AN IMPOSSIBILITY RESULT FOR PHYLOGENY RECONSTRUCTION FROM *k*-MER COUNTS

BY WAI-TONG LOUIS FAN[1,a], BRANDON LEGRIED[2,b] AND SEBASTIEN ROCH[3,c]

[1]*Department of Mathematics, Indiana University,* [a]*waifan@iu.edu*
[2]*Department of Statistics, University of Michigan,* [b]*blegried@umich.edu*
[3]*Department of Mathematics, University of Wisconsin–Madison,* [c]*roch@wisc.edu*

We consider phylogeny estimation under a two-state model of sequence evolution by site substitution on a tree. In the asymptotic regime where the sequence lengths tend to infinity, we show that for any fixed *k* no statistically consistent phylogeny estimation is possible from *k*-mer counts over the full leaf sequences alone. Formally, we establish that the joint distribution of *k*-mer counts over the entire leaf sequences on two distinct trees have total variation distance bounded away from 1 as the sequence length tends to infinity. Our impossibility result implies that statistical consistency requires more sophisticated use of *k*-mer count information, such as block techniques developed in previous theoretical work.

1. Introduction. Molecular sequence comparisons are fundamental to many bioinformatics methods [6, 10, 20]. In particular, the probabilistic analysis of sequences and their statistics has provided valuable insights, for instance, in comparative genomics [3, 23, 27, 34], population genetics [2, 31, 32, 39], and phylogenetics [13, 14, 28, 35, 37]. In this paper, we consider alignment-free phylogeny reconstruction [21, 42].

Alignment-free approaches are an important class of methods for estimating evolutionary trees that bypass the computationally hard multiple sequence alignment problem (depicted in Figure 1) and avoid the need for a reference genome. Typically, these methods construct pairwise distances between sequences based on match lengths [22, 41] or *k*-mer counts [15, 21, 33]. Here a *k*-mer refers to a consecutive substring of length *k* in an input sequence (see Figure 2 for an illustration). The pairwise distance matrix obtained is then used to reconstruct the phylogenetic relationships among the sequences. A variety of standard distance-based phylogenetic methods can be used for this purpose [38, 44]. Numerous popular pipelines are available that implement these alignment-free approaches [22, 24, 25, 30], although they do not offer rigorous guarantees of accurate reconstruction.

In this paper, we consider the problem of phylogeny estimation under a two-state symmetric model of sequence evolution by site substitutions on a leaf-labeled tree. In the asymptotic regime where the sequence length tends to infinity, we show that:

> *for any fixed k, no statistically consistent phylogeny estimation is possible from the k-mer counts of the entire input sequences alone.*

Formally, we establish that the joint distribution of *k*-mer counts over the entire leaf sequences on two distinct trees have total variation distance bounded away from 1 as the sequence length tends to infinity. Put differently, these two joint distributions have a nonvanishing overlap in that asymptotic regime. Our results are information-theoretic: since the reconstruction probability of any method is only as good as the worst total variation distance

(a) A A T G C T C A T T G A C G G G
(b) A A A G C T C A T T G A C G G G
(c) G T C A C T T T A C G T T

(a) A A T G C T C A T T G A C G G G
(b) A A A G C T C A T T G A C G G G
(c) G T C A C T - - T T - A C G T T

```
                    (a)
                    (b)
                    (c)
```

FIG. 1. *Standard steps in phylogeny estimation. Top: DNA sequences obtained from species (a), (b), and (c). While inherited from a common ancestor, the sequences and their lengths differ because of past mutations (including insertions and deletions). Middle: A multiple sequence alignment of the sequences, where gaps are inserted to align the columns as best as possible. Each column indicates inferred common ancestry (homology). Bottom: A rooted phylogenetic tree depicting the estimated evolutionary history of the sequences, with (a) and (b) being more closely related.*

(see [16], Lemma 3.2), our main claim (Theorem 1) implies an impossibility result for reconstruction methods using only k-mer counts across the entire sequences at the leaves. On the other hand, our results have no implications for reconstruction methods using k-mer counts in more elaborate ways, for example, through block decomposition. We come back to prior approaches of this type below in "related work."

To bound the total variation distance between the two distributions on well-chosen trees, our proof takes advantage of a multivariate local central limit theorem, an approach which is complicated by the probabilistic and linear dependencies of k-mers.

Related work. A related impossibility result was established in our previous work [17], where it was shown that no consistent distance estimation is possible from *sequence lengths alone* under the TKF91 model [40], a more complex model of sequence evolution which also allows for insertions and deletions (indels). On the other hand, sequence lengths are significantly simpler to analyze than k-mers and are not used in practice to infer phylogenies. Moreover the results in [17] only apply to distance-based phylogeny estimation methods, while our current results are more general.

In a separate line of work, a computationally efficient algorithm for alignment-free phylogenetic reconstruction was developed and analyzed in [8]. Rigorous sequence length guarantees for high-probability reconstruction under an indel model related to the TKF91 model were established. While this method is based on 1-mers, it first divides up the input sequences into blocks of an appropriately chosen length and then compares the 1-mer counts on each block across sequences. A block in the ancestral sequence gives rise to blocks in the descendant sequences that have the same position, so the comparison does not require an alignment technique. The weak correlation between the blocks allows the use of concentration inequalities on a notion of pairwise distance proposed in [8]. In particular, this reconstruction method
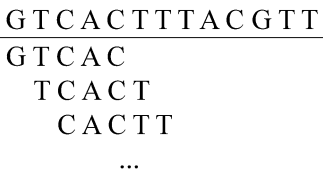
G T C A C T T T A C G T T
G T C A C
  T C A C T
    C A C T T
...

FIG. 2. *For a given sequence, the k-mer counts are obtained by reading words of length k starting from each site and then counting how many times each possible length-k word appears. Here $k = 5$.*

uses more information than 1-mer counts over the entire sequences (i.e., it uses 1-mer counts *over each separate block*), so that our results do not apply to it (in the limit of zero indel rate). However no practical implementation of this 1-mer-based approach is available. In [8], accurate phylogenetic reconstruction with high probability is shown to be achievable when the sequence length is of polynomial order in the number of leaves n. In a constrained regime of parameters, significantly improved bounds on the sequence length requirement were obtained in [19], who used techniques related to those of [8] and also exploited a well-studied connection to ancestral sequence reconstruction. In other related work, it was proved in [1] that the tree topology as well as mutation parameters can be identified from pairwise joint k-mer count distributions under more general substitution-only models of sequence evolution using an appropriately defined notion of distance. Block techniques can then be used to derive statistical consistency results. See also [11] for extensions to coalescent-based models. We emphasize that the results in [1, 8, 11, 19] do not contradict our main claim (Theorem 1), which excludes block decomposition.

Alignment-free sequence comparisons based on k-mer counts were also studied for independent sequences with i.i.d. sites or under certain hidden Markov models of sequences [3, 27, 34, 43]. Because they assume independent sequences, such results are not directly relevant to phylogeny reconstruction.

Organization. The paper is organized as follows. In Section 2, we state our main results after providing the necessary background and definitions. We also sketch the main steps of the proof. The details of the proof can be found in Section 3. A few auxiliary results are in the Appendix.

2. Definitions and main result. In this section, we state our result formally, after introducing the relevant concepts.

k-mers. Let k be a positive integer, fixed throughout. First, we define k-mers and introduce their frequencies in a binary sequence, which will serve as our main statistic.

DEFINITION 1. A *k-mer* is a binary string of length k, that is, $y \in \{0, 1\}^k$. For a binary sequence $\sigma = (\sigma_i)_{i=1}^m$ of length m, we let $f_\sigma(y) \in \mathbb{Z}_+$ be the number of times y appears in σ as a consecutive substring, where $\mathbb{Z}_+$ is the set of nonnegative integers. That is,

$$f_\sigma(y) = \sum_{i=0}^{m-k} \mathbf{1}\{(\sigma_{i+1}, \ldots, \sigma_{i+k}) = y\}.$$

The *frequency vector (or count vector) of k-mers in σ* is the vector

$$f_\sigma = (f_\sigma(y))_{y \in \{0,1\}^k} \in \mathbb{Z}_+^{2^k}.$$

The coordinates of f_σ are ordered such that the jth coordinate is the frequency of the k-mer that is the base-2 numeral representation of $j - 1$.

For example, when $k = 1$, the count vector of 1-mers of a binary sequence is (a, b) where a is the number of zeros and b is the number of ones. Hence, the count vector of 1-mers of 00111000 is $(5, 3)$. When $k = 2$, there are $2^k = 4$ binary strings, namely $\{(00), (01), (10), (11)\}$. So the count vector of 2-mers of the sequence 001111000 is $(3, 1, 1, 3)$ since (00) appears 3 times, (01) appears one time, etc. By convention, the count vector of k-mers for any binary sequence with length less than k is equal to $(0, \ldots, 0) \in \mathbb{Z}_+^{2^k}$.

Probabilistic model of sequence evolution. We consider a symmetric substitution model on phylogenies, also known as the Cavendar–Ferris–Neyman (CFN) model [4, 18, 29], for binary sequences of fixed length m. The CFN model on a single edge of a metric tree is a continuous-time Markov process with state space $\{0, 1\}^m$ such that (i) the m digits are independent and (ii) each of the m digits follows a continuous-time Markov process with two states $\{0, 1\}$ that switches state at rate 1.

We are interested in this process on a rooted metric tree T, that is, indexed by all points along the edges of T. We view an edge of length ℓ as the interval $[0, \ell]$ for the continuous-time substitution process. The root vertex ρ is assigned a state $\mathcal{X}_\rho \in \{0, 1\}^m$, drawn from the uniform distribution on $\{0, 1\}^m$. This state then evolves down the tree (away from the root) according to the following recursive process. Moving away from the root, along each edge $e = (u, v)$ starting at u, we run the CFN process for a time $\ell_{(u,v)}$ with initial state $\mathcal{X}_u$, described in the previous paragraph. Such processes along different edges starting at u are conditionally independent, given $\mathcal{X}_u$. Denote by $\mathcal{X}_t$ the resulting state at $t \in e$. Then the full process, denoted by $\{\mathcal{X}_t\}_{t \in T}$, is called the *CFN model on tree T*. In particular, the set of leaf states is $\mathcal{X}_{\partial T} = \{\mathcal{X}_v : v \in \partial T\}$. It is clear that, under this process, the m digits remain independent. For more background on the CFN model; see, for example, [38].

An impossibility result. Our main result is the following. Recall that the total variation distance between two probability measures ν_1 and ν_2 on a countable space E is defined by

$$
\|\nu_1 - \nu_2\|_{\text{TV}} = \sup_{E' \subseteq E} |\nu_1(E') - \nu_2(E')|. \tag{1}
$$

THEOREM 1. *Fix $k \in \mathbb{N}$. For any $n \geq 3$, there exists distinct trees $T_1 \neq T_2$ with n leaves such that*

$$
\limsup_{m \to \infty} \|\mathcal{L}_m^{(1)} - \mathcal{L}_m^{(2)}\|_{\text{TV}} < 1, \tag{2}
$$

where $\mathcal{L}_m^{(i)}$ is the law of the k-mer frequencies of the leaf sequences of length m under the CFN model on tree T_i. Furthermore, the trees $\{T_1, T_2\}$ can be chosen to be independent of k.

From (1), we see that (2) implies the following: using only the k-mer frequencies over the entire leaf sequences for a fixed $k \geq 1$, there is no statistical test that can distinguish between T_1 and T_2 with success probability going to 1 as the sequence length tends to $+\infty$. More precisely, by (2) and the reconstruction upper bound in part 1 of [16], Lemma 3.2, there exists $\epsilon > 0$ such that the probability that a tree estimator gives the correct estimate is at most $1 - \epsilon$, uniformly for all estimators and all integers $m \geq k$. We point out again that our results do not apply to block decomposition methods.

PROOF SKETCH. Since our goal is to prove a negative result, we get to pick the trees. We consider two trees $\{T_1, T_2\}$ that have the same set of n leaves and are the same except for the placement of a single edge leading to leaf A. These trees are depicted in Figure 3 and described in detail in Section 3.1 below. The topologies of $\{T_1, T_2\}$ differ only on the subtree containing three leaves $\{A, B, C\}$ that have the same distance from the root.

We seek to distinguish the law of the k-mer frequencies of the n leaf sequences between the two trees. This will be done in two steps, in Sections 3.2 and 3.3 respectively, and concluded in Section 3.4.

1. *Step 1 (Reductions):* By using the Markov property of the CFN process on trees, we first reduce the problem from n leaf sequences to only 3 sequences (Lemma 1). We can assume the sequence length m is a multiple of k (Lemma 2). Then k-mer frequencies are

functions of pairs of adjacent, nonoverlapping k-mers, together with the first and the last k-mers (Lemma 3). For short, we refer to these pairs as "adjacent k-mer pairs" and a precise definition is in (4) below. We can further reduce the problem to distinguishing the laws of adjacent k-mer pairs (Lemma 4). The collection of adjacent k-mer pairs satisfy certain linear relations (Lemma 5), which lead to redundancy that we need to address (Lemma 6). Summarizing, the problem is reduced to distinguishing the laws of nonredundant, adjacent k-mer pairs on three points $\{A, B', C'\}$ as depicted in Figure 4 (Lemma 7).

2. *Step 2 (Applying a local CLT)*: We apply a local central limit theorem for i.i.d. vectors to the law of nonredundant, adjacent k-mer pairs as $m \to \infty$ (Lemmas 8 and 9 and Theorem 2). Nonredundancy guarantees the nondegeneracy of the limit distribution (Lemmas 10, 11, 12, and 13). The two limit normal distributions, under the two trees respectively, have an overlap (Lemmas 14 and 15) and therefore so do the laws of nonredundant, adjacent k-mer pairs.

Section 3.4 concludes the proof of Theorem 1. $\square$

3. Proof. In this section, we give the details of the proof. Some standard results are stated in the Appendix.

3.1. *The two trees.* We consider two rooted metric trees $\{T_1, T_2\}$ as follows.

1. T_1 and T_2 have the same set of $n \geq 3$ leaves $\{A, B, C, X^4, \ldots, X^n\}$.
2. The subtree of T_1 restricted to the $n - 1$ leaves $\{B, C, X^4, \ldots, X^n\}$ is the same as that of T_2. Here the restriction of a metric tree T to a subset L of leaves is the metric tree obtained from T by keeping only those points lying on a path between two leaves in L.
3. The subtrees of T_1 and T_2 below the most recent common ancestor (MRCA) of $\{A, B, C\}$ contain none of $\{X^4, \ldots, X^n\}$.
4. Leaves $\{A, B, C\}$ satisfy, for $i \in \{1, 2\}$,

(3)
$$\text{dist}_{T_i}(\rho, C) = \text{dist}_{T_i}(\rho, B) \quad \text{and}$$
$$\text{dist}_{T_1}(A, C) = \text{dist}_{T_2}(A, B) < \text{dist}_{T_i}(B, C),$$

where $\text{dist}_{T_i}(x, y)$ denotes the sum of edge lengths along the path from x to y in the tree T_i.

These trees are depicted in Figure 3, where $X = (X^4, \ldots, X^n)$ refers to the set of all leaves other than $\{A, B, C\}$. In *Newick tree format* (see, e.g., [44]), the topology of T_1 restricted to $\{A, B, C\}$ is $((A, C), B)$, while the topology of T_2 restricted to $\{A, B, C\}$ is $((A, B), C)$. Clearly, $\{T_1, T_2\}$ does not depend on k, and their topologies differ only on the subtree containing three leaves $\{A, B, C\}$. The topology of the trees restricted to X is arbitrary and plays no role in the argument.

Notation. For $i \in \{1, 2\}$, we let $\mathbb{P}^{(i)} = \mathbb{P}^{(i),m}$ be the probability measure of the CFN model on T_i with sequence length m (recall that the root state is drawn from the uniform distribution on $\{0, 1\}^m$), and $\mathbb{P}_\Theta^{(i)}$ be the law of a random variable Θ under $\mathbb{P}^{(i)}$. For a binary sequence $\sigma_Z = \sigma(Z)$ at a point $Z \in T_i$ where $i \in \{1, 2\}$, we let $f_Z := f_{\sigma(Z)} \in \mathbb{Z}_+^{2^k}$ be the k-mer count vector in $\sigma(Z)$ (see Definition 1). For a finite ordered set of points $U = (u_j)$ on the tree T_i, we let $f_U = (f_{u_j}) \in \mathbb{Z}_+^{2^k \times |U|}$. With this notation, in Theorem 1, $\mathcal{L}_m^{(i)} = \mathbb{P}_{f_X, f_A, f_B, f_C}^{(i)}$. We also write $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

3.2. *Reductions.* Our argument proceeds through a series of reductions.

3.2.1. *Reduction to three vertices.* First we shall reduce the complexity of the problem from n to three vertices. For this we define two internal vertices B' and C' on both T_1 and T_2
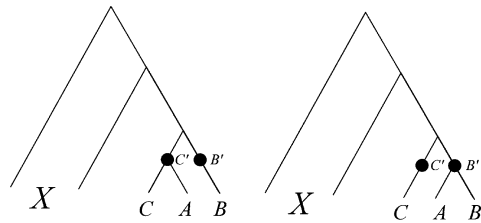
FIG. 3. *The trees T_1 (left) and T_2 (right) on n leaves with points C' and B' added. Here X refers to the remaining $n - 3$ leaves.*

as follows. Let C' be the MRCA of A and C on T_1, and label C' as well the point on the path between C and B on T_2 such that $\text{dist}_{T_1}(C, C') = \text{dist}_{T_2}(C, C')$. Similarly, we let B' be the MRCA of A and B on T_2, and label B' as well the point on the path between A and B on T_1 such that $\text{dist}_{T_1}(B, B') = \text{dist}_{T_2}(B, B')$. This setup is depicted in Figure 3.

Our first reduction lemma asserts that we can reduce the problem to one of distinguishing between the two three-vertex trees depicted in Figure 4.

LEMMA 1 (Reduction to 3 vertices). *Let T_1 and T_2 be the trees with points C' and B' as described above. Then, for all $m \in \mathbb{N}$,*

$$\|\mathcal{L}_m^{(1)} - \mathcal{L}_m^{(2)}\|_{\text{TV}} \leq \|\mathbb{P}_{f_A, f_{B'}, f_{C'}}^{(1)} - \mathbb{P}_{f_A, f_{B'}, f_{C'}}^{(2)}\|_{\text{TV}}.$$

PROOF. First, (f_X, f_A, f_B, f_C) is of course a function of $(f_X, f_A, f_B, f_C, f_{B'}, f_{C'})$, so Lemma 19 in the Appendix implies

$$\|\mathcal{L}_m^{(1)} - \mathcal{L}_m^{(2)}\|_{\text{TV}} = \|\mathbb{P}_{f_X, f_A, f_B, f_C}^{(1)} - \mathbb{P}_{f_X, f_A, f_B, f_C}^{(2)}\|_{\text{TV}}$$

$$\leq \|\mathbb{P}_{f_X, f_A, f_B, f_C, f_{B'}, f_{C'}}^{(1)} - \mathbb{P}_{f_X, f_A, f_B, f_C, f_{B'}, f_{C'}}^{(2)}\|_{\text{TV}}.$$

Also $f_{X,C,B} \to f_{B'C'} \to f_A$ forms a Markov chain under both $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$, satisfying all the conditions of Lemma 20 in the Appendix. Hence

$$\|\mathbb{P}_{f_X, f_A, f_B, f_C, f_{B'}, f_{C'}}^{(1)} - \mathbb{P}_{f_X, f_A, f_B, f_C, f_{B'}, f_{C'}}^{(2)}\|_{\text{TV}} = \|\mathbb{P}_{f_A, f_{B'}, f_{C'}}^{(1)} - \mathbb{P}_{f_A, f_{B'}, f_{C'}}^{(2)}\|_{\text{TV}},$$

giving the result. $\quad\square$

3.2.2. *Reduction to transitions between adjacent, nonoverlapping k-mers.* Due to the following lemma, we can assume $m = (\mu + 1)k$ for some $\mu \in \mathbb{N}$.
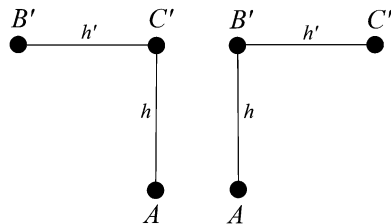

FIG. 4. *The three-vertex configurations for the measures $\mathbb{P}_{f_A, f_{B'}, f_{C'}}^{(1)}$ (left) and $\mathbb{P}_{f_A, f_{B'}, f_{C'}}^{(2)}$ (right) respectively. In this figure, $h' = \text{dist}_{T_1}(B', C') = \text{dist}_{T_2}(B', C')$ and $h = \text{dist}_{T_1}(A, C') = \text{dist}_{T_2}(A, B')$.*

LEMMA 2 (Reduction to multiples of k). *If $\bar{\mu}k < m < (\bar{\mu}+1)k$ where $\bar{\mu} \in \mathbb{N}$, then*

$$\|\mathbb{P}^{(1),m}_{f_A, f_{B'}, f_{C'}} - \mathbb{P}^{(2),m}_{f_A, f_{B'}, f_{C'}}\|_{\mathrm{TV}} \le \|\mathbb{P}^{(1),(\bar{\mu}+1)k}_{\widehat{f}_A, \widehat{f}_{B'}, \widehat{f}_{C'}} - \mathbb{P}^{(2),(\bar{\mu}+1)k}_{\widehat{f}_A, \widehat{f}_{B'}, \widehat{f}_{C'}}\|_{\mathrm{TV}},$$

where $\widehat{f}_V = (f_V, \sigma_V^{\mathrm{last}})$ is the k-mer count vector together with the last $2k$-digits σ_V^{last} in σ_V.

PROOF. Note that all digits are independent under the CFN model and

$$\|\mathbb{P}^{(1),m}_{f_A, f_{B'}, f_{C'}} - \mathbb{P}^{(2),m}_{f_A, f_{B'}, f_{C'}}\|_{\mathrm{TV}} = \|\mathbb{P}^{(1),(\bar{\mu}+1)k}_{f_A^m, f_{B'}^m, f_{C'}^m} - \mathbb{P}^{(2),(\bar{\mu}+1)k}_{f_A^m, f_{B'}^m, f_{C'}^m}\|_{\mathrm{TV}},$$

where f_V^m is the k-mer count vector of the first m digits of σ_V. The proof is complete by Lemma 19 since f_V^m is a function of f_V and the last $2k$-digits when σ_V has length $(\bar{\mu}+1)k$. $\square$

For $\sigma \in \{0,1\}^m$ where $m = (\mu+1)k$, we let $x_0^\sigma, \dots, x_\mu^\sigma \in \{0,1\}^k$ be the *adjacent, nonoverlapping k-mers* in σ. That is,

$$(4) \qquad \sigma = \underbrace{(\sigma_1, \dots, \sigma_k)}_{x_0^\sigma} \underbrace{(\sigma_{k+1}, \dots, \sigma_{2k})}_{x_1^\sigma} \cdots \underbrace{(\sigma_{\mu k+1}, \dots, \sigma_{(\mu+1)k})}_{x_\mu^\sigma} \in \{0,1\}^{(\mu+1)k}.$$

For $y, z \in \{0,1\}^k$, let $N_{y,z}^\sigma$ be the number of adjacent (y,z) pairs in this representation of σ:

$$(5) \qquad N_{y,z}^\sigma = \sum_{j=0}^{\mu-1} \mathbf{1}\{x_j^\sigma = y, x_{j+1}^\sigma = z\}.$$

We call $N_{y,z}^\sigma$ the number of adjacent transitions from y to z.

The following lemma and its proof give an expression for k-mer frequencies in terms of the numbers of adjacent k-mer pairs as well as the ending k-mers.

LEMMA 3 (k-mers as a function of adjacent transitions). *For any $\sigma \in \{0,1\}^{(\mu+1)k}$ and $\mu \in \mathbb{N}$, the frequency vector f_σ is a function of*

$$\big(x_\mu^\sigma, (N_{y,z}^\sigma)_{y,z \in \{0,1\}^k}\big).$$

PROOF. We split the set $\{0, 1, \dots, \mu k\}$ into the disjoint union $(\bigcup_{a=0}^{k-1} \Lambda_a) \cup \{\mu k\}$, where $\Lambda_a = \{a, k+a, 2k+a, \dots, (\mu-1)k+a\}$ contains μ integers with remainder a when divided by k. By definition, for $w = (w_1, \dots, w_k) \in \{0,1\}^k$,

$$(6) \qquad f_\sigma(w) = \sum_{i=0}^{\mu k} \mathbf{1}\{(\sigma_{i+1}, \dots, \sigma_{i+k}) = w\}$$

$$= \mathbf{1}\{x_\mu^\sigma = w\} + \sum_{a=0}^{k-1} \sum_{i \in \Lambda_a} \mathbf{1}\{(\sigma_{i+1}, \dots, \sigma_{i+k}) = w\}.$$

For $a = 0$, the set Λ_0 coincides with the multiples of k from 0 up to $\mu - 1$. So

$$(7) \qquad \sum_{i \in \Lambda_0} \mathbf{1}\{(\sigma_{i+1}, \dots, \sigma_{i+k}) = w\} = \sum_{i=0}^{\mu-1} \mathbf{1}\{x_i^\sigma = w\} = \sum_{z \in \{0,1\}^k} N_{w,z}^\sigma.$$

For $a \in \{1, \dots, k-1\}$,

$$(8) \qquad \sum_{i \in \Lambda_a} \mathbf{1}\{(\sigma_{i+1}, \dots, \sigma_{i+k}) = w\} = \sum_{(y,z) \in \Theta_a(w)} N_{y,z}^\sigma,$$

where $\Theta_a(w)$ is the set of all pairs of the form

$$((\theta_0, \ldots, \theta_{a-1}, w_1, \ldots, w_{k-a}), (w_{k-a+1}, \ldots, w_k, \theta_a, \ldots, \theta_{k-1})) \in \{0, 1\}^{2k},$$

where $(\theta_0, \ldots, \theta_{k-1})$ is an arbitrary element in $\{0, 1\}^k$.

The result then follows when we put (7) and (8) into (6), seeing that $f_\sigma(w)$ depends only on the specified value of $(x_\mu^\sigma, (N_{y,z}^\sigma)_{y,z \in \{0,1\}^k})$. This completes the proof of the lemma. $\quad\square$

For points $V \in \{A, B', C'\}$ on the trees we let

$$Z_V = ((x_0^{\sigma_V}, x_1^{\sigma_V}), (x_{\mu-1}^{\sigma_V}, x_\mu^{\sigma_V}), (N_{y,z}^{\sigma_V})_{y,z \in \{0,1\}^k}),$$

where σ_V is the binary sequence at V. Note that we included $x_0^{\sigma_V}$, $x_1^{\sigma_V}$ and $x_{\mu-1}^{\sigma_V}$ here for reasons that will become clear below.

LEMMA 4 (Reduction to adjacent transitions). *Suppose $m = (\mu + 1)k$ for some $\mu \in \mathbb{N}$.* *Then*

$$\|\mathbb{P}_{\widehat{f}_A, \widehat{f}_{B'}, \widehat{f}_{C'}}^{(1)} - \mathbb{P}_{\widehat{f}_A, \widehat{f}_{B'}, \widehat{f}_{C'}}^{(2)}\|_{\mathrm{TV}} \leq \|\mathbb{P}_{Z_A, Z_{B'}, Z_{C'}}^{(1)} - \mathbb{P}_{Z_A, Z_{B'}, Z_{C'}}^{(2)}\|_{\mathrm{TV}}.$$

PROOF. Recall the definition of $\widehat{f}_V = (f_V, \sigma_V^{\mathrm{last}})$ in Lemma 2. By Lemma 3, $(\widehat{f}_A, \widehat{f}_{B'}, \widehat{f}_{C'})$ is a function of $(Z_A, Z_{B'}, Z_{C'})$. Lemma 19 gives the result. $\quad\square$

3.2.3. *Dealing with redundancy.* The quantities $\{N_{y,z}^\sigma\}_{y,z \in \{0,1\}^k}$ satisfy certain linear relations described in Lemma 5 below. We will get rid of these redundancies in Lemma 6, which will be needed for a nondegeneracy condition in the local CLT; see Lemma 11 below.

LEMMA 5 (Combinatorial constraints). *For any $\sigma \in \{0, 1\}^{(\mu+1)k}$, $\mu \in \mathbb{N}$ and $z \in \{0, 1\}^k$,*

$$(9) \qquad \mathbf{1}\{x_0^\sigma = z\} + \sum_{y \in \{0,1\}^k : y \neq z} N_{y,z}^\sigma = \mathbf{1}\{x_\mu^\sigma = z\} + \sum_{y' \in \{0,1\}^k : y' \neq z} N_{z,y'}^\sigma.$$

Moreover

$$(10) \qquad \sum_{y,z \in \{0,1\}^k} N_{y,z}^\sigma = \mu.$$

PROOF. Equation (10) holds since the total number of adjacent transitions in σ is μ.

To verify (9), we observe that the total count $\sum_{i=0}^\mu \mathbf{1}\{x_i^\sigma = z\}$ can be computed two ways to give

$$\mathbf{1}\{x_0^\sigma = z\} + \sum_{y \in \{0,1\}^k} N_{y,z}^\sigma = \mathbf{1}\{x_\mu^\sigma = z\} + \sum_{y' \in \{0,1\}^k} N_{z,y'}^\sigma.$$

Subtracting $N_{z,z}^\sigma$ from both sides yields (9). $\quad\square$

There are actually only 2^k linearly independent equations among the $2^k + 1$ equations in (9)–(10), as can be seen from the proof of Lemma 6 below. To ensure a nondegenerate limit when applying the central limit theorem, we utilize these 2^k linearly independent equations to remove 2^k redundant variables. Specifically, we remove the transition counts corresponding to the pairs $\{(\vec{1}, z) : z \in \{0, 1\}^k\}$, where $\vec{1} = (1, \ldots, 1) \in \{0, 1\}^k$ is the all-1 string.

LEMMA 6 (Redundancy). *For any $\sigma \in \{0, 1\}^{(\mu+1)k}$ and $\mu \in \mathbb{N}$, the vector $(x_0^\sigma, x_\mu^\sigma,$* $(N_{y,z}^\sigma)_{y,z \in \{0,1\}^k})$ *is a function of $(x_0^\sigma, x_\mu^\sigma, (N_{y,z}^\sigma)_{(y,z) \in \mathcal{H}})$, where*

$$\mathcal{H} = \{(y, z) \in \{0, 1\}^k \times \{0, 1\}^k : y \neq \vec{1}\}.$$

PROOF. It suffices to show that for any $(y, z) \notin \mathcal{H}$, we can write $N_{y,z}^{\sigma}$ as a function of $x_0^{\sigma}, x_{\mu}^{\sigma}, \mu$, and $(N_{y,z}^{\sigma})_{(y,z) \in \mathcal{H}}$. We do this first for $N_{1,z}^{\sigma}$ where $z \neq \vec{1}$, and then for $N_{\vec{1},\vec{1}}^{\sigma}$.

Among the 2^k equations in (9), each one indexed by $z \neq \vec{1}$ has exactly one variable in $\mathcal{H}^c$, namely $N_{1,z}^{\sigma}$. Precisely, (9) gives

$$N_{1,z}^{\sigma} = \mathbf{1}\{x_{\mu}^{\sigma} = z\} - \mathbf{1}\{x_0^{\sigma} = z\} + \sum_{y' \neq z} N_{z,y'}^{\sigma} - \sum_{y \neq z, \vec{1}} N_{y,z}^{\sigma},$$

in which all terms on the right come from $\mathcal{H}$. Hence $N_{1,z}^{\sigma}$ can be written as a function of the required variables for each $z \neq \vec{1}$.

The variable $N_{\vec{1},\vec{1}}^{\sigma}$ is featured only in equation (10), and we obtain

$$N_{\vec{1},\vec{1}}^{\sigma} = \mu - \sum_{(y,z) \neq (\vec{1},\vec{1})} N_{y,z}^{\sigma}. \qquad \square$$

For points $V \in \{A, B', C'\}$ on the trees, we let

(11) $\qquad Z_V' = ((x_0^{\sigma_V}, x_1^{\sigma_V}), (x_{\mu-1}^{\sigma_V}, x_{\mu}^{\sigma_V}), N_{\mathcal{H}}^{\sigma_V}), \quad$ where $N_{\mathcal{H}}^{\sigma_V} = (N_{y,z}^{\sigma_V})_{(y,z) \in \mathcal{H}}.$

LEMMA 7 (Reduction to nonredundant transitions). *Suppose $m = (\mu + 1)k$ for some $\mu \in \mathbb{N}$. Then*

$$\|\mathbb{P}_{Z_A, Z_{B'}, Z_{C'}}^{(1)} - \mathbb{P}_{Z_A, Z_{B'}, Z_{C'}}^{(2)}\|_{\mathrm{TV}} \leq \|\mathbb{P}_{Z_A', Z_{B'}', Z_{C'}'}^{(1)} - \mathbb{P}_{Z_A', Z_{B'}', Z_{C'}'}^{(2)}\|_{\mathrm{TV}}.$$

PROOF. From Lemma 6, $(Z_A, Z_{B'}, Z_{C'})$ is a function of $(Z_A', Z_{B'}', Z_{C'}')$. Then the result follows from Lemma 19. $\square$

3.2.4. *Final reduction step.* By Lemmas 1, 2, 4, and 7 above, together with the second equality of Lemma 18 in the Appendix, to establish Theorem 1 it suffices to prove that

(12) $\qquad \liminf_{\mu \to \infty} \sum_{z_A', z_{B'}', z_{C'}'} \mathbb{P}_{Z_A', Z_{B'}', Z_{C'}'}^{(1)}(z_A', z_{B'}', z_{C'}') \wedge \mathbb{P}_{Z_A', Z_{B'}', Z_{C'}'}^{(2)}(z_A', z_{B'}', z_{C'}') > 0,$

where the sum is taken over the set $(\{0, 1\}^{2k} \times \{0, 1\}^{2k} \times \{0, 1, \ldots, \mu\}^{\mathcal{H}})^3$, and $m = (\mu+1)k$.

Our final reduction step in this section is to condition on the event

(13) $\qquad \widetilde{\mathcal{E}} = \{(x_0^{\sigma_A}, x_1^{\sigma_A}) = (x_0^{\sigma_{B'}}, x_1^{\sigma_{B'}}) = (x_0^{\sigma_{C'}}, x_1^{\sigma_{C'}}) = (\vec{0}, \vec{0})\},$

where $x_j^{\sigma_V} \in \{0, 1\}^k$ are the adjacent k-mers in the sequence σ_V at point $V \in \{A, B', C'\}$, defined in (4). Precisely, for $i \in \{1, 2\}$ we let $\widetilde{\mathbb{P}}^{(i)} = \widetilde{\mathbb{P}}^{(i),m}$ be the conditional measures under $\mathbb{P}^{(i)} = \mathbb{P}^{(i),m}$ given the event $\widetilde{\mathcal{E}}$. Then

$$\sum_{z_A', z_{B'}', z_{C'}'} \mathbb{P}_{Z_A', Z_{B'}', Z_{C'}'}^{(1)}(z_A', z_{B'}', z_{C'}') \wedge \mathbb{P}_{Z_A', Z_{B'}', Z_{C'}'}^{(2)}(z_A', z_{B'}', z_{C'}')$$

$$\geq \sum_{z_A', z_{B'}', z_{C'}'} (\widetilde{\mathbb{P}}_{Z_A', Z_{B'}', Z_{C'}'}^{(1)}(z_A', z_{B'}', z_{C'}')\mathbb{P}^{(1)}[\widetilde{\mathcal{E}}]) \wedge (\widetilde{\mathbb{P}}_{Z_A', Z_{B'}', Z_{C'}'}^{(2)}(z_A', z_{B'}', z_{C'}')\mathbb{P}^{(2)}[\widetilde{\mathcal{E}}])$$

$$\geq c_1 \sum_{z_A', z_{B'}', z_{C'}'} \widetilde{\mathbb{P}}_{Z_A', Z_{B'}', Z_{C'}'}^{(1)}(z_A', z_{B'}', z_{C'}') \wedge \widetilde{\mathbb{P}}_{Z_A', Z_{B'}', Z_{C'}'}^{(2)}(z_A', z_{B'}', z_{C'}'),$$

where $c_1 := \mathbb{P}^{(1)}[\widetilde{\mathcal{E}}] \wedge \mathbb{P}^{(2)}[\widetilde{\mathcal{E}}]$ is positive and does not depend on μ.

Hence, to show (12) it suffices to prove that

$$(14) \quad \liminf_{\mu \to \infty} \sum_{(z'_A, z'_{B'}, z'_{C'}) \in (\mathcal{S}_0^\mu)^3} \widetilde{\mathbb{P}}^{(1)}_{Z'_A, Z'_{B'}, Z'_{C'}} (z'_A, z'_{B'}, z'_{C'}) \wedge \widetilde{\mathbb{P}}^{(2)}_{Z'_A, Z'_{B'}, Z'_{C'}} (z'_A, z'_{B'}, z'_{C'}) > 0,$$

where $\mathcal{S}_0^\mu := \{(\vec{0}, \vec{0})\} \times \{0, 1\}^{2k} \times \{0, 1, \ldots, \mu\}^{\mathcal{H}}$.

For the rest of the proof, we shall establish (14) by obtaining suitable lower bounds on the probabilities in (14) through a local limit theorem.

3.3. *Applying a local limit theorem.*

It will be convenient to consider infinite sequences, since we shall employ a local limit theorem as the sequence length tends to infinity (i.e., $\mu \to \infty$). Let $\mathbb{P}^{(i),\infty}$ be the probability measure of the CFN model on T_i with infinite sequence length and $\widetilde{\mathbb{P}}^{(i),\infty}$ be the conditional measure under $\mathbb{P}^{(i),\infty}$, given the event $\widetilde{\mathcal{E}}$.

3.3.1. *Pairs of triplets as a Markov chain.*

We shall apply Doeblin's method (see, e.g., [7]). For $V \in \{A, B', C'\}$, we let $\sigma_V(n) = (x_0^V, \ldots, x_n^V) \in \{0, 1\}^{k(n+1)}$ be the first $n + 1$ adjacent k-mers of σ_V, where $0 \le n \le \mu$ if σ_V has length $(\mu + 1)k$ and $n \in \mathbb{Z}_+$ if $\sigma_V \in \{0, 1\}^{\mathbb{N}}$ has infinite length. For all such n, we consider the triples

$$(15) \qquad \vec{X}_n = (x_n^A, x_n^{B'}, x_n^{C'}) \in \{0, 1\}^{3k}.$$

Under $\widetilde{\mathbb{P}}^{(i),\infty}$, $\{\vec{X}_n\}_{n \in \mathbb{Z}_+}$ is a sequence of independent random vectors and the pairs $\vec{M}_n = (\vec{X}_n, \vec{X}_{n+1})$ form a Markov chain with a finite state space. This Markov chain is irreducible since the support of $(\vec{X}_n, \vec{X}_{n+1})$ is all of $\{0, 1\}^{3k} \times \{0, 1\}^{3k}$ for all n. The stationary distribution $\Theta_{\vec{M}}$ of $\{\vec{M}_n\}_{n \in \mathbb{Z}_+}$ is

$$(16) \qquad \Theta_{\vec{M}}(\vec{y}, \vec{z}) = \widetilde{\mathbb{P}}^{(i),\infty}(\vec{X}_2 = \vec{y}) \widetilde{\mathbb{P}}^{(i),\infty}(\vec{X}_2 = \vec{z}), \quad \text{for } \vec{y}, \vec{z} \in \{0, 1\}^{3k}.$$

Let $\tau_0 = 0$, let τ_1 be the first $n > 0$ such that $\vec{M}_n = (\vec{0}, \vec{0})$, and in general, for $\ell \ge 1$, let

$$(17) \qquad \tau_\ell = \inf\{n > \tau_{\ell-1} : \vec{M}_n = (\vec{0}, \vec{0})\},$$

where an infimum over an empty set is $+\infty$ by convention.

The connection between $\widetilde{\mathbb{P}}^{(i)} = \widetilde{\mathbb{P}}^{(i),m}$ and $\widetilde{\mathbb{P}}^{(i),\infty}$ that we will need is given by Lemma 8 below. We let

$$N_{y,z}^V(n) = N_{y,z}^{\sigma_V(n)} = \sum_{j=0}^{n-1} \mathbf{1}\{x_j^V = y, x_{j+1}^V = z\}$$

be the number of adjacent transitions from y to z up to x_n^V, as in (5), with the convention that $N_{y,z}^V(0) = 0$.

LEMMA 8 (Infinite sequences). *For all $\mu \in \mathbb{Z}_+$, $k \in \mathbb{N}$, $\ell \in \{1, 2, \ldots, \mu\}$, $(a, b, c) \in \mathbb{Z}_+^3$, and $i \in \{1, 2\}$, the event*

$$\{\tau_\ell = \mu, (N_{\mathcal{H}}^A(\tau_\ell), N_{\mathcal{H}}^{B'}(\tau_\ell), N_{\mathcal{H}}^{C'}(\tau_\ell)) = (a, b, c)\}$$

has the same probability under $\widetilde{\mathbb{P}}^{(i),m}$ and $\widetilde{\mathbb{P}}^{(i),\infty}$, where $m = (\mu + 1)k$.

This lemma follows directly from the construction of the CFN model, in which nonoverlapping, adjacent k-mers are independent. The rest of Section 3.3 concerns infinite sequences.

3.3.2. *Independent excursions and a multivariate local CLT.* We extract i.i.d. random variables from excursions of the Markov chain $\vec{M}$. Define, for $V \in \{A, B', C'\}$,

$$Y_V(\ell) = N_{\mathcal{H}}^V(\tau_\ell) - N_{\mathcal{H}}^V(\tau_{\ell-1}) = \left(N_{y,z}^V(\tau_\ell) - N_{y,z}^V(\tau_{\ell-1})\right)_{(y,z) \in \mathcal{H}}$$

and let

$$\mathbf{Y}(\ell) = \left(\tau_\ell - \tau_{\ell-1}, Y_A(\ell), Y_{B'}(\ell), Y_{C'}(\ell)\right).$$

Note that these random vectors take values in $\mathbb{N} \times (\mathbb{Z}_+^{\mathcal{H}})^3 \subset \mathbb{Z}_+^d$ where $d = 1 + 3(2^{2k} - 2^k)$, because $|\mathcal{H}| = 2^{2k} - 2^k$.

LEMMA 9. *The vectors $\{\mathbf{Y}(\ell)\}_{\ell=1}^\infty$ are i.i.d. under $\widetilde{\mathbb{P}}^{(i),\infty}$ for both $i = 1$ and 2. Further, their partial sum is equal to*

$$(18) \qquad \sum_{\iota=1}^{\ell} \mathbf{Y}(\iota) = \left(\tau_\ell, N_{\mathcal{H}}^A(\tau_\ell), N_{\mathcal{H}}^{B'}(\tau_\ell), N_{\mathcal{H}}^{C'}(\tau_\ell)\right).$$

PROOF. The first statement is obvious from the construction of the CFN model. The equality (18) follows from the definitions of $\mathbf{Y}$ and the conventions $\tau_0 = N_{y,z}^V(0) = 0$. $\square$

We will apply a multivariate local CLT of Davis and McDonald [9], Theorem 2.1, to the i.i.d. vectors $\{\mathbf{Y}(\ell)\}_{\ell=1}^\infty \subset \mathbb{Z}_+^d$ under $\widetilde{\mathbb{P}}^{(i),\infty}$. Theorem 2.1 of [9] works for an array of independent vectors. Here we need only a sequence of i.i.d. vectors so we state this result for the case of i.i.d. vectors in $\mathbb{Z}^d$.

THEOREM 2 ([9], Theorem 2.1). *Let $\{\mathbf{X}_j\}_{j=1}^\infty$ be a sequence of independent $\mathbb{Z}^d$-valued random variables with a common probability mass function f, finite mean $\mathbf{m} \in \mathbb{R}^d$, and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$. Suppose the following hold:*

(a) *For all $r \in \{1, 2, \ldots, d\}$, there exists $\mathbf{x}_r \in \mathbb{Z}^d$ such that*

$$f(\mathbf{x}_r) \wedge f(\mathbf{x}_r + \mathbf{e}_r) > 0,$$

where $\mathbf{e}_r \in \mathbb{Z}^d$ is the rth standard basis vector.
(b) *The sequence $\frac{\mathbf{S}_\ell - \ell \mathbf{m}}{\sqrt{\ell}}$ converges in distribution to the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ as $\ell \to \infty$, where $\mathbf{S}_\ell = \sum_{j=1}^\ell \mathbf{X}_j$.*
Then the following uniform convergence holds as $\ell \to \infty$:

$$\sup_{\mathbf{y} \in \mathbb{Z}^d} \left| \ell^{d/2} \mathbb{P}[\mathbf{S}_\ell = \mathbf{y}] - \varphi\left(\frac{\mathbf{y} - \ell \mathbf{m}}{\sqrt{\ell}}\right) \right| \to 0,$$

where φ is the probability density function of the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$.

Condition (a) of Theorem 2 implies that the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ is nondegenerate.

LEMMA 10. *Let f be a probability mass function on $\mathbb{Z}^d$ with finite mean and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$. Assume condition (a) of Theorem 2 holds. Then $\mathbf{\Sigma}$ is positive definite.*

PROOF. Let X and Y be two independent random vectors with distribution f. Then the covariance matrix of X can be written as

$$\mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T] = (1/2)\mathbb{E}[(X - Y)(X - Y)^T].$$

Let $\mathbf{x}_r$ be as in condition (a) of Theorem 2. Then for any nonzero vector $\mathbf{z} = (z_1, \ldots, z_d) \neq 0$ with, say, $z_r \neq 0$, we have

$$\mathbf{z}^T \mathbb{E}[(X - Y)(X - Y)^T]\mathbf{z} = \mathbb{E}[(\mathbf{z}^T(X - Y))^2]$$
$$\geq f(\mathbf{x}_r)f(\mathbf{x}_r + \mathbf{e}_r)z_r^2$$
$$> 0,$$

where the expression on the second line is the contribution to the expectation from the event that $X = \mathbf{x}_r + \mathbf{e}_r$ and $Y = \mathbf{x}_r$, and the third line follows from condition (a) of Theorem 2. Note that we used that each term contributing to the expectation is nonnegative.   □

We shall apply Theorem 2 to the i.i.d. vectors $\{\mathbf{Y}(\ell)\}_{\ell=1}^{\infty} \subset \mathbb{Z}_+^d$ under $\widetilde{\mathbb{P}}^{(i),\infty}$, for each of $i \in \{1, 2\}$.

3.3.3. *Checking conditions of the local CLT.*   In this section we verify that the i.i.d. vectors $\{\mathbf{Y}(\ell)\}_{\ell=1}^{\infty} \subset \mathbb{Z}_+^d$ satisfy all conditions of Theorem 2. We also show that they have the same mean under $\widetilde{\mathbb{P}}^{(1),\infty}$ and $\widetilde{\mathbb{P}}^{(2),\infty}$. For this we let $f^{(i)}$ be the probability mass function of

$$\mathbf{Y}(1) = (\tau, Y_A(1), Y_{B'}(1), Y_{C'}(1)) = (\tau, N_{\mathcal{H}}^A(\tau), N_{\mathcal{H}}^{B'}(\tau), N_{\mathcal{H}}^{C'}(\tau))$$

under $\widetilde{\mathbb{P}}^{(i),\infty}$ for $i \in \{1, 2\}$, where $\tau = \tau_1$ is defined in (17).

LEMMA 11 (Nondegeneracy).   *The distributions $f^{(1)}$ and $f^{(2)}$ both satisfy condition (a) of Theorem 2.*

PROOF.   Fix $i \in \{1, 2\}$. The proof relies crucially on the construction of the set $\mathcal{H}$ in Lemma 6. We write a point in $\mathbb{Z}_+^d$ as

$$\mathbf{x} = (t, (n_{yz}^A, n_{yz}^{B'}, n_{yz}^{C'})_{yz \in \mathcal{H}}), \quad \text{where } t \in \mathbb{Z}_+ \text{ and } n_{yz}^A, n_{yz}^{B'}, n_{yz}^{C'} \in \mathbb{Z}_+.$$

Recall that $\vec{0}$ and $\vec{1}$ refer to the all-0 and all-1 k-mers. A sequence of adjacent k-mer triples starting and ending with $(\vec{0}, \vec{0}, \vec{0})(\vec{0}, \vec{0}, \vec{0})$ will give rise to a unique point in $\mathbb{Z}_+^d$, in which t is the length of the sequence and n_{yz}^V counts the number of yz-transitions. By the definition of $\mathcal{H}$, we are not counting the transition from $\vec{1}$ to z for any $z \in \{0, 1\}^k$.

For $r = 1$ (corresponding to the t-coordinate), we consider the k-mer triple cycles of

$$\mathcal{C} = (\vec{0}, \vec{0}, \vec{0}), (\vec{0}, \vec{0}, \vec{0}), (\vec{1}, \vec{1}, \vec{1}), (\vec{0}, \vec{0}, \vec{0}), (\vec{0}, \vec{0}, \vec{0}) \quad \text{and}$$
$$\mathcal{C}^+ = (\vec{0}, \vec{0}, \vec{0}), (\vec{0}, \vec{0}, \vec{0}), (\vec{1}, \vec{1}, \vec{1}), (\vec{1}, \vec{1}, \vec{1}), (\vec{0}, \vec{0}, \vec{0}), (\vec{0}, \vec{0}, \vec{0}).$$

They give rise to $\mathbf{x}_r$ and $\mathbf{x}_r + \mathbf{e}_r$ respectively, where we take $\mathbf{x}_r$ to be the point in $\mathbb{Z}_+^d$ such that $t = 3$ and

$$(19) \qquad (n_{yz}^A, n_{yz}^{B'}, n_{yz}^{C'}) = \begin{cases} (2, 2, 2) & \text{if } (y, z) = (\vec{0}, \vec{0}), \\ (1, 1, 1) & \text{if } (y, z) = (\vec{0}, \vec{1}), \\ (0, 0, 0) & \text{if } (y, z) \in \mathcal{H} \setminus \{(\vec{0}, \vec{0}), (\vec{0}, \vec{1})\}, \end{cases}$$

and $\mathbf{x}_r + \mathbf{e}_r = (4, (n_{yz}^A, n_{yz}^{B'}, n_{yz}^{C'})_{yz \in \mathcal{H}})$. Recall that

$$\mathcal{H} = \{(y, z) \in \{0, 1\}^k \times \{0, 1\}^k : y \neq \vec{1}\},$$

so that, in particular, the transitions $(\vec{1}, \vec{1})$ are not counted. Then

$$f^{(i)}(\mathbf{x}_r) \geq \widetilde{\mathbb{P}}^{(i),\infty}((\vec{X}_n)_{n=0}^3 = \mathcal{C}) > 0 \quad \text{and}$$

$$f^{(i)}(\mathbf{x}_r + \mathbf{e}_r) \geq \widetilde{\mathbb{P}}^{(i),\infty}((\vec{X}_n)_{n=0}^4 = \mathcal{C}^+) > 0,$$

where $\vec{X}_n = (x_n^A, x_n^{B'}, x_n^{C'}) \in \{0, 1\}^{3k}$ as defined in (15).

For $r > 1$, we first suppose r corresponds to the coordinate n_{ab}^A where $(a, b) \in \mathcal{H}$. The cycles

$$\mathcal{C}_{ab}^A = (\vec{0}, \vec{0}, \vec{0}), (\vec{0}, \vec{0}, \vec{0}), (\vec{1}, \vec{1}, \vec{1}), (\vec{1}, \vec{1}, \vec{1}), (b, \vec{1}, \vec{1}), (\vec{0}, \vec{0}, \vec{0}), (\vec{0}, \vec{0}, \vec{0}) \quad \text{and}$$

$$\mathcal{C}_{ab}^{A+} = (\vec{0}, \vec{0}, \vec{0}), (\vec{0}, \vec{0}, \vec{0}), (\vec{1}, \vec{1}, \vec{1}), (a, \vec{1}, \vec{1}), (b, \vec{1}, \vec{1}), (\vec{0}, \vec{0}, \vec{0}), (\vec{0}, \vec{0}, \vec{0})$$

give rise to $\mathbf{x}_r$ and $\mathbf{x}_r + \mathbf{e}_r$ respectively, where $\mathbf{x}_r$ is the point on $\mathbb{Z}_+^d$ such that $t = 5$ and (19) hold. Hence both $f^{(i)}(\mathbf{x}_r)$ and $f^{(i)}(\mathbf{x}_r + \mathbf{e}_r)$ are positive, as before.

The proof for coordinates $n_{ab}^{B'}$ is the same, except that we replace $(a, \vec{1}, \vec{1})$ by $(\vec{1}, a, \vec{1})$ and $(b, \vec{1}, \vec{1})$ by $(\vec{1}, b, \vec{1})$. The proof for coordinates $n_{ab}^{C'}$ follows similarly. The proof is complete. $\square$

To verify condition (b) of Theorem 2, we let $\mathbf{m}^{(i)}$ and $\mathbf{\Sigma}^{(i)}$ be respectively the mean and the covariance matrix of $\mathbf{Y}(1)$ under $\widetilde{\mathbb{P}}^{(i),\infty}$. We also let $\mathbf{S}_\ell = \sum_{j=1}^\ell \mathbf{Y}(j)$.

LEMMA 12. *For $i \in \{1, 2\}$, under $\widetilde{\mathbb{P}}^{(i),\infty}$, the sequence $\frac{\mathbf{S}_\ell - \ell \mathbf{m}^{(i)}}{\sqrt{\ell}}$ converges in distribution to the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{(i)})$ as $\ell \to \infty$.*

PROOF. Fix $i \in \{1, 2\}$. Observe that $\mathbf{Y}(1) \leq (\tau_1, \tau_1, \ldots, \tau_1)$ coordinate-wise. Moreover, by construction, τ_1 is geometric and therefore has finite first and second moments. Hence $\mathbf{m}^{(i)}$ is finite and $\widetilde{\mathbb{E}}^{(i),\infty}[\|\mathbf{Y}(\ell)\|^2] < \infty$, from which we have that the entries of $\mathbf{\Sigma}^{(i)}$ are finite and hence $|\det(\mathbf{\Sigma}^{(i)})| < \infty$. Also $\mathbf{\Sigma}^{(i)}$ is positive definite by Lemmas 10 and 11. The claim follows from the multivariate central limit theorem (see, e.g., [12], Section 3.10). $\square$

Due to symmetry between T_1 and T_2, as well as the independence of nonoverlapping, adjacent k-mers under the CFN model, the expectations are the same, as we show formally next.

LEMMA 13 (Expectation). *The equality $\mathbf{m}^{(1)} = \mathbf{m}^{(2)} \in \mathbb{R}_+^d$ holds.*

PROOF. By symmetry (3), we have $\widetilde{\mathbb{P}}^{(1),\infty}_{(\tau,\sigma_A,\sigma_{B'},\sigma_{C'})} = \widetilde{\mathbb{P}}^{(2),\infty}_{(\tau,\sigma_A,\sigma_{C'},\sigma_{B'})}$. Hence

$$\widetilde{\mathbb{E}}^{(1),\infty}[(\tau, N_{\mathcal{H}}^A(\tau))] = \widetilde{\mathbb{E}}^{(2),\infty}[(\tau, N_{\mathcal{H}}^A(\tau))],$$

and

$$\widetilde{\mathbb{E}}^{(1),\infty}[(N_{\mathcal{H}}^{B'}(\tau), N_{\mathcal{H}}^{C'}(\tau))] = \widetilde{\mathbb{E}}^{(2),\infty}[(N_{\mathcal{H}}^{C'}(\tau), N_{\mathcal{H}}^{B'}(\tau))].$$

It remains to show that

(20) $$\widetilde{\mathbb{E}}^{(i),\infty}[N_{\mathcal{H}}^{B'}(\tau)] = \widetilde{\mathbb{E}}^{(i),\infty}[N_{\mathcal{H}}^{C'}(\tau)] \quad \text{for } i \in \{1, 2\}.$$

While $\widetilde{\mathbb{P}}^{(i),\infty}_{\sigma_{B'}} = \widetilde{\mathbb{P}}^{(i),\infty}_{\sigma_{C'}}$, equation (20) is not immediately clear because τ depends on all three sequences.

Using the notation of Section 3.3.1, for arbitrary $(y, z) \in \{0, 1\}^k \times \{0, 1\}^k$, we have

$$(21) \qquad N_{(y,z)}^{B'}(\tau) = \sum_{(\vec{y}, \vec{z}):(y_2, z_2)=(y,z)} \sum_{j=0}^{\tau-1} 1_{\vec{M}_j=(\vec{y}, \vec{z})},$$

where the sum is over the set of $(\vec{y}, \vec{z})$ with $y_2 = y$ and $z_2 = z$, with $\vec{y} = (y_1, y_2, y_3) \in \{0, 1\}^{3k}$ and $\vec{z} = (z_1, z_2, z_3) \in \{0, 1\}^{3k}$. Using standard Markov chain results (e.g., [12], Chapter 5),

$$(22) \qquad \widetilde{\mathbb{E}}^{(i),\infty}\left[\sum_{j=0}^{\tau-1} 1_{\vec{M}_j=(\vec{y}, \vec{z})}\right] = \widetilde{c}\,\Theta_{\vec{M}}(\vec{y}, \vec{z}),$$

where $\widetilde{c} = \widetilde{\mathbb{E}}^{(i),\infty}[\tau] \in (0, \infty)$ and the stationary distribution $\Theta_{\vec{M}}(\vec{y}, \vec{z})$ was computed in (16). Combining (16), (21), and (22), we have

$$\widetilde{\mathbb{E}}^{(i),\infty}[N_{y,z}^{B'}(\tau)] = \widetilde{c} \sum_{(\vec{y}, \vec{z}):(y_2, z_2)=(y,z)} \widetilde{\mathbb{P}}^{(i),\infty}(\vec{X}_2 = \vec{y})\widetilde{\mathbb{P}}^{(i),\infty}(\vec{X}_2 = \vec{z})$$

$$= \widetilde{c}\,\widetilde{\mathbb{P}}^{(i),\infty}(x_2^{B'} = y)\widetilde{\mathbb{P}}^{(i),\infty}(x_2^{B'} = z)$$

and, similarly for C',

$$\widetilde{\mathbb{E}}^{(i),\infty}[N_{y,z}^{C'}(\tau)] = \widetilde{c}\,\widetilde{\mathbb{P}}^{(i),\infty}(x_2^{C'} = y)\widetilde{\mathbb{P}}^{(i),\infty}(x_2^{C'} = z).$$

The two displayed equations are the same since $\widetilde{\mathbb{P}}_{\sigma_{B'}}^{(i),\infty} = \widetilde{\mathbb{P}}_{\sigma_{C'}}^{(i),\infty}$. $\quad\square$

3.3.4. *Applying the local CLT.* By Lemmas 11 and 12, we can apply Theorem 2 to the i.i.d. vectors $\{\mathbf{Y}(j)\}_{j=1}^{\infty}$ to obtain the following lower bound. Recall that $\mathbf{m}^{(1)} = \mathbf{m}^{(2)}$ by Lemma 13, and let $\mathbf{m} = \mathbf{m}^{(i)}$. Recall also that $\mathbf{S}_\ell = \sum_{j=1}^{\ell} \mathbf{Y}(j)$.

LEMMA 14 (Uniform lower bound). *There exist constants $c_1, c_2 \in (0, \infty)$ such that*

$$\inf_{y \in \mathcal{Y}_\ell^{(i)}} \widetilde{\mathbb{P}}^{(i),\infty}[\mathbf{S}_\ell = \mathbf{y}] \geq \frac{c_2}{\ell^{d/2}}$$

for all $\ell \geq c_1$ and $i \in \{1, 2\}$, where

$$(23) \qquad \mathcal{Y}_\ell^{(i)} := \{\mathbf{y} \in \mathbb{Z}_+^d : (\mathbf{y} - \ell\mathbf{m})^{\mathrm{T}}(\mathbf{\Sigma}^{(i)})^{-1}(\mathbf{y} - \ell\mathbf{m}) \leq 2\ell\}.$$

PROOF. By Theorem 2, for $i \in \{1, 2\}$, as $\ell \to \infty$,

$$(24) \qquad \sup_{\mathbf{y} \in \mathbb{Z}^d}\left|\ell^{d/2}\widetilde{\mathbb{P}}^{(i),\infty}[\mathbf{S}_\ell = \mathbf{y}] - \varphi^{(i)}\left(\frac{\mathbf{y} - \ell\mathbf{m}}{\sqrt{\ell}}\right)\right| \to 0,$$

where

$$\varphi^{(i)}(\mathbf{X}) = \frac{\exp\{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}(\mathbf{\Sigma}^{(i)})^{-1}\mathbf{x}\}}{\sqrt{(2\pi)^d \det(\mathbf{\Sigma}^{(i)})}}.$$

Therefore, for arbitrary $\epsilon > 0$, there exists ℓ_ϵ sufficiently large such that for all integers $\ell \geq \ell_\epsilon$ and all $\mathbf{y} \in \mathbb{Z}^d$,

$$\widetilde{\mathbb{P}}^{(i),\infty}[\mathbf{S}_\ell = \mathbf{y}] \geq \frac{1}{\ell^{d/2}}\left(\varphi^{(i)}\left(\frac{\mathbf{y} - \ell\mathbf{m}}{\sqrt{\ell}}\right) - \epsilon\right)$$

$$= \frac{1}{\ell^{d/2}}\left(\frac{\exp\{-\frac{1}{2}(\frac{\mathbf{y}-\ell\mathbf{m}}{\sqrt{\ell}})^{\mathrm{T}}(\mathbf{\Sigma}^{(i)})^{-1}(\frac{\mathbf{y}-\ell\mathbf{m}}{\sqrt{\ell}})\}}{\sqrt{(2\pi)^d \det(\mathbf{\Sigma}^{(i)})}} - \epsilon\right).$$

The bound in the definition of $\mathcal{Y}_\ell^{(i)}$ gives

$$\inf_{y \in \mathcal{Y}_\ell^{(i)}} \widetilde{\mathbb{P}}^{(i),\infty}[\mathbf{S}_\ell = \mathbf{y}] \geq \frac{1}{\ell^{d/2}} \left(\frac{e^{-1}}{\sqrt{(2\pi)^d [\det(\mathbf{\Sigma}^{(1)}) \vee \det(\mathbf{\Sigma}^{(2)})]}} - \epsilon \right)$$

for all $\ell \geq \ell_\epsilon$. The lemma follows by taking ϵ to be any fixed number small enough that depends only on $\det(\mathbf{\Sigma}^{(1)}) \vee \det(\mathbf{\Sigma}^{(2)})$. $\square$

Observe that the bound in Lemma 14 is uniform over the set $\mathcal{Y}_\ell^{(i)}$. Our use of Lemma 14 below will require a lower bound on the size of $\mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)}$.

LEMMA 15. *Let $\lambda_{\min}^{(i)}$ be the minimal eigenvalues of $\mathbf{\Sigma}^{(i)}$. Then*

$$(25) \qquad \{\mathbf{y} \in \mathbb{Z}_+^d : \|\mathbf{y} - \ell \mathbf{m}\|^2 \leq 2\ell(\lambda_{\min}^{(1)} \wedge \lambda_{\min}^{(2)})\} \subset \mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)},$$

where $\{\mathcal{Y}_\ell^{(i)}\}_{i=1}^2$ are defined in (23).

PROOF. Note that $0 < \lambda_{\min}^{(i)} < \infty$ by Lemma 12. Since λ is an eigenvalue of $\mathbf{\Sigma}^{(i)}$ if and only if $1/\lambda$ is an eigenvalue of $(\mathbf{\Sigma}^{(i)})^{-1}$, we have

$$(\mathbf{y} - \ell \mathbf{m})^{\mathrm{T}} (\mathbf{\Sigma}^{(i)})^{-1} (\mathbf{y} - \ell \mathbf{m}) \leq \frac{1}{\lambda_{\min}^{(i)}} \|\mathbf{y} - \ell \mathbf{m}\|^2.$$

This inequality implies (25). $\square$

In fact, we will need to control the size of subsets of $\mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)}$ whose first coordinates are sufficiently close to their expectation. Letting m_1 be the first coordinate of $\mathbf{m}$, by Lemma 13,

$$(26) \qquad m_1 = \widetilde{\mathbb{E}}^{(1),\infty}[\tau_1] = \widetilde{\mathbb{E}}^{(2),\infty}[\tau_1].$$

We consider the following set of pairs (μ, ℓ)

$$(27) \qquad \mathcal{L} = \{(\mu, \ell) \in \mathbb{N}^2 : |\mu - \ell m_1| \leq c_3 \sqrt{\ell}\}, \quad \text{where } c_3 = \sqrt{\lambda_{\min}^{(1)} \wedge \lambda_{\min}^{(2)}}.$$

The next two lemmas concern bounds on the level sets

$$\mathcal{L}|_\ell := \{\mu \in \mathbb{N} : (\mu, \ell) \in \mathcal{L}\} \quad \text{and} \quad \mathcal{L}|_\mu := \{\ell \in \mathbb{N} : (\mu, \ell) \in \mathcal{L}\}.$$

LEMMA 16. *Let $\mathbb{Z}_+^d(\mu)$ be the subset of $\mathbb{Z}_+^d$ whose first coordinate is μ. Then*

$$(28) \qquad \inf_{\mu \in \mathcal{L}|_\ell} |\mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)} \cap \mathbb{Z}_+^d(\mu)| \geq c_4 c_3^{d-1} \ell^{(d-1)/2}$$

for all $\ell \in \mathbb{N}$, where $c_4 \in (0, \infty)$ is a constant that depends only on d.

PROOF. By Lemma 15, the set $\mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)}$ contains all integer points of

$$B_d(\ell \mathbf{m}, c_3 \sqrt{2\ell}) \cap \mathbb{R}_+^d,$$

where $B_d(x, r) := \{y \in \mathbb{R}^d : \|y - x\|_{\mathbb{R}^d} \leq r\}$ is the d-dimensional Euclidean ball with center x and radius r. By Lemma 13, $\mathbf{m} = (m_1, m_2, \ldots, m_d) \in \mathbb{R}_+^d$. Hence $\mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)} \cap \mathbb{Z}_+^d(\mu)$ contains all integer points of $\widetilde{B}(r_\mu) \cap \mathbb{R}_+^d$, where

$$\widetilde{B}(r_\mu) := \{(\mu, y_2, \ldots, y_d) \in \mathbb{R}^d : \|(y_2, \ldots, y_d) - \ell(m_2, \ldots, m_d)\|_{\mathbb{R}^{d-1}} \leq r_\mu\}$$

with

$$(29) \qquad r_\mu := \sqrt{c_3^2 2\ell - (\mu - \ell m_1)^2} \geq c_3\sqrt{\ell}.$$

The last inequality follows whenever $(\mu, \ell) \in \mathcal{L}$.

Since $\{m_2, \ldots, m_d\}$ are all nonnegative, $\widetilde{B}(r_\mu) \cap \mathbb{R}_+^d$ contains a $(d-1)$-dimensional cube with side length $\frac{r_\mu}{\sqrt{d-1}} \geq \frac{c_3\sqrt{\ell}}{\sqrt{d-1}}$ by (29). This cube contains at least $c_4 c_3^{d-1} \ell^{(d-1)/2}$ many integer points for some $c_4 \in (0, \infty)$ that depends only on d, uniformly for all $(\mu, \ell) \in \mathcal{L}$. $\quad\square$

Finally, the following lemma gives a lower bound on the cardinality of $\mathcal{L}|_\mu$.

LEMMA 17. *There exists a constant $c_5 \in (0, \infty)$ that depends only on m_1 and c_3 such that, for μ large enough,*

$$\left[\frac{\mu}{m_1} - c_5\sqrt{\mu}, \frac{\mu}{m_1} + c_5\sqrt{\mu}\right] \subseteq \mathcal{L}|_\mu,$$

where $m_1 = \widetilde{\mathbb{E}}^{(1),\infty}[\tau_1] = \widetilde{\mathbb{E}}^{(2),\infty}[\tau_1]$.

PROOF. Suppose ℓ belongs to the interval on the left-hand side of the display in the statement of the lemma. Then $\ell \geq \frac{\mu}{m_1} - c_5\sqrt{\mu}$. Solving this quadratic inequality in $\sqrt{\mu}$ and then squaring gives

$$\sqrt{\mu} \leq \frac{c_5 m_1 + \sqrt{(c_5 m_1)^2 + 4 m_1 \ell}}{2},$$

and

$$\mu \leq \ell m_1 + \frac{1}{2}(c_5 m_1)^2 + \frac{c_5 m_1 \sqrt{(c_5 m_1)^2 + 4 m_1 \ell}}{2}.$$

From the last inequality, we see that $\mu \leq \ell m_1 + c_3\sqrt{\ell}$ for all $\ell \geq 1$, provided that $c_5 \in (0, \infty)$ is small enough (depending only on c_3 and m_1).

Similarly, by solving the inequality $\ell \leq \frac{\mu}{m_1} + c_5\sqrt{\mu}$ to yield

$$\sqrt{\mu} \geq \frac{-c_5 m_1 + \sqrt{(c_5 m_1)^2 + 4 m_1 \ell}}{2},$$

and

$$\mu \geq \ell m_1 + \frac{1}{2}(c_5 m_1)^2 - \frac{c_5 m_1 \sqrt{(c_5 m_1)^2 + 4 m_1 \ell}}{2}.$$

For sufficiently small $c_5 \in (0, \infty)$ (depending only on c_3 and m_1), we have $\mu \geq \ell m_1 - c_3\sqrt{\ell}$.

The desired subset relation is obtained. $\quad\square$

3.4. *Final bound on the total variation distance.* PROOF OF THEOREM 1. Now we finish the proof of Theorem 1 by establishing (14). That is, we now show that

$$(30) \qquad \liminf_{\mu \to \infty} \sum_{(z_A', z_{B'}', z_{C'}') \in (\mathcal{S}_0^\mu)^3} \widetilde{\mathbb{P}}^{(1)}_{Z_A', Z_{B'}', Z_{C'}'}(z_A', z_{B'}', z_{C'}') \wedge \widetilde{\mathbb{P}}^{(2)}_{Z_A', Z_{B'}', Z_{C'}'}(z_A', z_{B'}', z_{C'}') > 0,$$

where $\mathcal{S}_0^\mu := \{(\vec{0}, \vec{0})\} \times \{0, 1\}^{2k} \times \{0, 1, \ldots, \mu\}^{\mathcal{H}}$ and $m = (\mu + 1)k$. We further restrict the last pair of triples by considering $\mathcal{S}_{00}^\mu := \{(\vec{0}, \vec{0})\} \times \{(\vec{0}, \vec{0})\} \times \{0, 1, \ldots, \mu\}^{\mathcal{H}}$. Since $\mathcal{S}_{00}^\mu \subset \mathcal{S}_0^\mu$,

the sum $\sum_{(z'_A, z'_{B'}, z'_{C'}) \in (\mathcal{S}_0^\mu)^3}$ on the left of (30) is bounded below by

$$(31) \qquad W_{00} = \sum_{(z'_A, z'_{B'}, z'_{C'}) \in (\mathcal{S}_{00}^\mu)^3} \widetilde{\mathbb{P}}^{(1)}_{Z'_A, Z'_{B'}, Z'_{C'}}(z'_A, z'_{B'}, z'_{C'}) \wedge \widetilde{\mathbb{P}}^{(2)}_{Z'_A, Z'_{B'}, Z'_{C'}}(z'_A, z'_{B'}, z'_{C'}).$$

As an element of $\mathcal{S}_{00}^\mu$, $z'_V = ((\vec{0}, \vec{0}), (\vec{0}, \vec{0}), N'_V)$ for some $N'_V \in \{0, 1, \ldots, \mu\}^{\mathcal{H}}$, where $V \in \{A, B', C'\}$. Hence

$$\widetilde{\mathbb{P}}^{(i)}_{Z'_A, Z'_{B'}, Z'_{C'}}(z'_A, z'_{B'}, z'_{C'})$$

$$= \sum_{\ell=1}^\mu \widetilde{\mathbb{P}}^{(i)}\{(N_{\mathcal{H}}^{\sigma_A}(\tau_\ell), N_{\mathcal{H}}^{\sigma_{B'}}(\tau_\ell), N_{\mathcal{H}}^{\sigma_{C'}}(\tau_\ell)) = (N'_A, N'_{B'}, N'_{C'}), \tau_\ell = \mu\}$$

$$= \sum_{\ell=1}^\mu \widetilde{\mathbb{P}}^{(i),\infty}\{(N_{\mathcal{H}}^{\sigma_A}(\tau_\ell), N_{\mathcal{H}}^{\sigma_{B'}}(\tau_\ell), N_{\mathcal{H}}^{\sigma_{C'}}(\tau_\ell)) = (N'_A, N'_{B'}, N'_{C'}), \tau_\ell = \mu\}$$

$$= \sum_{\ell=1}^\mu \widetilde{\mathbb{P}}^{(i),\infty}\left\{\sum_{j=1}^\ell \mathbf{Y}(j) = (\mu, N'_A, N'_{B'}, N'_{C'})\right\},$$

where the second and the last equalities follow from Lemma 8 and (18) respectively. Therefore,

$$(32) \qquad W_{00} \geq \sum_{\ell=1}^\mu \sum_{\mathbf{y} \in \mathbb{Z}_+^d(\mu)} \widetilde{\mathbb{P}}^{(1),\infty}\left\{\sum_{j=1}^\ell \mathbf{Y}(j) = \mathbf{y}\right\} \wedge \widetilde{\mathbb{P}}^{(2),\infty}\left\{\sum_{j=1}^\ell \mathbf{Y}(j) = \mathbf{y}\right\},$$

where recall that $\mathbb{Z}_+^d(\mu)$ was defined in Lemma 16.

We further restrict the sums to be over $(\mu, N'_A, N'_{B'}, N'_{C'}) \in \mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)}$ and $\ell \geq c_1$, where recall that $\mathcal{Y}_\ell^{(i)}$ and c_1 were defined in Lemma 14. We obtain from Lemma 14 that the right-hand side of (32) is

$$\geq \sum_{\ell \in [c_1, \mu] \cap \mathbb{Z}_+} \sum_{\mathbf{y} \in \mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)} \cap \mathbb{Z}_+^d(\mu)} \frac{c_2}{\ell^{d/2}}$$

$$= c_2 \sum_{\ell \in [c_1, \mu] \cap \mathbb{Z}_+} \frac{|\mathcal{Y}_\ell^{(1)} \cap \mathcal{Y}_\ell^{(2)} \cap \mathbb{Z}_+^d(\mu)|}{\ell^{d/2}}$$

$$\geq c_2 c_3^{d-1} c_4 \sum_{\ell \in [c_1, \mu] \cap \mathcal{L}|_\mu} \frac{1}{\ell^{1/2}},$$

where the last inequality follows from Lemma 16 and the fact that $\ell \in \mathcal{L}|_\mu$ if and only if $\mu \in \mathcal{L}|_\ell$. Now by Lemma 17 and the fact that $m_1 \geq 1$ (recall that $m_1 = \widetilde{\mathbb{E}}^{(1),\infty}[\tau_1] = \widetilde{\mathbb{E}}^{(2),\infty}[\tau_1]$), we have for μ large enough that

$$\sum_{\ell \in [c_1, \mu] \cap \mathcal{L}|_\mu} \frac{1}{\ell^{1/2}} \geq \sum_{\ell \in [c_1, \mu] \cap [\frac{\mu}{m_1} - c_5\sqrt{\mu}, \frac{\mu}{m_1} + c_5\sqrt{\mu}]} \frac{1}{\ell^{1/2}}$$

$$= \sum_{\ell \in [\frac{\mu}{m_1} - c_5\sqrt{\mu}, \frac{\mu}{m_1} + c_5\sqrt{\mu}]} \frac{1}{\ell^{1/2}}$$

$$\geq \frac{2c_5\sqrt{\mu} - 1}{\sqrt{\frac{\mu}{m_1} + c_5\sqrt{\mu}}},$$

which tends to $2c_5\sqrt{m_1} > 0$ as $\mu \to \infty$. We finally obtain (30) by combining the last display with (32). $\square$

4. Concluding remarks. Our main result, Theorem 1, suggests that to develop statistically consistent k-mer-based methods under a standard model such as the CFN model, one cannot simply fix a k and use the k-mer frequencies of the entire leaf sequences. Instead, one has to look for more elaborate methods, such as block decomposition.

Another possible approach to achieve consistency is to allow k to increase with the sequence length m. It is an interesting open problem to determine the *smallest* growth rate of $k = k_m$ as a function of m for which consistency becomes possible (without recourse to block techniques). By standard phylogenetic reconstruction results for distance-based and likelihood-based methods (see, e.g., [5, 36]), statistical consistency is possible in the extreme case where $k_m = m$ (i.e., when the full sequence is observed). Formally, by the reconstruction upper bound [16], Lemma 3.2, it follows that

$$(33) \qquad \limsup_{m \to \infty} \| \mathcal{L}_m^{(1)} - \mathcal{L}_m^{(2)} \|_{\mathrm{TV}} = 1 \quad \text{if } k_m = m.$$

Hence the remaining question is: Is there a sequence $\{k_m\}$ such that, say, $\lim_{m \to \infty} \frac{k_m}{m} < 1$ or even $\lim_{m \to \infty} \frac{k_m}{m} = 0$, and such that we also have $\limsup_{m \to \infty} \| \mathcal{L}_m^{(1)} - \mathcal{L}_m^{(2)} \|_{\mathrm{TV}} = 1$? A key challenge to extend our proof is that, in our use of the local CLT, we must also control the convergence rate in terms of the dimension $d = 1 + 3(2^{2k} - 2^k)$.

We focused exclusively on the two-state symmetric model of single-site substitution. We conjecture that the techniques developed here can be used to analyze more complex substitution models as well (e.g., the four-state Jukes–Cantor model). Another open question is to show that our results hold under models of insertions and deletions, such as the TKF91 model [40]. See [17] for related results regarding estimators based on the sequence length alone.

APPENDIX: INFORMATION-THEORETIC BOUNDS

In this section we give details about some basic facts we used in the paper. Recall the definition of the total variation distance in (1). It is well known (see, e.g., [26]) that the supremum on the right-hand side of (1) is reached at the set $B = \{x \in E : \nu_1(x) \geq \nu_2(x)\}$ as well as its complement B^c, and that we have the following characterizations.

LEMMA 18. *Let ν_1 and ν_2 be probability measures on a countable space E:*

$$\| \nu_1 - \nu_2 \|_{\mathrm{TV}} = \frac{1}{2} \sum_{\sigma \in E} |\nu_1(\sigma) - \nu_2(\sigma)| = 1 - \sum_{\sigma \in E} \nu_1(\sigma) \wedge \nu_2(\sigma).$$

Let X be a measurable function on a measure space $(\Omega, \mathcal{F})$, and $\mathbb{P}$ and $\mathbb{P}'$ be two probability measures on $(\Omega, \mathcal{F})$. Denote by $\mathbb{P}_{g(X)}$ and $\mathbb{P}'_{g(X)}$ the probability distribution of $g(X)$ under $\mathbb{P}$ and $\mathbb{P}'$ respectively, where g is an arbitrary measurable function on the state space of X.

LEMMA 19. *Let g be a measurable map on the state space of X. Then*

$$\| \mathbb{P}_{g(X)} - \mathbb{P}'_{g(X)} \|_{\mathrm{TV}} \leq \| \mathbb{P}_X - \mathbb{P}'_X \|_{\mathrm{TV}}.$$

PROOF. Applying the definition (1) twice,

$$\| \mathbb{P}_{g(X)} - \mathbb{P}'_{g(X)} \|_{\mathrm{TV}} = \sup_A |\mathbb{P}(g(X) \in A) - \mathbb{P}'(g(X) \in A)|$$

$$= \sup_A |\mathbb{P}(X \in g^{-1}(A)) - \mathbb{P}'(X \in g^{-1}(A))| \leq \| \mathbb{P}_X - \mathbb{P}'_X \|_{\mathrm{TV}}. \qquad \square$$

Let X, Y, Z be measurable functions on a measure space $(\Omega, \mathcal{F})$, and $\mathbb{P}$ and $\mathbb{P}'$ be two probability measures on $(\Omega, \mathcal{F})$. We say that $X \to Y \to Z$ is a *Markov chain* under $\mathbb{P}$ if Z is conditionally independent of X given Y in the sense that

$$\mathbb{P}_{Z|X,Y} = \mathbb{P}_{Z|Y}, \tag{34}$$

where $\mathbb{P}_{Z|X,Y}$ is the conditional distribution of Z given (X, Y) and $\mathbb{P}_{Z|Y}$ is the conditional distribution of Z given Y. The law of total probability and (34) imply that

$$\mathbb{P}_{X,Y,Z} = \mathbb{P}_X \mathbb{P}_{Y|X} \mathbb{P}_{Z|Y}, \tag{35}$$

where $\mathbb{P}_{X,Y,Z}$ is the joint probability distribution of (X, Y, Z).

LEMMA 20. *Suppose $\mathbb{P}_X = \mathbb{P}'_X$, $\mathbb{P}_{Y|X} = \mathbb{P}'_{Y|X}$, and $X \to Y \to Z$ is a Markov chain under both $\mathbb{P}$ and $\mathbb{P}'$. Then*

$$\big\|\mathbb{P}_{X,Y,Z} - \mathbb{P}'_{X,Y,Z}\big\|_{\mathrm{TV}} = \big\|\mathbb{P}_{Y,Z} - \mathbb{P}'_{Y,Z}\big\|_{\mathrm{TV}}.$$

PROOF. By the first equality in Lemma 18,

$$\big\|\mathbb{P}_{X,Y,Z} - \mathbb{P}'_{X,Y,Z}\big\|_{\mathrm{TV}} = \frac{1}{2} \sum_{(a,b,c)} \big|\mathbb{P}\big((X, Y, Z) = (a, b, c)\big) - \mathbb{P}'\big((X, Y, Z) = (a, b, c)\big)\big|.$$

Applying (35) to $\mathbb{P}$ and $\mathbb{P}'$, we have

$$\mathbb{P}\big((X, Y, Z) = (a, b, c)\big) = \mathbb{P}(X = a)\mathbb{P}(Y = b|X = a)\mathbb{P}(Z = c|Y = b),$$

$$\mathbb{P}'\big((X, Y, Z) = (a, b, c)\big) = \mathbb{P}'(X = a)\mathbb{P}'(Y = b|X = a)\mathbb{P}'(Z = c|Y = b).$$

From the assumptions $\mathbb{P}_X = \mathbb{P}'_X$ and $\mathbb{P}_{Y|X} = \mathbb{P}'_{Y|X}$, it follows that $\mathbb{P}_{X,Y} = \mathbb{P}'_{X,Y}$ and $\mathbb{P}_Y = \mathbb{P}'_Y$. Using the displayed equations above gives

$$\big|\mathbb{P}\big((X, Y, Z) = (a, b, c)\big) - \mathbb{P}'\big((X, Y, Z) = (a, b, c)\big)\big|$$
$$= \mathbb{P}(X = a)\mathbb{P}(Y = b|X = a)\big|\mathbb{P}(Z = c|Y = b) - \mathbb{P}'(Z = c|Y = b)\big|$$
$$= \mathbb{P}(X = a, Y = b)\big|\mathbb{P}(Z = c|Y = b) - \mathbb{P}'(Z = c|Y = b)\big|.$$

Hence

$$\big\|\mathbb{P}_{X,Y,Z} - \mathbb{P}'_{X,Y,Z}\big\|_{\mathrm{TV}}$$
$$= \frac{1}{2} \sum_{(a,b,c)} \mathbb{P}(X = a, Y = b)\big|\mathbb{P}(Z = c|Y = b) - \mathbb{P}'(Z = c|Y = b)\big|$$
$$= \frac{1}{2} \sum_{(b,c)} \mathbb{P}(Y = b)\big|\mathbb{P}(Z = c|Y = b) - \mathbb{P}'(Z = c|Y = b)\big|$$
$$= \frac{1}{2} \sum_{(b,c)} \big|\mathbb{P}(Y = b)\mathbb{P}(Z = c|Y = b) - \mathbb{P}'(Y = b)\mathbb{P}'(Z = c|Y = b)\big|,$$

where we used $\mathbb{P}(Y = b) = \mathbb{P}'(Y = b)$ in the last equality. The expression on the last line is equal to $\|\mathbb{P}_{Y,Z} - \mathbb{P}'_{Y,Z}\|_{\mathrm{TV}}$ by the first equality in Lemma 18, establishing the claim. $\square$

REFERENCES

[1] ALLMAN, E. S., RHODES, J. A. and SULLIVANT, S. (2017). Statistically consistent k-mer methods for phylogenetic tree reconstruction. *J. Comput. Biol.* **24** 153–171. MR3607847 https://doi.org/10.1089/cmb.2015.0216

[2] BAIK, J., BEN AROUS, G. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. MR2165575 https://doi.org/10.1214/009117905000000233

[3] BARBOUR, A. D. and CHRYSSAPHINOU, O. (2001). Compound Poisson approximation: A user's guide. *Ann. Appl. Probab.* **11** 964–1002. MR1865030 https://doi.org/10.1214/aoap/1015345355

[4] CAVENDER, J. A. (1978). Taxonomy with confidence. *Math. Biosci.* **40** 271–280. MR0503936 https://doi.org/10.1016/0025-5564(78)90089-5

[5] CHANG, J. T. (1996). Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. *Math. Biosci.* **137** 51–73. MR1410044 https://doi.org/10.1016/S0025-5564(96)00075-2

[6] COMPEAU, P. and PEVZNER, P. (2018). *Bioinformatics Algorithms*: *An Active Learning Approach*. Active Learning Publishers, CA.

[7] CULANOVSKII, I. V. (1961). *Twenty-Five Papers on Statistics and Probability*. *Selected Translations in Mathematical Statistics and Probability Series*. Am. Math. Soc., Providence.

[8] DASKALAKIS, C. and ROCH, S. (2013). Alignment-free phylogenetic reconstruction: Sample complexity via a branching process analysis. *Ann. Appl. Probab.* **23** 693–721. MR3059273 https://doi.org/10.1214/12-AAP852

[9] DAVIS, B. and MCDONALD, D. (1995). An elementary proof of the local central limit theorem. *J. Theoret. Probab.* **8** 693–701. MR1340834 https://doi.org/10.1007/BF02218051

[10] DURBIN, R., EDDY, S. R., KROGH, A. and MITCHISON, G. (1998). *Biological Sequence Analysis*: *Probabilistic Models of Proteins and Nucleic Acids*. Cambridge Univ. Press, Cambridge.

[11] DURDEN, C. and SULLIVANT, S. (2019). Identifiability of phylogenetic parameters from k-mer data under the coalescent. *Bull. Math. Biol.* **81** 431–451. MR3902908 https://doi.org/10.1007/s11538-018-0399-1

[12] DURRETT, R. (2019). *Probability—Theory and Examples*, 5th ed. *Cambridge Series in Statistical and Probabilistic Mathematics* **49**. Cambridge Univ. Press, Cambridge. MR3930614 https://doi.org/10.1017/9781108591034

[13] ERDŐS, P. L., STEEL, M. A., SZÉKELY, L. A. and WARNOW, T. J. (1999). A few logs suffice to build (almost) all trees. I. *Random Structures Algorithms* **14** 153–184. MR1667319 https://doi.org/10.1002/(SICI)1098-2418(199903)14:2<153::AID-RSA3>3.3.CO;2-I

[14] EVANS, W., KENYON, C., PERES, Y. and SCHULMAN, L. J. (2000). Broadcasting on trees and the Ising model. *Ann. Appl. Probab.* **10** 410–433. MR1768240 https://doi.org/10.1214/aoap/1019487349

[15] FAN, H., IVES, A. R., SURGET-GROBA, Y. and CANNON, C. H. (2015). An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics* **16** 522.

[16] FAN, W.-T. and ROCH, S. (2018). Necessary and sufficient conditions for consistent root reconstruction in Markov models on trees. *Electron. J. Probab.* **23** Paper No. 47, 24. MR3814241 https://doi.org/10.1214/18-ejp165

[17] FAN, W.-T. L., LEGRIED, B. and ROCH, S. (2020). Impossibility of consistent distance estimation from sequence lengths under the TKF91 model. *Bull. Math. Biol.* **82** Paper No. 123, 12. MR4150869 https://doi.org/10.1007/s11538-020-00801-3

[18] FARRIS, J. S. (1973). A probability model for inferring evolutionary trees. *Syst. Biol.* **22** 250–256.

[19] GANESH, A. and ZHANG, Q. (2019). Optimal sequence length requirements for phylogenetic tree reconstruction with indels. In *STOC'19—Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* 721–732. ACM, New York. MR4003378 https://doi.org/10.1145/3313276.3316345

[20] GUSFIELD, D. (1997). *Algorithms on Strings*, *Trees*, *and Sequences*: *Computer Science and Computational Biology*. Cambridge Univ. Press, Cambridge. MR1460730 https://doi.org/10.1017/CBO9780511574931

[21] HAUBOLD, B. (2014). Alignment-free phylogenetics and population genetics. *Brief. Bioinform.* **15** 407–418.

[22] HAUBOLD, B., KLÖTZL, F. and PFAFFELHUBER ANDI, P. (2015). Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics* **31** 1169–1175.

[23] KARLIN, S. and ALTSCHUL, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* **87** 2264–2268.

[24] LEES, J. A., HARRIS, S. R., TONKIN-HILL, G., GLADSTONE, R. A., LO, S. W., WEISER, J. N., CORANDER, J., BENTLEY, S. D. and CROUCHER, N. J. (2019). Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* **29** 304–316. https://doi.org/10.1101/gr.241455.118

[25] LEES, J. A., KENDALL, M., PARKHILL, J., COLIJN, C., BENTLEY, S. D. and HARRIS, S. R. (2018). Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: A simulation based study. *Wellcome Open Res.* **3** 33. https://doi.org/10.12688/wellcomeopenres.14265.2

[26] LEVIN, D. A., PERES, Y. and WILMER, E. L. (2006). *Markov Chains and Mixing Times*. Amer. Math. Soc., Providence, RI.

[27] LIPPERT, R. A., HUANG, H. and WATERMAN, M. S. (2002). Distributional regimes for the number of k-word matches between two random sequences. *Proc. Natl. Acad. Sci. USA* **99** 13980–13989. MR1944413 https://doi.org/10.1073/pnas.202468099

[28] MOSSEL, E. (2004). Phase transitions in phylogeny. *Trans. Amer. Math. Soc.* **356** 2379–2404. MR2048522 https://doi.org/10.1090/S0002-9947-03-03382-8

[29] NEYMAN, J. (1971). Molecular studies of evolution: A source of novel statistical problems. In *Statistical Decision Theory and Related Topics* (*Proc. Sympos.*, *Purdue Univ.*, *Lafayette*, *Ind.*, 1970) 1–27. MR0327321

[30] ONDOV, B. D., TREANGEN, T. J., MELSTED, P., MALLONEE, A. B., BERGMAN, N. H., KOREN, S. and PHILLIPPY, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17** 132. https://doi.org/10.1186/s13059-016-0997-x

[31] PATTERSON, N., PRICE, A. L. and REICH, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* **2** e190.

[32] PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909.

[33] QI, J., WANG, B. and HAO, B.-I. (2004). Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. *Mol. Biol. Evol.* **58** 1–11. https://doi.org/10.1007/s00239-003-2493-7

[34] REINERT, G., CHEW, D., SUN, F. and WATERMAN, M. S. (2009). Alignment-free sequence comparison. I. Statistics and power. *J. Comput. Biol.* **16** 1615–1634. MR2578699 https://doi.org/10.1089/cmb.2009.0198

[35] ROCH, S. and SLY, A. (2017). Phase transition in the sample complexity of likelihood-based phylogeny inference. *Probab. Theory Related Fields* **169** 3–62. MR3704765 https://doi.org/10.1007/s00440-017-0793-x

[36] ROYCHOUDHURY, A., WILLIS, A. and BUNGE, J. (2015). Consistency of a phylogenetic tree maximum likelihood estimator. *J. Statist. Plann. Inference* **161** 73–80. MR3316551 https://doi.org/10.1016/j.jspi.2015.01.001

[37] STEEL, M. (1994). Recovering a tree from the leaf colourations it generates under a Markov model. *Appl. Math. Lett.* **7** 19–23. MR1350139 https://doi.org/10.1016/0893-9659(94)90024-8

[38] STEEL, M. (2016). *Phylogeny—Discrete and Random Processes in Evolution*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **89**. SIAM, Philadelphia, PA. MR3601108 https://doi.org/10.1137/1.9781611974485.ch1

[39] TAVARÉ, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.* **26** 119–164. MR0770050 https://doi.org/10.1016/0040-5809(84)90027-3

[40] THORNE, J. L., KISHINO, H. and FELSENSTEIN, J. (1991). An evolutionary model for maximum likelihood alignment of dna sequences. *J. Mol. Evol.* **33** 114–124.

[41] ULITSKY, I., BURSTEIN, D., TULLER, T. and CHOR, B. (2006). The average common substring approach to phylogenomic reconstruction. *J. Comput. Biol.* **13** 336–350. MR2255263 https://doi.org/10.1089/cmb.2006.13.336

[42] VINGA, S. and ALMEIDA, J. (2003). Alignment-free sequence comparison—a review. *Bioinformatics* **19** 513–523.

[43] WAN, L., REINERT, G., SUN, F. and WATERMAN, M. S. (2010). Alignment-free sequence comparison (II): Theoretical power of comparison statistics. *J. Comput. Biol.* **17** 1467–1490. MR2739743 https://doi.org/10.1089/cmb.2010.0056

[44] WARNOW, T. (2017). *Computational Phylogenetics*: *An Introduction to Designing Methods for Phylogeny Estimation*, 1st ed. Cambridge Univ. Press, USA.