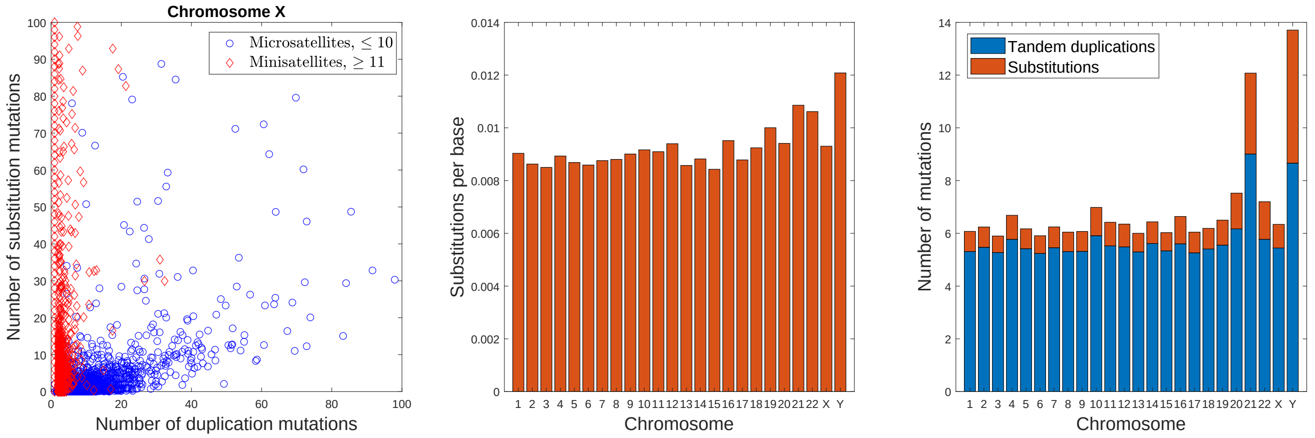
We use the Tandem Repeats Database (TRDB) [17], which provides the set of tandem repeats in each chromosome, as identified by the Tandem Repeat Finder (TRF) algorithm, and related information such as the length of the repeat unit and indel (insertion/deletion) percentage. As a preprocessing step, among overlapping repeats, we keep only one. We also



(a)

(b)

Figure 4: Comparison of SMTR Estimation (SM) and DTSCORE (DT): Error of $\hat{\mathbf{q}}'$, (a), and the average execution time for an instance of the problem on an Intel Core i7-7700 CPU with 16 GB of RAM, (b).



(a)

(b)

(c)

Figure 5: The mutation profile of tandem repeats in chromosome X (a) and mutation variation across chromosomes in microsatellites (pattern length ≤ 10): Mean of the ratio of the number of substitutions to the length of the tandem repeat sequence (b) and mean of the total number of mutations per tandem repeat sequence for each chromosome (c).

remove repeats with unknown (N) bases and those with copy number less than 2. Finally, we discard repeats whose indel percentage is nonzero, as our model does not include insertion and deletion mutations. As an example, the number of repeat sequences in chromosome 1 reduces from $93,626$ to $38,628$ as a result of preprocessing.

We applied the SMTR algorithm to tandem repeats in each chromosome. The results for chromosome X are given in Figure 5a. Each point in this plot corresponds to a tandem repeat sequence. The position of each point is determined by the estimated number of tandem duplications and substitutions that occurred to create the sequence. It can be observed that tandem repeat sequences can roughly be divided into two clusters with different behaviors: one dominated by tandem duplication mutations and the other by substitution mutations. This difference in behavior matches well with the classification of tandem repeats as microsatellites and minisatellites, with pattern lengths of 1-10 and 11-100 bases, respectively. Other chromosomes exhibit behavior similar to chromosome X illustrated here. Among all chromosomes, the minimum Kendall tau correlation coefficient between the rankings of repeats based on length of the pattern and based on the fraction of mutations that are substitutions was 0.5160. Given the large number of tandem repeats in each chromosome, such high correlation coefficients lead to p-values that are practically zero (as computed with MATLAB).

The different mutation profiles suggest that these two types of tandem repeats may result from different mutation mechanisms. This is compatible with previous findings, where polymerase slippage is thought to give rise to microsatellites while unequal recombination is believed to cause the heterogeneity observed in minisatellites [18]. Our method is only designed to model slippage and not recombination. The fact that it generally estimates the number of substitutions to be higher for minisatellites than microsatellites can be the result of higher raw heterogeneity that is observed in microsatellites and/or caused by model mismatch. The results of this analysis suggests that it is possible to design statistical test to decide the origin of tandem repeat sequences, as a means of classifying them, rather than relying on classification merely based on pattern length. We leave developing such tests to future work.

We now turn our attention to evaluating the variation of mutation rates across chromosomes. Through comparison with the chimpanzee genome [19], [20], [21], it is known that mutation rates vary across chromosomes. To see whether this variation can also be observed in repeated regions, we study the number of mutations in tandem repeat sequences across chromosomes. Since our model represents replication slippage, we only consider tandem repeats with short patterns. Specifically, for tandem repeats with pattern length ≤ 10, we estimate the number of substitution and duplication mutations. As a measure of mutation activity, we find the average of the ratio of the the number of substitutions to the length of the tandem repeat sequence for each chromosome (Figure 5b). The top five chromosomes that have the highest substitution rates are Y, 21, 22, 19, and 16. Based on comparison with the chimpanzee genome [21], the five chromosomes with highest mutation activity are Y, 21, 19, 22, and 16. The fact that the top five match (p-value=0.00002) suggests a strong relationship between substitutions in repeated regions and overall mutation activity in chromosomes. On the other hand, the results are not exactly aligned. For example, while chromosome X has the smallest divergence from chimpanzee, here it does not have the smallest normalized number of substitutions. We also repeated this analysis for repeats with maximum pattern lengths of 8, 9, 11, and 12, and in all cases, at least four of the top five matched the result from comparison with chimpanzee [21]. Overall, our results suggest estimation of mutation activity based on tandem repeats can be a powerful tool in studying mutations since unlike existing methods it relies on a single genome rather than on comparison of genomes from different species.

We also considered the average number of mutations per tandem repeat for each chromosome (Figure 5c). On average, tandem repeats in chromosome 21 have a higher number of mutations than other autosomes. The reason for this behavior is unknown to us but it is interesting to note that individuals with trisomy 21 can survive into adulthood, which suggests that mutations in chromosome 21 are relatively better tolerated. The average number of duplication mutations is in fact estimated to be higher in chromosome 21 than the Y chromosome. The higher number of mutations in chromosome 21 compared to other autosomes is also observed if we set the upper bound on the length of the patterns that are considered at 8, 9, 11, and 12.

In Figure 5c, 3 of the 5 chromosomes with the highest total number of mutations in microsatellites, Y, 21, and 22, match the result from [21]. While this also suggests a higher mutation activity in these chromosomes, care should be taken in interpreting results about mutation counts that are not normalized by the length of the sequence. The opportunity for mutation increases with length and copy number. In particular, increased copy number may increase the probability of misalignment during replication [22]. Another factor that can affect the number of mutations in a complex manner is the interplay between substitution mutations and tandem duplication mutations: if many substitutions occur, the copies become more heterogeneous, which may decrease the possibility of misalignment. This interaction is not taken into account in our model and left to future work.

IV. Conclusion

In this paper, we introduced a new stochastic model for tandem duplication and substitution mutations, and analyzed it via stochastic approximation. In particular, we fully characterized the limit set of the stochastic process described by the model. In addition to enabling us to predict the behavior of a sequence that undergoes tandem duplication and substitution mutations, this characterization allowed us to derive a minimization problem whose solutions are estimates of the mutation probabilities for tandem duplication and substitution. We showed further that it is possible to estimate the total number of mutations. Finally, we evaluated the estimation method via simulation by generating random sequences and comparing the estimated probabilities with the true values and also applied it to the human genome, where it demonstrated the differing behavior of micro- and mini-satellites as well as the variability of mutation activity across chromosomes.

Advantages of our method include its scalability and the fact that it relies on a single sequence to infer occurrences of mutations. While with this method, we can learn only about mutations in tandem repeat regions, our results show that the findings may be applicable to surrounding regions and can be of use in forming hypotheses about mutation activity, for example, about factors that increase or decrease activity.

There still exist many open problems in stochastic modeling and estimation for tandem repeats. For example, the model presented here does not take into account deletions nor the fact that the level of heterogeneity may affect the probability of tandem duplication. Further, we only analyzed it in the asymptotic regime and left finite-time behavior to future work. Finally, further work is needed to accurately model mutations other than DNA slippage that cause duplication, especially those that lead to minisatellite repeats.

Acknowledgment: The authors would like to thank Han Mao Kiah for helpful discussions related to Lemma 2.

References

[1] 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.

[2] D. Pumpernik, B. Oblak, and B. Bortnik, "Replication slippage versus point mutation rates in short tandem repeats of the human genome," *Molecular Genetics and Genomics*, vol. 279, no. 1, pp. 53–61, 2008.

[3] T. B. Sonay, T. Carvalho, M. D. Robinson, M. P. Greminger, M. Krützen, D. Comas, G. Highnam, D. Mittelman, A. Sharp, T. Marques-Bonet, and A. Wagner, "Tandem repeat variation in human and great ape populations and its impact on gene expression divergence," *Genome Research*, vol. 25, pp. 1591–1599, Jan. 2015.

[4] J. M. Butler, "Genetics and Genomics of Core Short Tandem Repeat Loci Used in Human Identity Testing," *Journal of Forensic Sciences*, vol. 51, no. 2, pp. 253–265, 2006.

[5] K. Usdin, "The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases," *Genome Research*, vol. 18, pp. 1011–1019, July 2008.

[6] J. W. Fondon and H. R. Garner, "Molecular origins of rapid and continuous morphological evolution," *Proceedings of the National Academy of Sciences*, vol. 101, no. 52, pp. 18058–18063, 2004.

[7] G. Levinson and G. A. Gutman, "Slipped-strand mispairing: a major mechanism for DNA sequence evolution," *Molecular Biology and Evolution*, vol. 4, no. 3, pp. 203–221, 1987.

[8] C. Schlötterer, "Evolutionary dynamics of microsatellite DNA," *Chromosoma*, vol. 109, pp. 365–371, Sept. 2000.

[9] Y. Lai and F. Sun, "The Relationship Between Microsatellite Slippage Mutation Rate and the Number of Repeat Units," *Molecular Biology and Evolution*, vol. 20, no. 12, pp. 2123–2131, 2003.

[10] S. Kruglyak, R. T. Durrett, M. D. Schug, and C. F. Aquadro, "Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations," *Proceedings of the National Academy of Sciences*, vol. 95, no. 18, pp. 10774–10778, 1998.

[11] R. Durrett and S. Kruglyak, "A new stochastic model of microsatellite evolution," *Journal of Applied Probability*, vol. 36, pp. 621–631, Sept. 1999.

[12] G. Benson and L. Dong, "Reconstructing the duplication history of a tandem repeat.," in *ISMB*, pp. 44–53, 1999.

[13] M. Tang, M. Waterman, and S. Yooseph, "Zinc finger gene clusters and tandem gene duplication," *Journal of Computational Biology*, vol. 9, no. 2, pp. 429–446, 2002.

[14] Olivier Gascuel, Denis Bertrand, and Olivier Elemento, "Reconstructing the duplication history of tandemly repeated sequences," in *Mathematics of Evolution and Phylogeny* (Olivier Gascuel, ed.), ch. 8, Oxford, New York: Oxford University Press, May 2005.

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

[16] O. Elemento and O. Gascuel, "An efficient and accurate distance based algorithm to reconstruct tandem duplication trees," *Bioinformatics*, vol. 18, pp. S92–S99, Oct. 2002.

[17] Y. Gelfand, A. Rodriguez, and G. Benson, "TRDB–the Tandem Repeats Database," *Nucleic Acids Research*, vol. 35, pp. D80–87, Jan. 2007.

[18] H. Debrauwere, C. G. Gendrel, S. Lechat, and M. Dutreix, "Differences and similarities between various tandem repeat sequences: Minisatellites and microsatellites," *Biochimie*, vol. 79, pp. 577–586, Oct. 1997.

[19] A. Hodgkinson and A. Eyre-Walker, "Variation in the mutation rate across mammalian genomes," *Nature Reviews Genetics*, vol. 12, pp. 756–766, Nov. 2011.

[20] I. Ebersberger, D. Metzler, C. Schwarz, and S. Pääbo, "Genomewide Comparison of DNA Sequences between Humans and Chimpanzees," *American Journal of Human Genetics*, vol. 70, pp. 1490–1497, June 2002.

[21] The Chimpanzee Sequencing and Analysis Consortium, R. H. Waterson, E. S. Lander, and R. K. Wilson, "Initial sequence of the chimpanzee genome and comparison with the human genome," *Nature*, vol. 437, pp. 69–87, Sept. 2005.

[22] M. Wierdl, M. Dominska, and T. D. Petes, "Microsatellite instability in yeast: Dependence on the length of the microsatellite," *Genetics*, vol. 146, pp. 769–779, July 1997.