Synthetic and Systems Biotechnology 4 (2019) 150–156

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: http://www.keaipublishing.com/synbio

The statistical power of *k*-mer based aggregative statistics for alignment-free detection of horizontal gene transfer

Guan-Da Huang[a,1], Xue-Mei Liu[a,1], Tian-Lai Huang[a], Li- C. Xia[b,*]

[a] *School of Physics and Optoelectronics, South China University of Technology, Guangzhou, 510640, China*
[b] *Department of Medicine, Stanford University School of Medicine, Stanford, CA, 94305, USA*

ARTICLE INFO

Keywords:
Alignment-free sequence comparison
k-mer
Horizontal gene transfer
Statistical power

ABSTRACT

Alignment-based database search and sequence comparison are commonly used to detect horizontal gene transfer (HGT). However, with the rapid increase of sequencing depth, hundreds of thousands of contigs are routinely assembled from metagenomics studies, which challenges alignment-based HGT analysis by overwhelming the known reference sequences. Detecting HGT by *k*-mer statistics thus becomes an attractive alternative. These alignment-free statistics have been demonstrated in high performance and efficiency in whole-genome and transcriptome comparisons. To adapt *k*-mer statistics for HGT detection, we developed two aggregative statistics T^S_{sum} and T^*_{sum}, which subsample metagenome contigs by their representative regions, and summarize the regional D^S_2 and D^*_2 metrics by their upper bounds. We systematically studied the aggregative statistics' power at different *k*-mer size using simulations. Our analysis showed that, in general, the power of T^S_{sum} and T^*_{sum} increases with sequencing coverage, and reaches a maximum power > 80% at $k = 6$, with 5% Type-I error and the coverage ratio > 0.2x. The statistical power of T^S_{sum} and T^*_{sum} was evaluated with realistic simulations of HGT mechanism, sequencing depth, read length, and base error. We expect these statistics to be useful distance metrics for identifying HGT in metagenomic studies.

1. Introduction

Horizontal gene transfers (HGT) is the transversal exchange of genetic material between different organisms, cells, or organelles, as well as that between organisms and environment. It accelerates the speed of biological adaptation to the environment, promotes the evolution of organisms, and promotes the convergence of species [1]. The main modes of HGT in prokaryotic organisms are conjugation, transformation and transduction [2–4], while the mode of HGT for eukaryotic organisms are complex, with increased potential in the presence of viral transfection, host parasites or direct contacts of symbiotic organisms [5].

Researchers are highly interested in identifying and analyzing HGT in metagenomics sequence data, so as to understand how it reshapes and rebuilds the microbial community in response to a changing environment. The computational methods for predicting HGT genes from biological sequences include: (1) phylogenetic methods, which uses bootstrap resampling [6,7] and maximum likelihood framework [8,9] for identifying the significant differences between phylogenetic trees; (2) parametric methods, which extract the characteristics of genes from a genome as its tags [10]. These tags can include sequence structure-based features [11], statistical features [12,13], and information entropy-based features [14]; and (3) sequence distance-based methods, which detects HGT using sequence comparison.

The distance-based methods were predominated by alignment-based metrics. These methods require computing the optimal pairwise or multiple alignment (locally or globally) using dynamic programming, such as algorithms implemented by ClustalW [15] and Muscle [16], or performing alignment-based database search, such as algorithms implemented by BWA [17], Blat [18], Blast [19] and Fasta [20]. However, in either cases, finding the optimal alignment demands a significant amount of computation both in time and memory space. Consequently, these alignment-based methods are limited by their throughput in comparatively analyzing a large number of contigs as seen in today's metagenomics studies.

Alternatively, alignment-free statistics are potential useful distance metrics, which are more efficient computationally and do not require high sequence homology [21,22]. In principle, alignment-free methods

Peer review under responsibility of KeAi Communications Co., Ltd.

[*] Corresponding author.

E-mail addresses: 1003589085@qq.com (G.-D. Huang), liuxm@scut.edu.cn (X.-M. Liu), 632609144@qq.com (T.-L. Huang), l.c.xia@stanford.edu (L.-.C. Xia).

[1] Equal contribution.

https://doi.org/10.1016/j.synbio.2019.08.001
Received 6 May 2019; Received in revised form 14 July 2019; Accepted 5 August 2019

decompose the DNA or protein sequences into short subsequences (or k-mers). They first compute k-mer frequency vectors based on k-mer occurrence and then calculate the sequence distances based on these vectors. These methods originated from the application of comparing computer programs [23] and were introduced to molecular sequence analysis by the early works of Hao et al. [24] and others [25] dating back to the early 2000s. Notably, Dr. Hao's [26] work in CV-Tree methods represents the first systematic whole-genome phylogenetic attempt using alignment-free k-mer statistics.

Nowadays, k-mer based statistics have become a cornerstone in alignment-free whole-genome and gene-based sequence comparisons [27–30]. It was proved useful in predicting protein structure and function [31], in predicting HGT [32,33], and in constructing phylogenetic trees [34]. Those applications typically involve comparing conserved sites or homology regions between two sequences and quantifying molecular evolutionary relationships at the sequence level. Those successful applications were built upon the good performance that k-mer based statistics and distance metrics have in clustering and classifying amino acid and nucleic acid sequences.

In this paper, we specifically extended the aggregative k-mer statistics T_{sum} previously developed by Liu et al. [35] to the task of detecting HGT in metagenomics sequence data. We developed a sequence subsampling scheme that computes and summarizes the k-mer statistic metrics from the regional contigs using their upper bounds with aggregative statistics. The rationale is that the identified regions of the two contigs that maximize such k-mer metrics will also have the highest potential for being HGT. We parameterized the subsampling based on realistic assumptions of sequencing platform and HGT mechanisms. We tested the developed aggregative statistics in the cases of seven different k-mer metrics and with a k-mer size ranging from 4 to 8 by simulations. We found T^S_{sum} and T^*_{sum}, as parameterized by D^S_2 and D^*_2 metrics respectively, at the k-mer size 6 have the best power for detecting HGT events.

2. Materials and methods

k-mer based alignment-free statistics, first count the k-mer frequency and then correlate these k-mer frequency vectors (or their transformations), to evaluate the similarity, or homology, between two biological sequences. A k-mer is defined as a short subsequence of the DNA (nucleic acid) or protein (amino acid) sequence. The length of a k-mer is defined by the value of k (e.g., if $k = 4$, then the k-mer length is 4 base pair). In the case of DNA sequences, a k-mer is composed of k random letters from the alphabet $\{A, T, G, C\}$. When a k is chosen, there are 4^k possible k-mers and their occurrence in a sequence can be recorded in a 4^k dimension k-mer frequency vector. All alignment-free k-mer statistics based themselves on the observed fact that these k-mer frequency vectors will show higher similarity if the two biological sequences being compared are more evolutionarily related.

2.1. Alignment-free k-mer statistics

Formally, we can define two molecular sequences as X and Y and their shared sequence length n. For a given k, the occurrence of all possible k-mer words, as denoted by $w = w_1 w_2 \cdots w_k$, is counted for the sequence X and recorded in the k -mer frequency vector X_w. Similarly, the k-mer words are counted and recorded for the sequence Y in Y_w. Note, both X_w and Y_w are 4^k dimensional vectors made up of the occurrence numbers of all possible 4-mers, which is denoted by Λ^k. Therefore, the similarity, or relatedness, between the two sequences can be measured by correlating the two k-mer frequency vectors X_w and Y_w. This similarity or relatedness metric can be transformed to obtain proper dissimilarity or distance metrics.

In this setting, Torney et al. [36], derived the similarity metric D_2 statistic as:

$$D_2 = \sum_{w \in \Lambda^k} X_w Y_w \tag{1}$$

Kantorovitz et al. [37] extended D_2 to D_2Z, which allows a generalized subtraction of background distribution from the D_2 statistic. The technique proved very effective for predicting the regulatory region of seemingly unrelated genes. E.g., Foret et al. [38,39] found that the D_2 statistic is superior to both Blast and exact k-mer matching in the evolutionary comparison of sequences. It can be seen from Eq. (1) that D_2 correlates the occurrence numbers of all k-mers directly, without adjusting for the total number of k-mers presented in individual sequences, which means that D_2 is subject to bias in sequence length as well as background sequencing noise incorporated in individual sequence.

Reinert et al., [40] then identified that the performance of D_2 in large-scale biological sequences comparison could be improved by adjusting for the noise as introduced by individual sequences. They developed two generalized versions of D_2, namely D^S_2 and D^*_2 [41,42] and proved that they are advantageous to D_2 in comparing sequences. These statistics were derived following the theorem by Shepp et al. [43], which states that if the independent variables U and V are normally distributed and their means are zero, then $UV/(U^2 + V^2)^{1/2}$ is also normally distributed. Based on that fact, Reinert and Liu standardized the D_2 statistic for raw k-mer frequency vectors X_w and Y_w, as the follows:

$$\widetilde{X_w} = X_w - (n - k + 1)p_w \tag{2}$$

$$\widetilde{Y_w} = Y_w - (n - k + 1)p_w \tag{3}$$

$$D^S_2 = \sum_{w \in \Lambda^k} \left(\widetilde{X_w}\widetilde{Y_w} / \sqrt{\widetilde{X_w}^2 + \widetilde{Y_w}^2} \right) \tag{4}$$

$$D^*_2 = \sum_{w \in \Lambda^k} (\widetilde{X_w}\widetilde{Y_w} / (n - k + 1)p_w) \tag{5}$$

Here, p_w is the probability of the word w (w= $w_1 w_2 \cdots w_k$) occurs in the sequence and $p_w = p_{w_1} p_{w_2} p_{w_3} \ldots p_{w_k}$ under an independent and identical distribution (*i.i.d.*) model.

2.2. Aggregative measures T^S_{sum} and T^*_{sum}

To extend D^S_2 and D^*_2 to whole-genome sequence comparison, in which high intra-sequence heterogeneity could alter local k-mer distributions, Liu (2011) et al. introduced an aggregative measure T_{sum}. T_{sum} subsamples the sequence distances using sliding windows and summarizes them using local upper bounding of D^S_2 and D^*_2 statistics. The subsampling strategy was configurable by the window and shift sizes. Using simulations, Liu et al. found that the statistical power of T_{sum}, which aggregates D^S_2 and D^*_2 statistics, was maximized when the whole-genome sequence was subsampled with non-overlapping sliding windows.

Following the procedure of Liu et al. (see Fig. 1), we defined F_1 to F_N as the subsampled fragments, and G the size of subsampling gap. We defined X^S_i as the maxima of D^S_2 between the ith subsampled fragment of X and all subsampled fragments of Y. The same was true for Y^S_i. Thus,

$$X^S_i = \max_{1 \le j \le N} D^S_2(F^X_i, F^Y_j) \tag{6}$$

$$Y^S_j = \max_{1 \le i \le N} D^S_2(F^X_i, F^Y_j) \tag{7}$$

Next, by summarizing over all X^S_i’s and Y^S_i’s, we defined the aggregative k-mer statistic T_{sum} between X and Y as:

$$T^S_{sum} = \sum_{i=1}^{N} X^S_i + \sum_{j=1}^{N} Y^S_j \tag{8}$$

Similarly, we can derive X^*_i, Y^*_i and T^*_{sum} – the aggregative k-mer statistic for X^*_i’s and Y^*_i’s when D^*_2 is used.

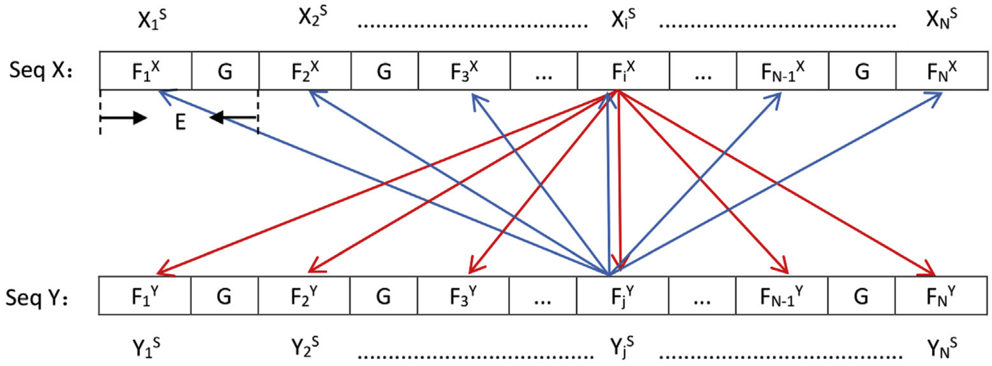


Fig. 1. The T_{sum} resampling scheme. (On Seq X and Seq Y, F_1 to F_N are the subsampled fragments, and G is gap. X_i^S is the maxima of D_2^S between the *i*th subsampled fragment of X and all subsampled fragments of Y. The same for Y_i^S is the maxima of D_2^S between the *i*th subsampled fragment of Y and all subsampled fragments of X.)

To simply the simulation, we defined the parameter genome coverage rate R as:

$$R = \mathrm{F}/(\ \mathrm{F}+\ \mathrm{G}) \times 100\% \tag{9}$$

and the combined gap and fragment length, i.e. the entire genome segment length E as:

$$E = F + G \tag{10}$$

It is thus convenient to vary E and R in our simulation with this new parameterization, where R represents the fraction of unknown parts of sequences and E represents the expected genome segment given full sequence information. We then evaluated the statistical power of T_{sum}^S and T_{sum}^* by simulating a large number of HGT and non-HGT sequence pairs with a range of parameters representing real NGS metagenomics data.

2.3. Simulation model and statistical power

We built two computer models to simulate the background and foreground biological sequences respectively. We simulated the background sequences as pairs of independently and identically distributed (*i.i.d.*) random letter sequences of the same length from the nucleic acid alphabet $\{A, T, G, C\}$. These pairs of sequences were considered as unrelated and assigned to the control group.

For the foreground simulation, we considered both uniform and non-uniform distributions of random letters. In the case of uniform distribution, the probability of observing A, C, G, and T in a sequence is all the same at 1/4. We refer to this simulated distribution as the equally independent and identical distribution (*e.i.i.d*) scheme. In the case of non-uniform distribution, we assign a probability of 1/3 to observing G and C, and a probability of 1/6 to observing A and C. We refer to this simulated non-equally independent and identical distribution (*n.i.i.d*) as the *n.i.i.d* (gc-rich) scheme because of its enrichment of G and C bases in the resultant sequences.

Reinert and Sun et al. (2009) proposed two procedures to simulate related sequence pairs representing true biological relationships, one for mimicking the real sequence compositions as seen in cis-regulatory modules (**CRM**) and the other for mimicking that as seen in horizontal gene transfer. As we are studying the HGT in prokaryotes, we simulated the foreground sequence pairs by adapting their HGT procedure (see Fig. 2). Specifically, we randomly selected a set of motif sites in the first background sequence of a pair. We then transferred these motifs to the other background sequence by using them to replace the corresponding sites in the second sequence. These site positions were randomly selected by generating the Bernoulli random numbers (1s for being selected) with the probability that a position gets selected was 0.05. The length of a motif L was set to 5.

We computed the T_{sum}^S and T_{sum}^* statistics for all such generated background and foreground sequence pairs. To compute the statistical power, we regarded background sequence pairs as true negatives and foreground pairs as true positives. We set the required statistical significance level (Type-I error rate) to $a = 5\%$, which is the fraction of true negatives that were also declared positive by the statistic. The statistical significance threshold t was learned from a statistic's 1-$a = 95\%$ percentile value (ranked in ascendance) from 10,000 simulated background sequence pairs. As we applied the threshold t to the collection of T_{sum} (either T_{sum}^* or T_{sum}^S) statistics computed for the foreground pairs, we estimated the fraction of true positives which were not declared positive by the threshold, which is the Type-II error, i.e. 1-beta. Beta is then the statistical power with a range from 0 to 1 and we computed it as follows:

$$\mathrm{Power}\ (\beta) = \mathrm{N}_{T_{sum} \geq t}/10000 \tag{11}$$

The individual simulation element is analogous to the hypothesis testing for each pair by which the H_0 states the two sequences is not related (null), while the H_1 states otherwise. When the statistic computed is larger than t, the alternative hypothesis H_1 is accepted, which is correct if the pair is foreground, or wrong if the pair is background. Conversely, if the statistic is less than t, the null hypothesis H_0 is retained, which is correct if the pair is background, or wrong if it is foreground.

Our overall simulation process was as follows: first, we set the common motif length L in the range of (4,5,6,7,8) and k-mer size k in the range of (4,5,6,7,8); Next, we calculated the power values of T_{sum}^S and T_{sum}^* for each k and L parameter combination with the coverage rate R ranging from 25% to 75%; Then, we iterated the power computation for the genome segment length ranging from 1000 to 10,000 with the increment of 1000; In each iteration, we simulated 10,000 pairs of background and foreground sequences to compute power; Finally, the obtained power values were plotted in curves where the X-axis is the genome length and the Y-axis is the power. Finally, we identified the optimal k for different L values based on genome segment length and the obtained maximum statistical power.

2.4. Simulation and statistical software

We developed a software package for carrying out the power analysis, termed SeqPowerK, using the C# language. SeqPowerK can compute the statistical power for generic k-mer statistics with versatile parameter specifications as detailed in its user's manual (https://github.com/liuxuemeiscut/SeqPowerK).

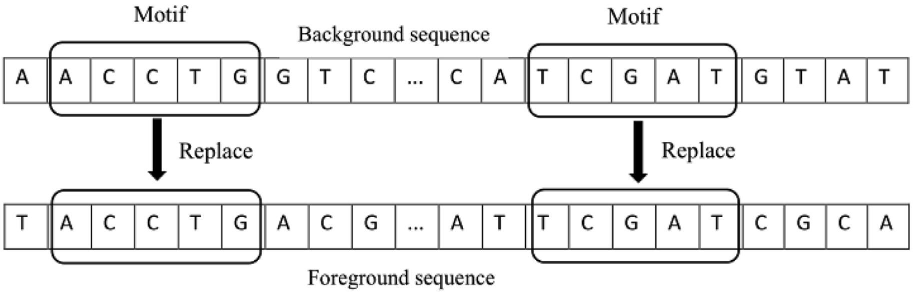
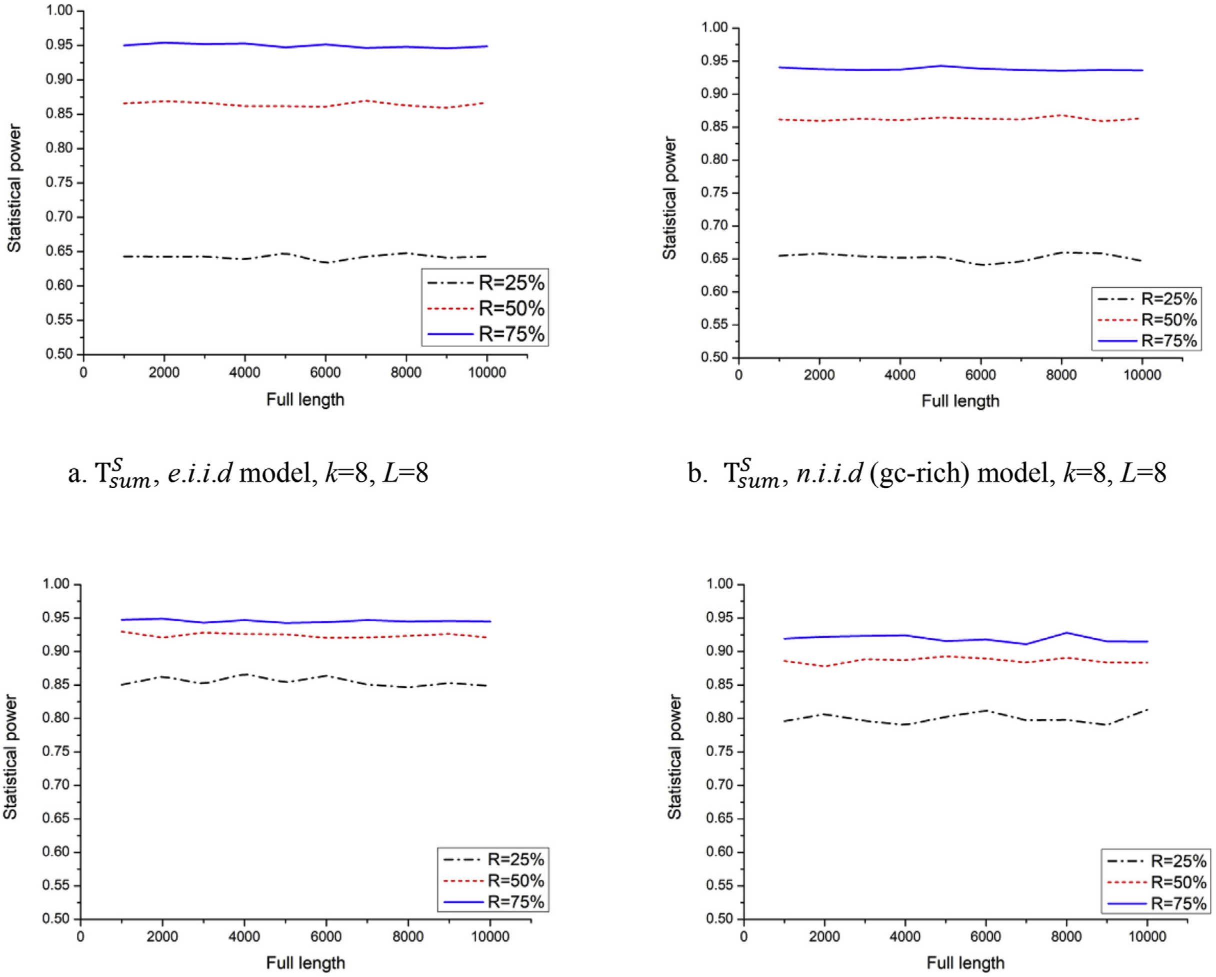

Motif
Background sequence
Motif
A A C C T G G T C ... C A T C G A T G T A T
Replace
Replace
T A C C T G A C G ... A T T C G A T C G C A
Foreground sequence

Fig. 2. Simulating the foreground sequence pairs using the Horizontal Gene Transfer (HGT) procedure with motif length $L = 5$.

3. Results and discussion

3.1. The effect of subsampling coverage rate *R* on the power of T^{S}_{sum} and T^{*}_{sum}

The subsampling coverage rate *R* is an important parameter because it represents the fraction of genomes available for alignment-free analysis. Under different combinations of *k* and *L*, also considering both the *e.i.i.d* and *n.i.i.d* (gc-rich) background schemes, we explored the statistical power of T^{S}_{sum} and T^{*}_{sum} with a varying subsampling coverage rate *R* from 25%, 50%, to 75% (see Fig. 3). For both T^{S}_{sum} and T^{*}_{sum}, when the coverage rate *R* is high, the statistical power is high. For example, at $R = 25\%$, the statistical power is only 65%, which is significantly lower than that of $R = 50\%$ and 75%, when the powers are > 85%. The statistical power at $R = 50\%$ and 75% are reasonably close, suggesting that the T_{sum} statistic efficiency is saturated by a reasonable high coverage rate at ~50%.



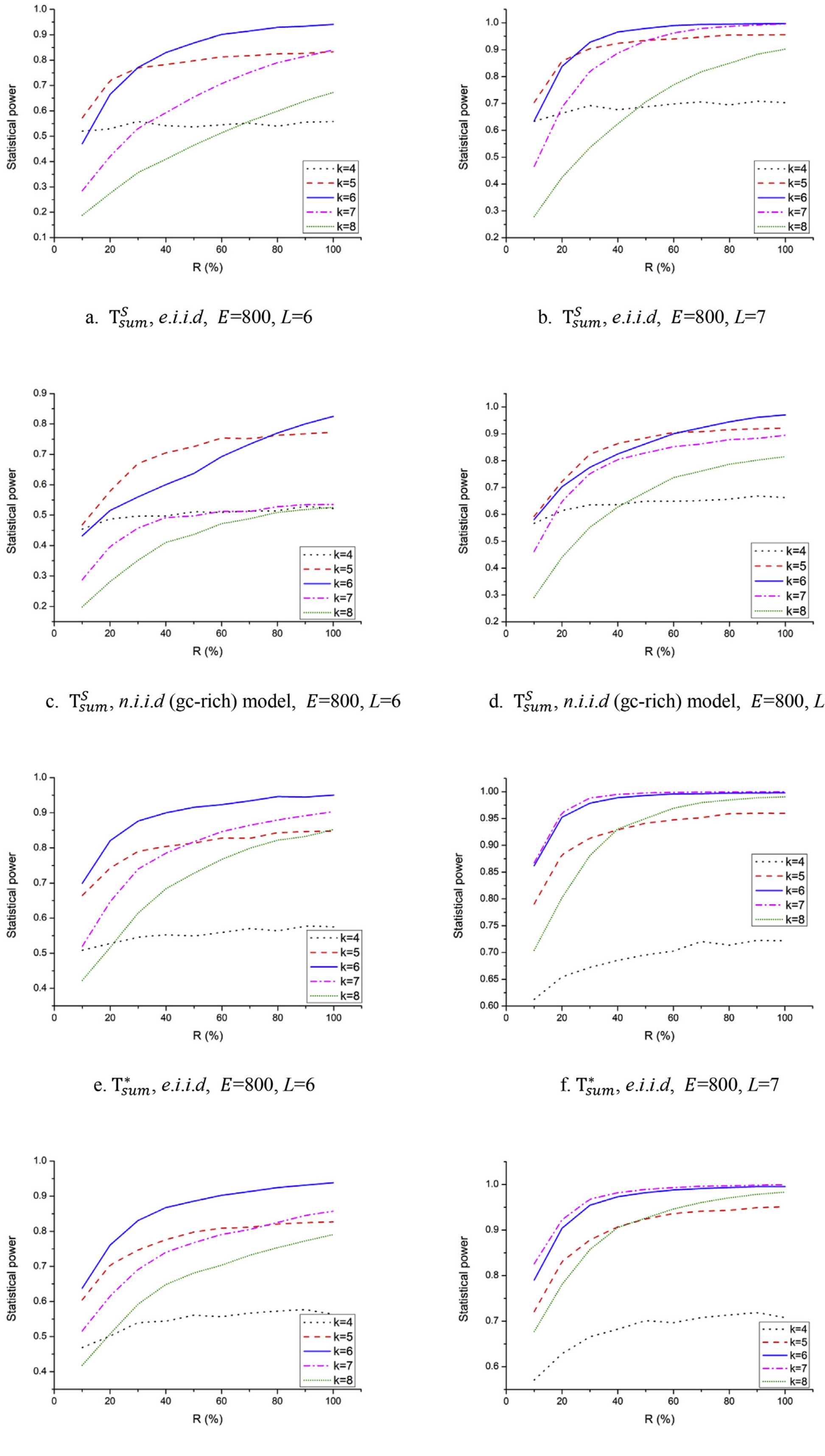
a. T^{S}_{sum}, *e.i.i.d* model, *k*=8, *L*=8

b. T^{S}_{sum}, *n.i.i.d* (gc-rich) model, *k*=8, *L*=8

c. T^{*}_{sum}, *e.i.i.d* model, *k*=6, *L*=6

d. T^{*}_{sum}, *n.i.i.d* (gc-rich) model, *k*=6, *L*=6

Fig. 3. Statistical power of T^{S}_{sum} and T^{*}_{sum} when the coverage rate $R = 25\%$, 50%, 75%. (Two statistics T^{S}_{sum} and T^{*}_{sum} are used in this figure. Full length is the length of whole sequence, *R* is subsampling coverage rate, *k* is *k*-mer's length and *L* is the length of motif. The probability of four bases is $P(A) = P(C) = P(G) = P(T) = 1/4$ in *e.i.i.d* model, and $P(A) = P(T) = 1/6$, $P(G) = P(C) = 1/3$ in *n.i.i.d* (gc-rich) model.).



a. T^{S}_{sum}, *e.i.i.d*, *E*=800, *L*=6

b. T^{S}_{sum}, *e.i.i.d*, *E*=800, *L*=7

c. T^{S}_{sum}, *n.i.i.d* (gc-rich) model, *E*=800, *L*=6

d. T^{S}_{sum}, *n.i.i.d* (gc-rich) model, *E*=800, *L*=7

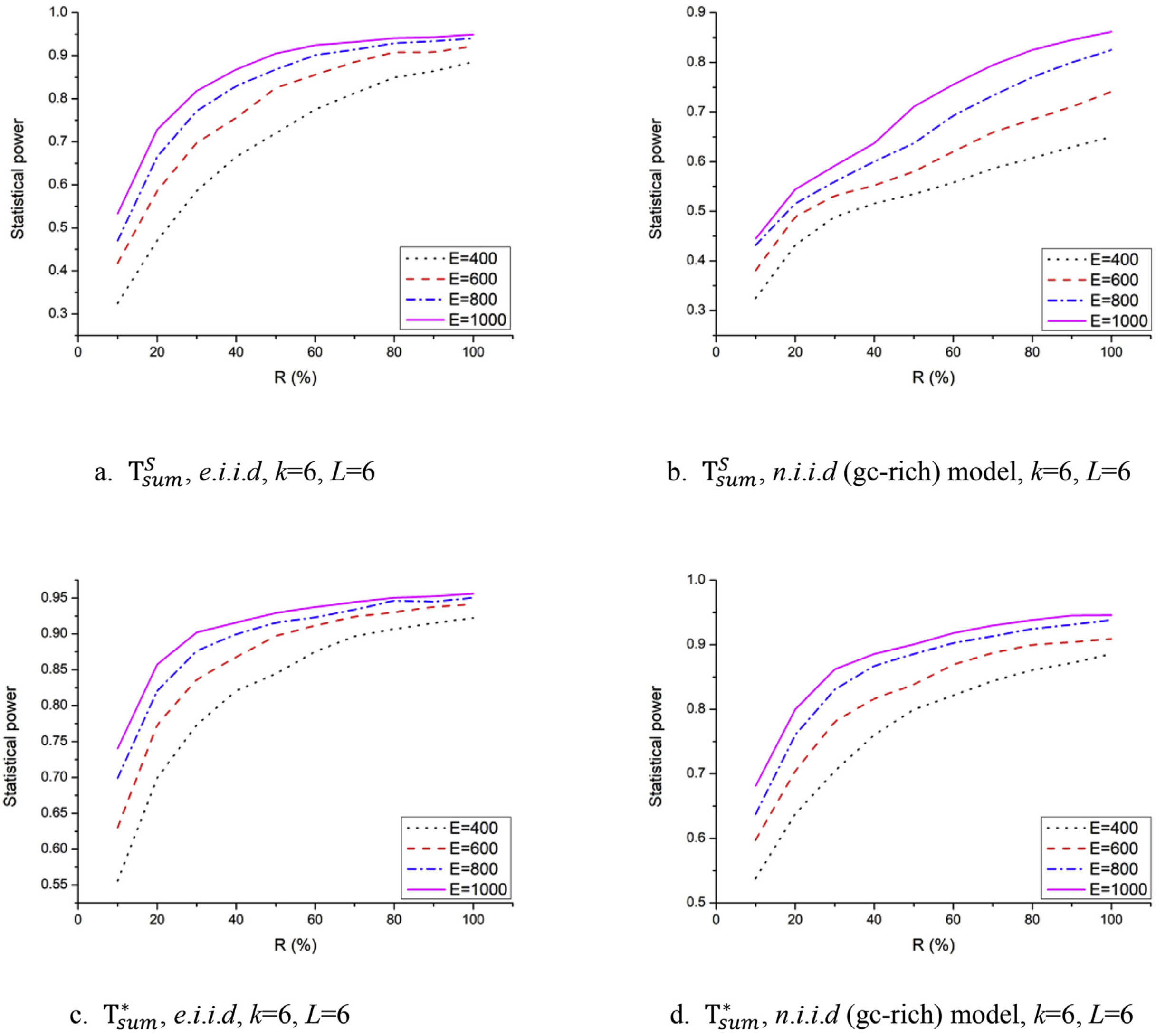
e. T^{*}_{sum}, *e.i.i.d*, *E*=800, *L*=6

f. T^{*}_{sum}, *e.i.i.d*, *E*=800, *L*=7

g. T^{*}_{sum}, *n.i.i.d* (gc-rich) model, *E*=800, *L*=6

h. T^{*}_{sum}, *n.i.i.d* (gc-rich) model, *E*=800, *L*=7

Fig. 4. Statistical power of T^{S}_{sum} and T^{*}_{sum} when k = 4, 5, 6, 7, 8. (Two statistics T^{S}_{sum} and T^{*}_{sum} are used in this figure. *E* is the entire genome segment length, R is subsampling coverage rate, k is k-mer's length and L is the length of motif. The probability of four bases is P(A) = P(C)=P(G) = P(T) = 1/4 in *e.i.i.d* model, and P(A) = P(T) = 1/6, P(G) = P(C) = 1/3 in *n.i.i.d* (gc-rich) model.)



a. T^S_{sum}, *e.i.i.d*, *k*=6, *L*=6

b. T^S_{sum}, *n.i.i.d* (gc-rich) model, *k*=6, *L*=6

c. T^*_{sum}, *e.i.i.d*, *k*=6, *L*=6

d. T^*_{sum}, *n.i.i.d* (gc-rich) model, *k*=6, *L*=6

Fig. 5. Statistical power of T^S_{sum} and T^*_{sum} when E = 400, 600, 800, 1000. (Two statistics T^S_{sum} and T^*_{sum} are used in this figure. E is the entire genome segment length, R is subsampling coverage rate, k is k-mer's length and L is the length of motif. The probability of four bases is $P(A) = P(C) = P(G) = P(T) = 1/4$ in *e.i.i.d* model, and $P(A) = P(T) = 1/6$, $P(G) = P(C) = 1/3$ in *n.i.i.d* (gc-rich) model.)

3.2. The optimal k-mer size for achieving the best power of T^S_{sum} and T^*_{sum}

The k-mer size choice is typically the first and foremost important parameter influencing alignment-free sequence comparison. Power simulations are useful ways to find the optimal k value given realistic parameter settings. We simulated HGT sequence data with motif length L in (4,5,6,7,8) and fixed $E = 800$. We compared and identified the optimal k when the statistical power is the highest given these parameters (see Fig. 4). We first observed that if the k is too small, the statistics power is low. When $k = 4$, the highest statistical power is < 70%. When the k is larger than 5, the statistical power is generally better than 80%. And the optimal k depends on coverage rate and the underlying sequence relatedness as represented by L.

We also observed that the statistical power of T^S_{sum} and T^*_{sum} differed slightly on k values. For both *e.i.i.d* and *n.i.i.d* (gc-rich) background models, no matter of the values of L, the optimal k for the highest power of T^S_{sum} is either 5 or 6 (Fig. 4a–d), while the optimal k of T^*_{sum} is 6 or 7 (Fig. 4e–h). Based on these facts, $k = 6$ is likely the best choice for general alignment-free sequence analysis, for example, the statistical power at $k = 6$ is consistently > 80% when the coverage rate is > 20%, which is at least comparable to and often times far better than what were achievable by other k values for either T^S_{sum} or T^*_{sum}.

We also compared the classification performance between T^S_{sum} and T^*_{sum}. Overall, the power of T^*_{sum} was higher than that of T^S_{sum}. As we known, T^*_{sum} was based on D_2^* and T^S_{sum} was based on D_2^S, which could mean that D_2^* is better than D_2^S in this type of application. This is because the subsampling procedure divides the sequence into a series of genome segments. Previous research based on D_2 statistics showed that, in general, D_2^S and D_2^* have higher power than D_2 and that D_2^S is more suitable for longer sequence comparisons, while D_2^* is more suitable for shorter sequences. However, in practice, we often do not know the genome segment length of the related subsequences. In this respect, the robustness of a statistic is required, and T^S_{sum} is generally better than T^*_{sum}.

3.3. The effect of genome segment length E on the power of T^S_{sum} and T^*_{sum}

We also studied the statistical power with different segment length E under $k = 6$ and $L = 6$. We varied genome segment size E form 400, 600, 800 to 1000. Fig. 5 showed that the statistical power of T^S_{sum} and T^*_{sum} positively correlates with E. However, compared with the coverage rate R, the genome segment length E has only a modest effect on the power. The maximal difference in power is < 0.1. Segment length E represents the degree of fragmentations in the subsampling process.

Segment length E determines the amount of sequences per fragment that is available for estimating *k*-mer frequency. Thus the smaller E the lessser accurate such estimates are, which in turn reduces statistical power.

4. Conclusions

HGT describes how organisms transfer genetic material to other cells rather than offspring. HGT plays a key role in the evolution of species and microbial genome diversity. With the increasing number of NGS data, the prediction of HGT is of great practical significance for better understanding the impact of environment on community structure. So far, the many methods relying upon sequence alignment to identify HGT were burdened by computation challenges. *k*-mer based alignment-free sequence comparison were effective in comparing multiple sequences and identifying HGTs without incurring significant computational cost.

In this paper, we proposed two new aggregative *k*-mer statistics T^{S}_{sum} and T^{*}_{sum} by subsampling and finding the upper bounds of underlying D^{S}_{2} and D^{*}_{2} statistics to identify HGT. We conducted an extensive simulation benchmark to evaluate the statistical power of various *k*-mer distances using these aggregative metrics. Our power analysis showed that, in general, the statistical power of *k*-mer statistics increases with sequencing coverage. We found that the optimal *k*-mer size for sequence comparison is 6 for both T^{S}_{sum} and T^{*}_{sum}, which ensures a statistical power > 80% given the assumed Type-I error rate of 5% and the genome coverage rate > 0.2x. To summarize, we propose T^{S}_{sum} and $T^{*}_{sum}$with D^{S}_{2} and D^{*}_{2} metrics and *k*-mer size 6 as the best aggregative statistics for detecting HGT events, a conclusion may help guide further research in this direction.

Acknowledgment

L.C.X. was supported by the Innovation in Cancer Informatics Fund.

References

[1] Doolittle WF. Phylogenetic classification and the universal tree. Science 1999;284:2124–9.

[2] Burrus V, Waldor MK. Shaping bacterial genomes with integrative and conjugative elements. Res Microbiol 2004;155:376–86.

[3] Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol 2005;3:722–32.

[4] Kelly BG, Vespermann A, Bolton DJ. The role of horizontal gene transfers in the evolution of selected foodborne bacterial pathogens. Food Chem Toxicol 2009;47:951–68.

[5] Andersson JO. Lateral gene transfer in eukaryotes. Cell Mol Life Sci 2005;62:1182–97.

[6] Lawrence JG, Hartl DL. Inference of horizontal genetic transfer from molecular data: an approach using the bootstrap. Genetics 1992;131:753–60.

[7] Makarenkov V, Boc A, Xie JX, Peres-Neto P, Lapointe FJ, Legendre P. Weighted bootstrapping: a correction method for assessing the robustness of phylogenetic trees. BMC Evol Biol 2010;10:250.

[8] Huelsenbeck JP, Crandall KA. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu Rev Ecol Systemat 1997;28:437–66.

[9] Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 2003;52:696–704.

[10] Azad RK, Lawrence JG. Use of artificial genomes in assessing methods for atypical gene detection. PLoS Comput Biol 2005;1:e56.

[11] Zhou FF, Olman V, Xu Y. Barcodes for genomes and applications. BMC Bioinf 2008;9:546.

[12] Tsirigos A, Rigoutsos I. A new computational method for the detection of horizontal gene transfer events. Nucleic Acids Res 2005;33:922–33.

[13] Tang KJ, Lu YY, Sun FZ. Background adjusted alignment-free dissimilarity measures improve the detection of horizontal gene transfer. Front Microbiol 2018;9:711.

[14] Bohlin J, Passel MWV, Snipen L, Kristoffersen AB, Ussery D, Hardy SP. Relative entropy differences in bacterial chromosomes, plasmids, phages and genomic islands. BMC Genomics 2012;13:66.

[15] Thompson JD, Higgins DG, Gibson TJ, CLUSTAL W. Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–80.

[16] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32:1792–7.

[17] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2010;25:1754–60.

[18] Kent WJ. BLAT: the BLAST-like alignment tool. Genome Res 2002;12(4):656–64.

[19] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.

[20] Goldsteint T, Studer C, Baraniuk R. A field guide to forward-backward splitting with a fasta implementation. Comput. Sci. 2014.

[21] Domazet- Lošo M, Haubold B. Alignment-free detection of horizontal gene transfers between closely related bacterial genomes. Mob Genet Elem 2011;1:230–5.

[22] Bromberg R, Grishin NV, Otwinowski Z. Phylogeny reconstruction with alignment free method that corrects for horizontal gene transfer. PLoS Comput Biol 2016;12(6):e1004985.

[23] Bonham-Carter O, Steele J, Bastola D. Alignment-free genetic sequence comparison: a review of recent approaches by word analysis. Briefings Bioinf 2014;15:890–905.

[24] Hao BL, Qi J, Wang B. Prokaryotic phylogeny based on complete genomes without sequence alignment. Mod Phys Lett B 2003;17:91–4.

[25] Blaisdell B. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc. Natl. Acad. Sci. U.S.A. 1986;83:5155–9.

[26] Qi J, Luo H, Hao BL. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Res 2004;32:W45–7.

[27] Lu YY, Tang K, Ren J, Fuhrman JA, Waterman MS, Sun F. CAFE: accelerated alignment-free sequence analysis. Nucleic Acids Res 2017;45:W554–9.

[28] Qi J, Wang B, Hao BL. Whole proteome prokaryote phylogeny without sequence alignment: a k-string composition approach. J Mol Evol 2004;58:1–11.

[29] Yang L, Zhang X, Zhu H. Alignment free comparison: similarity distribution between the DNA primary sequences based on the shortest absent word. J Theor Biol 2012;295:125–31.

[30] Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol 2017;18:186.

[31] Madera M, Calmus R, Thiltgen G, Karplus K, Gough J. Improving protein secondary structure prediction using a simple k-mer model. Bioinformatics 2010;26:596–602.

[32] Cong Y, Chan YB, Ragan MA. A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF. Sci Rep 2016;6:30308.

[33] Hao BL, Qi J. Vertical heredity vs. horizontal gene transfer: a challenge to bacterial classification. J Syst Sci Complex 2003;16:307–14.

[34] Zuo GH, Xu Z, Hao BL. Phylogeny and taxonomy of archaea: a comparison of the whole-genome-based CVTree approach with 16S rRNA sequence analysis. Life 2015;5:949–68.

[35] Liu XM, Wan L, Li J, Reinert G, Waterman MS, Sun FZ. New powerful statistics for alignment-free sequence comparison under a pattern transfer model. J Theor Biol 2011;284:106–16.

[36] Torney DC, Burks C, Davison D, Sirotkin KM. Computation of d2: a measure of sequence dissimilarity. Computers and DNA: The Interface Between Computation Science & Nucleic Acid Sequencing Workshop. 1990. p. 109–25.

[37] Kantorovitz MR, Robinson GE, Sinha S. A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics 2007;23:i249–55.

[38] Forêt S, Wilson SR, Burden CJ. Characterizing the D2statistic: word matches in biological sequences. Stat. Appl. Genet. Mol. Biol. 2009;8:1–21.

[39] Forêt S, Kantorovitz MR, Burden CJ. Asymptotic behavior and optimal word size for exact and approximate word matches between random sequences. BMC Bioinf 2006;7:S5–21.

[40] Reinert G, Chew D, Sun FZ, Waterman MS. Alignment-free sequence comparison (I): statistics and power. J Comput Biol 2009;16:1615–34.

[41] Song K, Ren J, Zhai ZY, Liu XM, Deng MH, Sun FZ. Alignment-free sequence comparison based on next generation sequencing reads. J Comput Biol 2013;20:64–79.

[42] Song K, Ren J, Reinert G, Deng MH, Waterman MS, Sun FZ. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. Briefings Bioinf 2014;15:343–53.

[43] Shepp L. Normal functions of normal random variables. SIAM Rev 1964;6:459–60.