Bioinformatics Advance Access published June 30, 2015

Inference of Markovian Properties of Molecular Sequences from NGS Data and Applications to Comparative Genomics

Jie Ren [1], Kai Song [2], Minghua Deng [2], Gesine Reinert [3], Charles H. Cannon [4,5] and Fengzhu Sun [1,6] *

[1]Molecular and Computational Biology Program, University of Southern California, Los Angeles, California, USA;
[2]School of Mathematical Sciences, Peking University, Beijing, China;
[3]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK;
[4]Department of Biological Sciences, Texas Tech University, TX 79409-3131, USA;
[5]Xishuangbanna Tropical Botanic Garden, Chinese Academy of Sciences, Yunnan, China;
[6]Centre for Computational Systems Biology, School of Mathematical Sciences, Fudan University, Shanghai, China.

Associate Editor: Prof. Benjamin Raphael

ABSTRACT

Motivation: Next Generation Sequencing (NGS) technologies generate large amounts of short read data for many different organisms. The fact that NGS reads are generally short makes it challenging to assemble the reads and reconstruct the original genome sequence. For clustering genomes using such NGS data, word-count based alignment-free sequence comparison is a promising approach, but for this approach, the underlying expected word counts are essential.

A plausible model for this underlying distribution of word counts is given through modelling the DNA sequence as a Markov chain (MC). For single long sequences, efficient statistics are available to estimate the order of MCs and the transition probability matrix for the sequences. As NGS data do not provide a single long sequence, inference methods on Markovian properties of sequences based on single long sequences cannot be directly used for NGS short read data.

Results: Here we derive a normal approximation for such word counts. We also show that the traditional Chi-square statistic has an approximate gamma distribution, using the Lander-Waterman model for physical mapping. We propose several methods to estimate the order of the MC based on NGS reads and evaluate them using simulations.

We illustrate the applications of our results by clustering genomic sequences of several vertebrate and tree species based on NGS reads using alignment-free sequence dissimilarity measures. We find that the estimated order of the MC has a considerable effect on the clustering results, and that the clustering results that use a MC of the estimated order give a plausible clustering of the species.

Availability: Our implementation of the statistics developed here is available as R package "NGS.MC" at `http://www-rcf.usc.edu/~fsun/Programs/NGS-MC/NGS-MC.html`.

Contact: fsun@usc.edu

*to whom correspondence should be addressed

1 INTRODUCTION

NGS technologies generate large amounts of overlapping short read data for many different organisms; for example a read is a subsequence of less than 400 bps for Illumina and 700 bps for 454 sequencing technologies, and can sometimes be much shorter. The fact that NGS reads are generally short makes it challenging to reconstruct the original genome sequence.

Recently several word-count based alignment-free sequence comparison methods have been applied to infer the relationship among different species (Yi and Jin, 2013; Song *et al.*, 2013) and metagenomic samples (Jiang *et al.*, 2012; Wang *et al.*, 2014; Behnam and Smith, 2014; Hurwitz *et al.*, 2014) based on NGS reads without assembly. Our alignment-free sequence dissimilarity measures, d_2^* and d_2^S (Song *et al.*, 2013, 2014), and their variants (Liu *et al.*, 2011; Behnam *et al.*, 2013; Ren *et al.*, 2013) have shown promise. These methods require the knowledge about the approximate distribution of word counts in the underlying sequences. While a model which assumes that all letters in the sequence are equally likely is relatively straightforward to analyse, see Reinert *et al.* (2009), a Markov model for the underlying sequences is more realistic.

Markov chains (MC) have been widely used to model molecular sequences (Almagor, 1983) with many applications including the study of dependencies between the bases (Blaisdell, 1985), the enrichment and depletion of certain word patterns (Pevzner *et al.*, 1989), prediction of occurrences of long word patterns from short patterns (Hong, 1990; Arnold *et al.*, 1988), and the detection of

signals in introns (Avery, 1987). Narlikar *et al.* (2013) studied the effect of the order of MCs on several biological problems including phylogenetic analysis, assignment of sequence fragments to different genomes in metagnomic studies, motif discovery, and functional classification of promoters. These applications showed the importance of accurate specification of the order of MCs. Reliable estimators for the order of the MC and the transition probability matrix based on the sequence data are crucial.

Based on relatively long molecular sequences, for a general finite state MC sequence of letters from a finite alphabet $\mathcal{A} = \{1, 2, \cdots, L\}$ of size L, Hoel (1954) showed, under the hypothesis that the long sequence follows a $(k-2)$-th order MC, that twice the log-likelihood ratio of the probability of the sequence under a $(k-1)$-th order MC versus that under the $(k-2)$-th order MC model follows approximately a χ^2-distribution with $df_k = (L-1)^2 L^{k-2}$ degrees of freedom under general conditions. He also approximated the log-likelihood ratio by the Pearson-type statistic

$$S_k = \sum_{\mathbf{w} \in \mathcal{A}^k} \frac{(N_{\mathbf{w}} - E_{\mathbf{w}})^2}{E_{\mathbf{w}}}, \tag{1}$$

which is also approximately χ^2-distributed with the same degrees of freedom. Here, $\mathbf{w} = w_1 w_2 \cdots w_k$ denotes a k-word formed of letters $w_i \in \mathcal{A}$, ${}^-\mathbf{w} = w_2 \cdots w_k$, $\mathbf{w}^- = w_1 w_2 \cdots w_{k-1}$, and ${}^-\mathbf{w}^- = w_2 \cdots w_{k-1}$; $N_{\mathbf{w}}$ denotes the count of the word $\mathbf{w}$ in the sequence, and $E_{\mathbf{w}} = \frac{N_{-\mathbf{w}} N_{\mathbf{w}^-}}{N_{-\mathbf{w}^-}}$ is the estimated expected count of $\mathbf{w}$ if the sequence is generated by a MC of order $k-2$. Here $k \geq 3$; see also Avery and Henderson (1999) for a detailed study, Billingsley (1961a,b) for an an excellent exposition of statistical issues related to MCs, as well as Waterman (1995); Reinert *et al.* (2000, 2005); Ewens and Grant (2005) for applications to sequence analysis.

The Chi-square statistic (1) and the log-likelihood ratio statistics can be used to test the order of a MC, using all k-words $\mathbf{w} \in \mathcal{A}^k$. When a particular order of MC is rejected, we can identify particular word patterns that are exceptional, through the approximate distribution of $N_{\mathbf{w}}$. The approximate distributions of $N_{\mathbf{w}}$ in long sequences is well understood, see for example Waterman (1995); Reinert *et al.* (2005, 2000). In particular, suppose that the sequence follows a stationary $(k-2)$-th order MC and let

$$\hat{\sigma}^2_{\mathbf{w}} = E_{\mathbf{w}} \left(1 - \frac{N_{-\mathbf{w}}}{N_{-\mathbf{w}^-}}\right) \left(1 - \frac{N_{\mathbf{w}^-}}{N_{-\mathbf{w}^-}}\right).$$

For

$$Z_{\mathbf{w}} = \frac{N_{\mathbf{w}} - E_{\mathbf{w}}}{\hat{\sigma}_{\mathbf{w}}}, \tag{2}$$

Theorem 6.4.2 in Reinert *et al.* (2005) gives that, as sequence length goes to infinity, for all real values x, $\mathbb{P}(Z_{\mathbf{w}} \leq x) \to \Phi(x)$, where Φ denotes the cumulative distribution function of a standard normal variable. We also say that $Z_{\mathbf{w}}$ converges to the standard normal distribution $N(0,1)$ *in distribution*. This asymptotic result can then be used to find exceptional words in long sequences.

Given an NGS short read sample, it is tempting to use the test statistic S_k defined in (1) to test the order of a MC by simply counting the number of the occurrences of words in short read data. However, as the short reads from NGS data are sampled randomly from the genome, some parts of the genome are possibly not sampled and some parts are possibly sampled extensively. The sampling process introduces additional randomness to the statistic, and makes S_k deviate from its traditional χ^2 -distribution. Similarly, the approximate distribution of $Z_{\mathbf{w}}$ given in (2) will be different from the standard normal distribution.

In this paper, we study these approximate distributions, both theoretically and by simulations. First we extend the statistics S_k and $Z_{\mathbf{w}}$ for a MC sequence to S_k^R and $Z_{\mathbf{w}}^R$ for the NGS read data. Our underlying model for the distribution of reads along the genome is the potentially inhomogeneous Lander-Waterman model for physical mapping (Lander and Waterman, 1988). We discover that for a set of short reads sampled from a $(k-2)$-th order MC sequence, the statistic S_k^R follows approximately a gamma distribution with shape parameter $df_k/2$ and scale parameter $2d$, where d is a factor related to the distribution of the reads along the genome. We also show that, with the same factor d, the distribution of the single word statistic $Z_{\mathbf{w}}^R/\sqrt{d}$ tends to the standard normal distribution. Based on the theoretical results, we introduce an estimator for the order of the MC using NGS data. For practical purposes, we also give an estimator for the factor d when the underlying reads sampling distribution is unknown. To the best of our knowledge, this is the first study of the Markovian properties of molecular sequences based on NGS read data.

To illustrate our theoretical results and our estimators, we first carry out a simulation study based on transition probability matrices which are estimated from cis-regulatory module (CRM) DNA sequences, and insert repeats. We simulate different read lengths, numbers of reads, inhomogeneous sampling, as well as sequencing errors, and we include a regime where the sampling rate depends on the GC content. If the GC bias is not very strong or the sequencing depth is not very low, then the simulation results agree with our theoretical predictions despite the theoretical assumptions being slightly violated.

Next we apply our methods to cluster 28 vertebrate species using our alignment-free dissimilarity measures d_2^* and d_2^S under different MC models which are estimated from NGS read samples. The estimated orders based on NGS data without assembly are found to be consistent with those inferred directly from the long genome sequences. The clustering performs best when using MCs around the estimated order. Applying the same analysis to 13 tropical tree species whose genomes are unknown, based on their NGS read samples, the most plausible clustering is achieved when using a MC model of order close to the one estimated from the NGS reads.

The paper is organized as follows. The "Methods" section contains the probabilistic models of generating the MC sequence and sampling the short reads, as well as the theorem for the approximate distributions of S_k^R and $Z_{\mathbf{w}}^R$ for NGS data. This theorem is used to derive our estimators for the order of the MC and for the factor d. In the "Results" section, we first provide extensive simulation studies including the comparison of the theoretical approximate distributions and the simulated results for S_k^R and $Z_{\mathbf{w}}^R$, the effect of inhomogeneous sampling and sequencing errors, the efficiency of the estimator of the factor d, and the evaluations of the methods for estimating the MC order. Second, we estimate the orders of the MCs for 28 vertebrate species based on the simulated whole genome NGS samples. We then use our dissimilarity measures d_2^* and d_2^S to cluster the NGS samples of the 28 species under different MC orders to see the effect on the performance of the clustering. The applications show that our new methods are effective for the inference of relationships among sequences based on NGS

reads. Finally, we use our methods to study the relationships among 13 tree species whose complete genomic sequences as well as their phylogenetic relationships are unknown. Our clustering results are consistent with the physical characteristics of the tree species. The paper concludes with some discussion of the study.

2 METHODS

2.1 Probabilistic modeling of a MC sequence and random sampling of the reads using NGS

In NGS, a large number of reads are randomly sampled from the genome. Hence two random processes are involved in the generation of the short read data: the generation of the underlying genome sequence and the random sampling of the reads.

We use an r-th order homogeneous ergodic MC to model the underlying genome sequence with each letter taking values in a finite alphabet set $\mathcal{A}$ of size L. Since our study is based on genomic sequences, $L = 4$. As in Lander and Waterman (1988); Zhang *et al.* (2008); Zhai *et al.* (2012); Daley and Smith (2013); Simpson (2014), we assume that the genome is continuous and that the distribution of reads along the genome follows a potentially *inhomogeneous* Poisson process with rate $c(x)$ at position x. If $c(x) = c$ for all x, we refer to the sampling of the reads as *homogeneous*. We assume that all sampled reads have the same length of β bps. A total of M reads are independently sampled from the genome of length G bps.

We extend the statistics S_k and $Z_{\mathbf{w}}$ in (1) and (2) to NGS short read data accordingly. Let $N^R_{\mathbf{w}}$ be the number of occurrences of the k-word $\mathbf{w}$ in the short read data, where the superscript R refers to the "read" data, and define

$$S^R_k = \sum_{\mathbf{w}\in\mathcal{A}^k} \frac{\left(N^R_{\mathbf{w}} - E^R_{\mathbf{w}}\right)^2}{E^R_{\mathbf{w}}}, \tag{3}$$

$$Z^R_{\mathbf{w}} = \frac{N^R_{\mathbf{w}} - E^R_{\mathbf{w}}}{\hat{\sigma}^R_{\mathbf{w}}}, \tag{4}$$

where

$$E^R_{\mathbf{w}} = \frac{N^R_{-\mathbf{w}} N^R_{\mathbf{w}-}}{N^R_{-\mathbf{w}-}} \text{ and } (\hat{\sigma}^R_{\mathbf{w}})^2 = E^R_{\mathbf{w}} \left(1 - \frac{N^R_{-\mathbf{w}}}{N^R_{-\mathbf{w}-}}\right)\left(1 - \frac{N^R_{\mathbf{w}-}}{N^R_{-\mathbf{w}-}}\right).$$

We have the following theorem on the approximate distributions of S^R_k and $Z^R_{\mathbf{w}}$; the proof is given in the Supplementary Materials. Note that for each read we discard the last $k-1$ positions as they would lead to words of length less than k; the error made with this approximation is asymptotically negligible when k is small relative to β.

THEOREM 1. *Assume that the underlying genome follows a $(k-2)$-th order MC which assigns non-zero probability to every k-word $\mathbf{w}$. Let S^R_k and $Z^R_{\mathbf{w}}$ be defined as in (3) and (4), respectively. Suppose that the genome of length G can be divided into (not necessarily contiguous) regions with constant coverage r_i for the i-th region, so that every base is covered exactly r_i times, based on the first $\beta - k + 1$ positions of the reads. Let G_i be the length of the i-th region that changes with G in a way such that $\lim_{G\to\infty} G_i/G = f_i > 0$ for the i-th region, $i = 1, 2, \cdots$. Let*

$$d = \frac{\sum_i r_i^2 f_i}{\sum_i r_i f_i}. \tag{5}$$

Then, as $G \to \infty$,

a) For each k-word $\mathbf{w}$, in distribution, $Z^R_{\mathbf{w}}/\sqrt{d} \to N(0,1)$.

b) The statistic S^R_k/d has an approximate χ^2-distribution with $df_k = (L-1)^2 L^{k-2}$ degrees of freedom; equivalently, the statistic S^R_k has an approximate gamma distribution with shape parameter $df_k/2$ and scale parameter $2d$.

If the M reads are sampled homogeneously along the genome with coverage c based on the first $\beta - k + 1$ positions of the reads along the genome, i.e. $c = \frac{M(\beta-k+1)}{G-k+1}$, the Lander-Waterman formula (Lander and Waterman, 1988) shows that the fraction of genome covered $r_i = i$ times is $f_i = \exp(-c)c^i/i!$. Under this assumption, we obtain

$$d = \frac{\sum_i i^2 f_i}{\sum_i i f_i} = \frac{c^2 + c}{c} = c + 1.$$

The results in Theorem 1 continue to hold when taking $d = c + 1$.

In the Lander-Waterman model for physical mapping (Lander and Waterman, 1988), the factor $c = \frac{M\beta}{G}$ is the coverage of the genome. Hence we refer to d from (5) as the *effective coverage* of the reads along the genome based on the first $\beta - k + 1$ positions of each read.

2.2 Estimating the order of the MC based on NGS reads

Based on Theorem 1, we can estimate the order r of a MC sequence using NGS reads. First, we test the null hypothesis that the sequence follows an independent identically distributed (i.i.d; MC order = 0) model. For a test at significance level α, if S^R_2/d is higher than the $1-\alpha$ quantile of the χ^2-distribution with $df = (L-1)^2$ degrees of freedom, the i.i.d hypothesis is rejected. If this null hypothesis is rejected, then here we propose an estimator for the order of a MC; it is an analog of a corresponding established estimator of MC orders based on long sequences that has been shown to be effective. In the Supplementary Materials we present four related estimators as well as estimators based on the AIC and BIC information criteria; the one presented here has the best performance in simulation studies.

We assume that the word length $k \geq 2$ and that the assumptions of Theorem 1 are satisfied. Then, for $k \geq r + 2$, S^R_k/d has approximately a χ^2-distribution with $(L-1)^2 L^{k-2}$ degrees of freedom. If $k < r + 2$, then S^R_k/d will typically be larger than expected from this χ^2-distribution. For $k \geq r + 2$, the law of large numbers gives that $\frac{S^R_{k+1}}{LS^R_k} \to 1$ for $G \to \infty$; if $k < r + 2$ then the ratio will be much larger than 1 in the limit. Therefore we can estimate r as follows:

$$\hat{r}_{S_k} = \text{argmin}_k \left\{\frac{S^R_{k+1}}{S^R_k}\right\} - 1. \tag{6}$$

In general, we want the value of $\min_k \left\{\frac{S^R_{k+1}}{S^R_k}\right\}$ to be very small, e.g, less than 0.01.

Using the law of large numbers it can be shown that under our assumptions this estimator is consistent, in the sense that $\hat{r}_{S_k}$ tends to r in probability as G tends to infinity.

2.3 Estimating the effective coverage d

Often the effective coverage d is not known and we would like to estimate the effective coverage d using NGS short read data. From Theorem 1, we can see that, under the general conditions stated in the theorem, $(Z^R_{\mathbf{w}})^2/d$ follows a χ^2-distribution with one degree of freedom. Since the median of the χ^2-distribution with one degree of freedom is about 0.456, we can use the scaled median as a robust estimator for d;

$$\hat{d} = \text{median}\{(Z^R_{\mathbf{w}})^2, \ \ \mathbf{w} \in \mathcal{A}^k\}/0.456. \tag{7}$$

When we assume that the underlying long sequence follows a MC of order at most m, we use $(m+2)$-words to estimate d using (7).

Note that for an i.i.d. model sequence, the set of 2-words would not yield meaningful results as there are only 16 different 2-words and the median based on 16 numbers is generally not reliable. As an underlying genome sequence following an r-th order MC can also be seen as an $(r+1)$-th, $(r+2)$-th, ..., and higher order MC sequence, we can use k-words with relatively large k ($\geq r+2$) to estimate the factor d, if the maximum order of a MC is unknown beforehand.

2.4 Simulation study

For the simulation study, we first generate MCs of different orders. For realistic parameter values, the transition probability matrices of the MCs are based on real cis-regulatory module (CRM) DNA sequences in mouse forebrain from Blow *et al.* (2010). We use CRM sequences here because CRM sequences are often used to study the effectiveness of alignment-free sequence dissimilarity measures (Göke *et al.*, 2012; Song *et al.*, 2014; Ren *et al.*, 2013). To take into consideration that in real genomic sequences, many repeat regions are present, we insert repeats into the generated MCs. We simulate NGS data by sampling a varying number of reads of different lengths from the MC, varying genome length as well as coverage.

We include homogeneous and inhomogeneous sampling of the reads as well as sequencing errors. We also let the sampling rate of the reads depend on the GC content of the fragments based on data from the current sequencing technologies (Benjamini and Speed, 2012). We set the sequencing error rate at 10%, which is relatively high compared to the true sequencing error rate in real sequencing in order to clearly distinguish among the estimators with regards to their robustness to sequencing errors. When a sequencing error occurs at a position, the nucleotide base is changed to one of the other three nucleotides with equal probability.

Once the NGS reads are generated, we calculate the statistics S^R_k and $Z^R_{\mathbf{w}}$ for each word $\mathbf{w}$, the order estimator $\hat{r}_{S_k}$ and the estimator for effective coverage d based on (7); each procedure is repeated 1000 times. In each repeat experiment, we let the order estimator choose the model from 1st, 2nd, $\cdots$, 5th order MCs; we estimate the effective coverage d by (7), using 3-tuples for a first order MC, and 4-tuples for a second order MC. The details are given in the Supplementary Materials.

2.5 Applications to the study of relationships among organisms

We test our methods on real and simulated NGS data from 28 vertebrate species whose complete genomic sequences are available and that are comprehensively studied in (Miller *et al.*, 2007; Karolchik *et al.*, 2008). We download the genomes of the 28 vertebrate species from UCSC Genome Browser, and then use MetaSim (Richter *et al.*, 2008) to simulate reads from each of the 28 vertebrate species. In simulations the accuracy of the order estimation increases with read coverage. To reflect a worst-case scenario, we set the read coverage to be 1 as a lower bound for the performance although the current sequencing technology can generate data with very high read coverage. We set MetaSim to generate reads of length 62bp under the error rate which is estimated by Illumina in our simulations.

To estimate the order of MC based on the NGS sample for each of the 28 species, we apply the order estimator $\hat{r}_{S_k}$ in (6); there is no sharp ratio transition found over $k = 2, \cdots, 14$. Given that real genomes consist of multiple types of regions (coding, non-coding and regulatory regions) and each type may fit to different MC models, the result indicates that no suitable MC model can adequately fit all the patterns in the genome. Instead, we fit the data with a MC model that can explain the majority (say 80%) of the word patterns in the genome. Motivated by the normal approximation of a particular word statistic in Theorem 1, we study the fraction of k-words whose occurrences can be explained using the statistic $(Z^R_{\mathbf{w}})^2/d$ by comparison to a χ^2-distribution with one degree of freedom with type I error 0.01. We estimate the order of MC to be the smallest $k-2$ under which more than 80% of k-words can be explained by the $(k-2)$-th order MC.

To cluster the organisms, we use the inferred MC models to estimate the expected number of occurrences of word patterns and then study the relationships among the organisms using our dissimilarity measures d_2^* and d_2^S. We briefly present their definitions below, please see Song *et al.* (2013, 2014) for details. Then we apply a similar approach to study the relationships among 13 tree species with NGS reads, for which neither the complete genome sequences nor their relationships are known. To estimate the unknown effective coverage d using k-words by (7), we let k to be relatively large and use $\hat{d}$ as the value at which the estimated d stabilizes as k increases.

2.6 Alignment-free sequence comparison dissimilarity measures

Consider two sets of NGS reads from two genomes. We use superscripts (1) and (2) to denote the first and the second read set, respectively. Suppose that $M^{(i)}$ reads of length $\beta^{(i)}$ are in the i-th data set. Since the reads can come from either the forward strand or the reverse strand of the genome in NGS, we supplement the observed reads by their complements and refer to the joint set of the reads and the complements as the read set.

Let $N^{(i)}_{\mathbf{w}}$ be the count of the word $\mathbf{w}$ in the i-th data set. We define $EN^{(i)}_{\mathbf{w}}$ to be the expected number of occurrences of word $\mathbf{w}$ based on either the i.i.d model or a Markov model, $EN^{(i)}_{\mathbf{w}} = M^{(i)}(\beta^{(i)}-k+1)(p^{(i)}_{\mathbf{w}} + p^{(i)}_{\bar{\mathbf{w}}})$, where $M^{(i)}(\beta^{(i)}-k+1)$ is the total number of k-word in the i-th sample, $\bar{\mathbf{w}}$ is the complement of word $\mathbf{w}$, and $p^{(i)}_{\mathbf{w}}$ is the probability of word $\mathbf{w}$ in the i-th genome

under a specific model. Then we define D_2^* and D_2^S as follows,

$$D_2^* = \sum_{\mathbf{w}\in\mathcal{A}^k} \frac{\tilde{N}_{\mathbf{w}}^{(1)}\tilde{N}_{\mathbf{w}}^{(2)}}{\sqrt{EN_{\mathbf{w}}^{(1)}EN_{\mathbf{w}}^{(2)}}}, \text{and } D_2^S = \sum_{\mathbf{w}\in\mathcal{A}^k} \frac{\tilde{N}_{\mathbf{w}}^{(1)}\tilde{N}_{\mathbf{w}}^{(2)}}{\sqrt{\left(\tilde{N}_{\mathbf{w}}^{(1)}\right)^2+\left(\tilde{N}_{\mathbf{w}}^{(2)}\right)^2}},$$

where $\tilde{N}_{\mathbf{w}}^{(i)} = N_{\mathbf{w}}^{(i)} - EN_{\mathbf{w}}^{(i)}$, $i = 1, 2$. Further, the dissimilarity measures d_2^* and d_2^S, ranging from 0 to 1, are defined as,

$$d_2^* = \frac{1}{2}\left(1 - \frac{D_2^*}{\sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k}\frac{\left(\tilde{N}_{\mathbf{w}}^{(1)}\right)^2}{EN_{\mathbf{w}}^{(1)}}}\sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k}\frac{\left(\tilde{N}_{\mathbf{w}}^{(2)}\right)^2}{EN_{\mathbf{w}}^{(2)}}}}\right), \text{and}$$

$$d_2^S = \frac{1}{2}\left(1 - \frac{D_2^S}{\sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k}\frac{\left(\tilde{N}_{\mathbf{w}}^{(1)}\right)^2}{\sqrt{\left(\tilde{N}_{\mathbf{w}}^{(1)}\right)^2+\left(\tilde{N}_{\mathbf{w}}^{(2)}\right)^2}}}\sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k}\frac{\left(\tilde{N}_{\mathbf{w}}^{(2)}\right)^2}{\sqrt{\left(\tilde{N}_{\mathbf{w}}^{(1)}\right)^2+\left(\tilde{N}_{\mathbf{w}}^{(2)}\right)^2}}}}\right).$$

For comparison, we also use a simplistic dissimilarity measure based on the non-centered correlation of the word frequencies defined as $d_2 = \frac{1}{2}\left(1 - \frac{\sum_{\mathbf{w}\in\mathcal{A}^k} N_{\mathbf{w}}^{(1)}N_{\mathbf{w}}^{(2)}}{\sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k}\left(N_{\mathbf{w}}^{(1)}\right)^2}\sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k}\left(N_{\mathbf{w}}^{(2)}\right)^2}}\right)$.

3 RESULTS

3.1 Summary of simulation results

Due to page limitations, we summarize the simulation results here; details are given in the Supplementary Materials. Our extensive simulations show that the simulated mean, standard deviation and distributions of S_k^R and $Z_{\mathbf{w}}^R$ are very close to their corresponding theoretical approximations given by Theorem 1. Both the effective coverage and the MC order can be estimated accurately under the parameter settings of the current sequencing technologies.

3.2 The relationship among 28 vertebrate species

Table S4 shows the estimated orders of MCs for a group of 28 vertebrate species that are studied in (Miller *et al.*, 2007; Karolchik *et al.*, 2008) based on simulated NGS short reads. For each of the 28 species, we compute the fraction of the k-words that have $(Z_{\mathbf{w}}^R)^2/\hat{d}$ within the 99% of a χ^2-distribution with one degree of freedom, for $k = 8, 9, \ldots, 14$. Using 80% as a threshold, we estimate the order of MC for each species to be the smallest $k-2$ under which the fraction of words that can be explained by the $(k-2)$-th order MC is greater than the threshold.

Comparing our results with the results in Narlikar *et al.* (2013), where the order of MCs for a selection of vertebrate genomes was estimated by AIC and BIC criteria using whole genome sequences, we find that the estimated order based on NGS read data are almost the same as that estimated based on the whole genome sequences in Narlikar *et al.* (2013). Our proposed methods of estimating the order of MC based on short reads of NGS data achieve the same accuracy as that based on whole genome sequences.

For a given value of k, we compute d_2^* and d_2^S using an r-th order MC, $r = 0$ (i.i.d model), $\ldots, (k-2)$ for each pair of species, yielding a 28×28 pairwise dissimilarity matrix under each MC model. To evaluate the dissimilarity measures, we use the pairwise distance matrix obtained from Figure S1 in Miller *et al.* (2007) as the gold standard for the dissimilarity between each pair of the 28 species; the matrix is given as Table S5 in the Supplementary Materials. Note that the estimated orders of the 28 species range from 7 to 11, and the average order is 10. To study the performance of the dissimilarity measures under different orders of MC, we choose $k = 14$ such that we can study the results under the MC model with orders up to 12.

Table 1 shows Spearman's rank correlation coefficient (SPCC) between the standard distance and the dissimilarity estimated by the d_2-type measures under MC models of various orders; higher SPCC indicates better performance. Both measures, d_2^* and d_2^S, achieve their best results of SPCC=0.92 when using a MC of order 12. Note that using a simplistic dissimilarity measure d_2 only gives SPCC=0.08.

In general both d_2^* and d_2^S obtain higher SPCC with the standard matrix as the order of MC increases, except for d_2^S at order 9. In particular, the measure d_2^* has negative correlation coefficient with the standard distance under the i.i.d model. The SPCC becomes stable when the order of the MC used for the analysis is close to 11, the maximum estimated MC orders over the 28 species. Here d_2^S is less affected by the order of the MC than d_2^*. When the appropriate order of MC is used, d_2^* and d_2^S perform similarly and much better than d_2.

d_2-type	order=0	order=5	order=9	order=10	order=11	order=12
d_2^*	-0.21	-0.16	0.85	0.89	0.90	0.92
d_2^S	0.86	0.87	0.85	0.88	0.90	0.92

Table 1. The Spearman's rank correlation coefficient (SPCC) between the true distance matrix and the dissimilarity matrix by d_2-type dissimilarity measures under MC models with order 0 (i.i.d), 5, 9, 10, 11 and 12. The length of the k-tuple word is 14.

3.3 The relationship among 13 tropical tree species with unknown reference genomes

We also apply our method to the 13 tree species based on the NGS shotgun read data sets in Cannon *et al.* (2010). The reference genome sequences for the 13 tree species are unknown. Our objective is to cluster these tree species using d_2^* and d_2^S with MCs for the sequences.

The estimated order of the MC for all the 13 tree species is 8. We use the dissimilarity measures d_2^* and d_2^S under various orders of MC as the background model to cluster the 13 tree species from their NGS reads. We choose $k = 11$ so that we explore the MC with order up to 9. We use the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) to cluster the tree species.

The 13 trees species can be generally classified into two groups: 5 tree species from *Moraceae* and 8 tree species from *Fagaceae*. The two *Moraceaes*, *Ficus altissima* and *Ficus microcarpa*, should cluster together because they are known to be closely related and are both large hemiepiphytic trees while the other three *Moraceae* species are small dioecious shrubs. Within the *Fagaceae* group, the two *Castanopsis* species should cluster together, and the five *Lithocarpus* species should also form a subgroup. *Trigonobalanus doichangensis (Fagaceae)* is an ancestral genus that is very divergent from the rest of the family and has undergone considerable

sequence evolution. It should not group within the class of *Castanopsis* and *Lithocarpus* in *Fagaceae*.

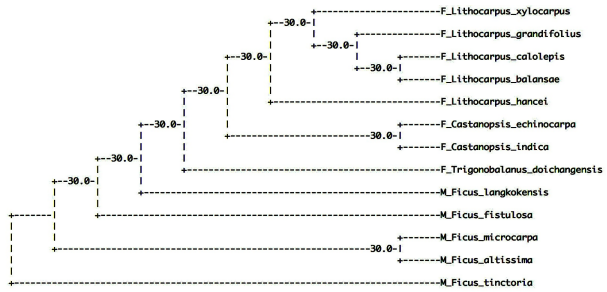
Figure 1 shows the clustering results of the 13 tree species using d_2^* under MCs of order 0 (i.i.d), 4, 8 and 9. The trees are built based on all the reads. From the results we can see, under the i.i.d model, *Lithocarpus* mixes up with *Castanopsis*; *T. doichangensis* can not be separated from the rest of *Fagaceae*, while under the MC of order greater than 4, *T. doichangensis* is successfully separated from the rest of the *Fagaceae*. Moreover, within the *Moraceae* group, *Ficus fistulas* and *Ficus langkokensis* form a subgroup under the i.i.d model, and they are separated under the MC with order greater than 4. While *F. langkokensis* is the closest *Maraceae* to the *Fagaceae* under 4th order MC, *F. fistulosa* becomes the closest species to the *Fagaceaes* under 8th and 9th order MCs.

In order to see whether the clustering of the tree species can be correctly inferred using only a portion of the shotgun read data, we randomly sample 10% of the total read data for each tree species to cluster them. To study the variation of the clusters due to random sampling of the reads, we repeat the sampling process 30 times and calculate the frequencies of each internal branch of the clustering using all the reads occurring among the 30 clusterings. The number on the branch refers to the frequency of the branch occurring among the 30 clusterings based on random sampled 10% reads. Three branches of the tree under MC of order 9 have frequencies of occurrence less than 30. When using the MC of a very high order, the clustering becomes unstable.

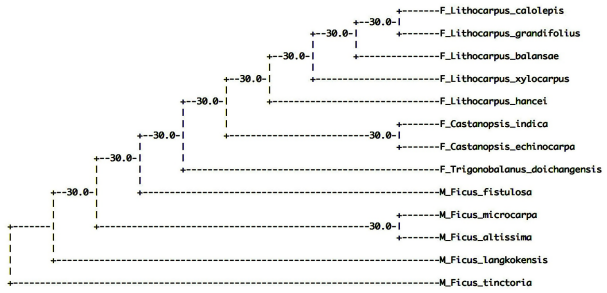
For the clustering results using d_2^S, see Figure S7. Under MC with all four orders, the two *Castanopsis* and the five *Lithocarpus* species are grouped separately, and *F. altissima (Moraceae)* and *F.microcarpa (Moraceae)* are clustered together. Under the i.i.d model, *T.doichangenesis (Fagaceae)* is successfully separated from *Lithocarpus*, but it is not the most outside species in the *Fagaceae* group. When the MC order is greater than 4, *T.doichangenesis (Fagaceae)* gets separated from the rest of the *Fagaceaes*. It can also be seen that when using the i.i.d model, or a MC with order 8 or greater, some of the branches becomes unstable.

In general, the results show that the clustering becomes more accurate as the order of MC increases using both d_2^* and d_2^S. Under the i.i.d model, the clustering based on d_2^* does not correctly separate *Castanopsis* from *Lithocarpus*, while the clustering based on d_2^S groups the two types separately. With higher order MCs, d_2^* successfully separates *Castanopsis* from *Lithocarpus*. The general clustering structure among *Lithocarpus*, *Castanopsis*, *Trigonobalanus* and *Ficus* stays correct when order is greater than 4 for both measures. When using the MC with order higher than the estimated order, the clustering is unstable and indeed the branch for *L.Hancei (Fagaceae)* is not supported on the last tree when using only 10% of the data. With a large number of parameters to estimate, 10% of the data does not suffice to capture the information in the data. The best clustering is achieved under a MC of order 8 and 9.

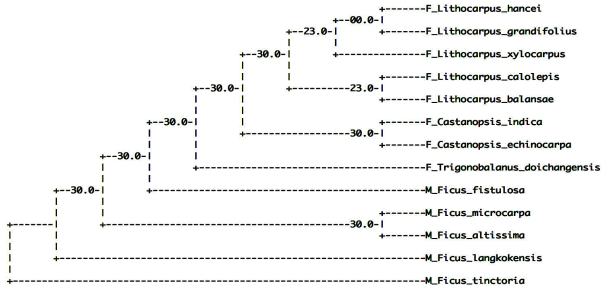
(a) k11, order0, d_2^*



(b) k11, order4, d_2^*



(c) k11, order8, d_2^*



(d) k11, order9, d_2^*

Fig. 1. The clustering of the 13 tropical tree species using d_2^* under MC with order 0 (i.i.d), 4, 8 and 9. The number on the branch refers to the frequency of the branch occurring among the 30 clusterings based on random sampled 10% reads. The letter 'F' at the beginning of the names represents *Fagaceae*; similarly the letter 'M' represents *Maraceae*.

4 DISCUSSION

Next generation sequencing technologies provide large amount of data in the form of short reads. Assembly of the millions of short reads to recover the long sequence is challenging, because the relative short length of the reads makes it difficult to resolve the repeat regions, not all regions may be covered, and assembly is time consuming. While multiple sequence alignment may be prohibitive, we can use word-count based dissimilarity measures to cluster the underlying species. These measures require an underlying probability model for the sequences; Markov chains are a reasonable model for such sequences. While transition probabilities can be

estimated directly from count data, estimating the order of a MC here is not straightforward.

Methods for estimating the order of a MC of a long sequence have been developed since the 1950s, but estimating the order of a MC directly from a set of short reads without assembly has not been studied yet. In this paper, we develop two statistics S_k^R and $Z_{\mathbf{w}}^R$ and show that both S_k^R and $Z_{\mathbf{w}}^R$ have surprisingly simple approximate distributions with only two parameters, one of them depending on the order of the original long MC sequence, and the other one depending on the distribution of the reads along the sequence. Intriguingly, one of these parameters is $d = c + 1$ under homogeneous sampling, where c is the coverage of the reads along the genome based on the first $\beta - k + 1$ positions of each read.

Based on the property of S_k^R and $Z_{\mathbf{w}}^R$, we develop an estimator for the order of a MC as well as an estimator for the parameter d based on NGS data. Extensive simulation studies are carried out to verify the theorem and evaluate the estimator.

Finally, we apply the estimation methods to two NGS data sets. Since the real genome sequences consist of coding, non-coding and various regulatory regions, single standard MC models do not fit the data well. Moreover, some enriched patterns, such as the motif sequences, are widespread throughout the genomes and violate the simple MC model for the whole genome sequence. Hence studying the fraction of k-words whose occurrences can be explained using the statistic $(Z_{\mathbf{w}}^R)^2/d$ by comparison to a χ_1^2 distribution is a more realistic way to determine the order of the MC for a real genome sequence. The estimated orders are consistent with the orders estimated directly from the full genome sequences using BIC methods.

Our primary motivation for this study is alignment-free genome comparison using NGS data. Further, we cluster the 28 species based on the NGS data using MC models with various orders. The results show that the clustering performs best and gives stable results when using a MC model with order on and above the estimated order. In addition, we apply the same analysis to 13 tropical tree species whose reference genomes are unknown; again the best clustering is achieved under a MC with the order within the estimated range.

When the sequence length is short or the sequencing depth is low, the numbers of occurrences of some k-words become small or even zero. Then the assumption of non-zero variance for all word counts which underlies the gamma approximation for S_k^R no longer holds and the gamma approximation may not work well. In such a situation an exact test for the order of MCs in the spirit of Besag and Mondal (2013) could be very helpful. In this paper we have only made a start on the Markov chain modelling of NGS data. An exhaustive study of errors in the data, in the form of power studies, could help to further understand the application range of our results. Finally, in this work we take the estimation of the transition probabilities for granted, once the order of the MC is determined. While the estimation of the transition probabilities of the MC model of a long sequence has been studied by Anderson and Goodman (1957) and Baum and Petrie (1966), it would be interesting to extend these methods to NGS data.

ACKNOWLEDGMENTS

The authors would like to thank anonymous referees for helpful comments on this work and on previous related work. We thank Dr. Xiaohui Xie from UCI for explanation of the 28 vertebrate species. The research is supported by National Natural Science Foundation of China (No.10871009, 10721403), and National Key Basic Research Project of China (No.2009CB918503). FS is partially supported by US NIH P50 HG 002790 and NSF DMS-1043075 and OCE 1136818. GDR is partially supported by EPSRC EP/K032402/1.

Conflict of Interest: : None declared.

REFERENCES

Almagor, H. (1983). A Markov analysis of DNA sequences. *Journal of Theoretical Biology*, **104**(4), 633–645.

Anderson, T. W. and Goodman, L. A. (1957). Statistical inference about Markov chains. *The Annals of Mathematical Statistics*, **28**(4), 89–110.

Arnold, J., Cuticchia, A. J., Newsome, D. A., Jennings, W. W., and Ivarie, R. (1988). Mono-through hexanucleotide composition of the sense strand of yeast DNA: a Markov chain analysis. *Nucleic Acids Research*, **16**(14), 7145–7158.

Avery, P. J. (1987). The analysis of intron data and their use in the detection of short signals. *Journal of Molecular Evolution*, **26**(4), 335–340.

Avery, P. J. and Henderson, D. A. (1999). Fitting Markov chain models to discrete state series such as DNA sequences. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **48**(1), 53–61.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, **37**(6), 1554–1563.

Behnam, E. and Smith, A. D. (2014). The amordad database engine for metagenomics. *Bioinformatics*, **30**(20), 2949–2955.

Behnam, E., Waterman, M. S., and Smith, A. D. (2013). A geometric interpretation for local alignment-free sequence comparison. *Journal of Computational Biology*, **20**(7), 471–485.

Benjamini, Y. and Speed, T. (2012). Summarizing and correcting the gc content bias in high-throughput sequencing. *Nucleic Acids Research*, **40**(10), e72.

Besag, J. and Mondal, D. (2013). Exact goodness-of-fit tests for Markov chains. *Biometrics*, **69**(2), 488–496.

Billingsley, P. (1961a). *Statistical Inference for Markov Processes*, volume 2. University of Chicago Press Chicago.

Billingsley, P. (1961b). Statistical methods in Markov chains. *The Annals of Mathematical Statistics*, **32**(1), 12–40.

Blaisdell, B. E. (1985). Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding. *Journal of Molecular Evolution*, **21**(3), 278–288.

Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., *et al.* (2010). Chip-seq identification of weakly conserved heart enhancers. *Nature Genetics*, **42**(9), 806–810.

Cannon, C. H., Kua, C. S., Zhang, D., and Harting, J. (2010). Assembly free comparative genomics of short-read sequence data discovers the needles in the haystack. *Molecular Ecology*, **19**(Suppl. 1), 146–160.

Daley, T. and Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. *Nature Methods*, **10**(4), 325–327.

Ewens, W. J. and Grant, G. R. (2005). *Statistical methods in bioinformatics: an introduction*. Springer.

Göke, J., Schulz, M. H., Lasserre, J., and Vingron, M. (2012). Estimation of pairwise sequence similarity of mammalian enhancers with word neighbourhood counts. *Bioinformatics*, **28**(5), 656–663.

Hoel, P. G. (1954). A test for Markov chains. *Biometrika*, **41**(3/4), 430–433.

Hong, J. (1990). Prediction of oligonucleotide frequencies based upon dinucleotide frequencies obtained from the nearest neighbor analysis. *Nucleic Acids Research*, **18**(6), 1625–1628.

Hurwitz, B. L., Westveld, A. H., Brum, J. R., and Sullivan, M. B. (2014). Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. *Proceedings of the National Academy of Sciences*, **111**(29), 10714–10719.

Jiang, B., Song, K., Ren, J., Deng, M., Sun, F., and Zhang, X. (2012). Comparison of metagenomic samples using sequence signatures. *BMC Genomics*, **13**(1), 730.

Karolchik, D., Kuhn, R. M., Baertsch, R., Barber, G. P., Clawson, H., Diekhans, M., Giardine, B., Harte, R. A., Hinrichs, A. S., Hsu, F., *et al.* (2008). The UCSC genome browser database: 2008 update. *Nucleic Acids Research*, **36**(suppl 1), D773–D779.

Lander, E. S. and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**(3), 231–239.

Liu, X., Wan, L., Li, J., Reinert, G., Waterman, M., and Sun, F. (2011). New powerful statistics for alignment-free sequence comparison under a pattern transfer model. *Journal of Theoretical Biology*, **284**(1), 106–116.

Miller, W., Rosenbloom, K., Hardison, R., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D., Baertsch, R., Blankenberg, D., *et al.* (2007). 28-way vertebrate alignment and conservation track in the UCSC genome browser. *Genome Research*, **17**(12), 1797–1808.

Narlikar, L., Mehta, N., Galande, S., and Arjunwadkar, M. (2013). One size does not fit all: On how markov model order dictates performance of genomic sequence analyses. *Nucleic Acids Research*, **41**(3), 1416–1424.

Pevzner, P. A., Borodovsky, M. Y., and Mironov, A. A. (1989). Linguistics of nucleotide sequences i: the significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *Journal of Biomolecular Structure and Dynamics*, **6**(5), 1013–1026.

Reinert, G., Schbath, S., and Waterman, M. (2000). Probabilistic and statistical properties of words: an overview. *Journal of Computational Biology*, **7**(1-2), 1–46.

Reinert, G., Schbath, S., and Waterman, M. S. (2005). Statistics on words with applications to biological sequences. *Lothaire: Applied Combinatorics on Words, J. Berstel and D. Perrin, eds.*, **105**, 251–328.

Reinert, G., Chew, D., Sun, F. Z., and Waterman, M. S. (2009). Alignment-free sequence comparison (I): Statistics and power. *Journal of Computational Biology*, **16**(12), 1615–1634.

Ren, J., Song, K., Sun, F., Deng, M., and Reinert, G. (2013). Multiple alignment-free sequence comparison. *Bioinformatics*, **29**(21), 2690–2698.

Richter, D., Ott, F., Auch, A., Schmid, R., and Huson, D. (2008). MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One*, **3**(10), e3373.

Simpson, J. T. (2014). Exploring genome characteristics and sequence quality without a reference. *Bioinformatics*, **30**(9), 1228–1235.

Song, K., Ren, J., Zhai, Z., Liu, X., Deng, M., and Sun, F. (2013). Alignment-free sequence comparison based on next-generation sequencing reads. *Journal of Computational Biology*, **20**(2), 64–79.

Song, K., Ren, J., Reinert, G., Deng, M., Waterman, M. S., and Sun, F. (2014). New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings in Bioinformatics*, **15**(3), 343–353.

Wang, Y., Liu, L., Chen, L., Chen, T., and Sun, F. (2014). Comparison of metatranscriptomic samples based on k-tuple frequencies. *PloS One*, **9**(1), e84348.

Waterman, M. S. (1995). *Introduction to Computational Biology: Maps, Sequences and Genomes*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis.

Yi, H. and Jin, L. (2013). Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Research*, **41**(7), e75.

Zhai, Z., Reinert, G., Song, K., Waterman, M. S., Luan, Y., and Sun, F. (2012). Normal and compound poisson approximations for pattern occurrences in ngs reads. *Journal of Computational Biology*, **19**(6), 839–854.

Zhang, Z. D., Rozowsky, J., Snyder, M., Chang, J., and Gerstein, M. (2008). Modeling chip sequencing in silico with applications. *PLoS Computational Biology*, **4**(8), e1000158.