Journal of Theoretical Biology 284 (2011) 106–116

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

New powerful statistics for alignment-free sequence comparison under a pattern transfer model

Xuemei Liu [a,b,1], Lin Wan [b,1], Jing Li [b], Gesine Reinert [c], Michael S. Waterman [b,d], Fengzhu Sun [b,d,*]

[a] School of Physics, South China University of Technology, Guangzhou, PR China
[b] Molecular and Computational Biology Program, University of Southern California, Los Angeles, CA 90089, USA
[c] Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK
[d] TNLIST/Department of Automation, Tsinghua University, Beijing, PR China

ARTICLE INFO

Article history:
Received 16 February 2011
Received in revised form
30 May 2011
Accepted 17 June 2011
Available online 25 June 2011

Keywords:
Alignment-free sequence comparison
D_2
Pattern transfer model

ABSTRACT

Alignment-free sequence comparison is widely used for comparing gene regulatory regions and for identifying horizontally transferred genes. Recent studies on the power of a widely used alignment-free comparison statistic D_2 and its variants D_2^* and D_2^s showed that their power approximates a limit smaller than 1 as the sequence length tends to infinity under a pattern transfer model. We develop new alignment-free statistics based on D_2, D_2^* and D_2^s by comparing local sequence pairs and then summing over all the local sequence pairs of certain length. We show that the new statistics are much more powerful than the corresponding statistics and the power tends to 1 as the sequence length tends to infinity under the pattern transfer model.

1. Introduction

Alignment-free sequence comparison is frequently used to compare genomic sequences and, in particular, gene regulatory regions. Gene regulatory regions are generally not highly conserved making alignment-based methods for the identification of gene regulatory regions less efficient (Ivan et al., 2008; Leung and Eisen, 2009). Several alignment-free sequence comparison statistics have been developed for the identification of gene regulatory regions (Kantorovitz et al., 2007b; Koohy et al., 2010; Leung and Eisen, 2009). Alignment-free sequence comparison has a relatively long history starting in the mid-1980s (Blaisdell, 1986); see for example the review in Vinga and Almeida (2003). The earliest and most widely studied alignment-free statistic is D_2, an uncentered correlation between the number of occurrences of k-tuples (or k-grams) between two sequences (Blaisdell, 1886). Several investigators studied the approximate distribution of D_2 for two unrelated independent identically distributed (i.i.d) or Markovian sequences (Burden et al., 2008; Forêt et al., 2009a,b; Forêt et al., 2006; Kantorovitz et al., 2007a; Lippert et al., 2002). These studies are important for defining threshold values for detecting relationships between two sequences when D_2 is used as a statistic. However, it was pointed out in Lippert et al. (2002) that the D_2 statistic is dominated by the stochastic noise in each sequence and is not appropriate for detecting relationships between two sequences. By normalizing D_2 through its mean and variance under the null model, a new statistic $D2z$ (Kantorovitz et al., 2007b) was developed to compare gene regulatory sequences and the $D2z$ statistic was used to identify cis-regulatory modules in Drosophila (Ivan et al., 2008). Although the $D2z$ (Kantorovitz et al., 2007b) statistic improves the performance over D_2, it is still mainly dominated by the variation of each pattern from the background (Reinert et al., 2009; Wan et al., 2010). Several investigators used the frequency of word patterns to study evolutionary relationships among different organisms (Gao and Qi, 2007; Sims et al., 2009; Jun et al., 2010; Wu et al., 2009; Wang et al., 2009; Qi et al., 2004), compared the advantages and disadvantages of alignment-free sequence comparison methods (Dai and Wang, 2008; Wu et al., 2005), and studied the optimal size of the pattern for comparing genomic sequences (Sims et al., 2009; Wu et al., 2005). Recent simulation studies have shown that alignment-free distance measures based on k-tuple frequencies can give even more accurate trees than phylogenetic tree construction methods based on multiple sequence alignment (Dai and Wang, 2008; Yang and Zhang, 2008), in particular, when the species diverged long ago. Dai and Wang (2008) proposed more complex dissimilarity measures between two sequences and showed that alignment-free measures can be a powerful tool for sequence comparison of highly diverged sequences.

Another application of alignment-free sequence comparison involves the identification of horizontally transferred genes between different organisms (Dalevi et al., 2006; Dufraigne et al., 2005; Sandberg et al., 2001; Suzuki et al., 2008). If highly homologous

* Corresponding author at: Molecular and Computational Biology Program, University of Southern California, Los Angeles, CA 90089, USA.
E-mail address: fsun@usc.edu (F. Sun).
[1] Contribute equally to the work.

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.06.020

genes are detected in distantly related species, these genes may have been horizontally transferred from one organism to another. Horizontally transferred genes can be detected by comparing gene trees and species trees (Dalevi et al., 2006; Sandberg et al., 2001). Sandberg et al. (2001) calculated the distribution of k-tuples in fragments of one genome, compared with those of the other genomes, and assigned the chosen fragment to the species with k-tuple frequency most similar to that of the chosen fragment. Using a similar idea, Dalevi et al. (2006) proposed a Bayesian approach to study horizontal gene transfer among different species. Similar approaches have also been used to study horizontal gene transfer among different environments (Hooper et al., 2008, 2009). It was shown that viruses and their hosts tend to have similar k-tuple distributions (Suzuki et al., 2008). Suzuki et al. (2008) used Mahalanobis distance to study potential horizontal gene transfer between plasmids and their hosts and showed that Mahalanobis distance outperforms the δ-distance, i.e. the average absolute dinucleotide relative abundance difference, in identifying the hosts of the plasmids. Wu et al. (1997) showed that the Mahalanobis distance outperforms Euclidean and standardized Euclidean distance in identifying homologous proteins. However, the computation of Mahalanobis distance can be computationally challenging for $k>5$. Using k-tuple distributions, investigators have also shown that microbial organisms transfer genetic material from one location to the other (Hooper et al., 2008). Furthermore, it has recently been observed that genomic DNA can transfer from donor cells to host cells (Ehnfors et al., 2009; Waterhouse et al., 2011).

Despite the large amount of applications for alignment-free sequence comparison, it is not clear, given an evolutionary scenario, which of the various statistics is most powerful for detecting relationships between the two sequences. We recently carried out a statistical power study of D_2 and its two variations D_2^* and D_2^s. These three statistics are defined as follows. Let $X_{\mathbf{w}}$ and $Y_{\mathbf{w}}$ be the numbers of occurrences of word $\mathbf{w}$ of length k in the first and the second sequences of letters from an alphabet $\mathcal{A}$, respectively. Here both sequences are assumed to have the same length n for simplicity. The D_2 statistic is defined as

$$D_2 \equiv \sum_{\mathbf{w}\in\mathcal{A}^k} X_{\mathbf{w}}Y_{\mathbf{w}}.$$

To define D_2^* and D_2^s as in Reinert et al. (2009) and Wan et al. (2010), we first introduce the centralized count variables by

$$\tilde{X}_{\mathbf{w}} = X_{\mathbf{w}} - (n-k+1)p_{\mathbf{w}} \quad \text{and} \quad \tilde{Y}_{\mathbf{w}} = Y_{\mathbf{w}} - (n-k+1)p_{\mathbf{w}},$$

where $p_{\mathbf{w}}$ is the probability of word $\mathbf{w}$ under the null model. Then D_2^* and D_2^s are defined by

$$D_2^* = \sum_{\mathbf{w}\in\mathcal{A}^k} \frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{(n-k+1)p_{\mathbf{w}}} \quad \text{and} \quad D_2^s = \sum_{\mathbf{w}\in\mathcal{A}^k} \frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{\sqrt{\tilde{X}_{\mathbf{w}}^2+\tilde{Y}_{\mathbf{w}}^2}}.$$

Here we set $\frac{0}{0}=0$.

We studied the power of D_2, D_2^*, and D_2^s under two different models (Reinert et al., 2009; Wan et al., 2010): a common motif model, where the two sequences are related by sharing random instances of a common motif, and a pattern transfer model, where random fragments in the first sequence are transferred to the second sequence, rendering the two sequences related. The pattern transfer model tries to simulate a simple horizontal gene transfer between different species, and it may model sequences with a distant common ancestor which have preserved limited sequence patterns. In particular, we would expect to see that some regions have high mutation rate and some regions are highly conserved. Although the statistical power of both D_2^* and D_2^s increases with the sequence length and tends to 1 as the sequence length tends to infinity under a common motif model, their power approaches a limit which is generally smaller than 1 as the sequence length tends to infinity (Reinert et al., 2009; Wan et al., 2010) under the pattern transfer model. Thus, D_2^* and D_2^s are not ideal for detecting relationships between two sequences under the pattern transfer model.

The objective of this study is to provide new powerful alignment-free statistics for comparing two sequences related through the pattern transfer model. We show through simulations that the new statistics are generally much more powerful than both D_2^* and D_2^s in this setting and their power can approach 1 as the sequence length tends to infinity.

The organization of the paper is as follows. In the "Methods" section, the pattern transfer model that was originally proposed in Reinert et al. (2009) is introduced. Second, the basic ideas behind the new statistics are presented and clear definitions of the new statistics are given. Third, computational issues involved in the calculation of the new statistics are discussed. Fourth, simulation methods to evaluate the power of the new statistics are described and the statistics are used to analyze HIV-1 sequence data. In the "Results" section, we compare the power of the new statistics with the global statistics D_2^* and D_2^s and study the effect of tuple length, window size, shift size (which are defined in the "Methods" section), and evolutionary time after pattern transfer on the power of the new statistics. We also used the new statistics together with D_2^* and D_2^s to analyze HIV-1 sequences. We show that the association between the new statistics and sequence alignment similarity is much higher than the association of D_2^* or D_2^s with sequence alignment similarity when the sequence alignment similarity is around 80%. On the other hand, data show a trend that the converse holds when the alignment similarity is above 83%; however, the difference is not statistically significant. The paper concludes with some discussion and directions for future studies.

2. Methods: the pattern transfer model, new statistics, and simulations

2.1. Introduction of the pattern transfer model

The pattern transfer model was first introduced in Reinert et al. (2009) and a hidden Markov model for it was later developed (Wan et al., 2010). For completeness, we briefly describe the model here. In the pattern transfer model, we randomly choose subsequences of length k_0 from the first sequence and use them to replace corresponding word patterns in the second sequence. More precisely, two independent sequences $\mathbf{A}=A_1A_2\cdots A_n$ and $\mathbf{B}^{(0)}=B_1^{(0)}B_2^{(0)}\cdots B_n^{(0)}$ are initially generated according to the i.i.d model with given nucleotide frequencies (p_A,p_C,p_G,p_T). Then Bernoulli random variables $Z_1,Z_2,\ldots,$ with $P(Z_i=1)=1-\lambda$ are generated for $i=1,2,\ldots,n-k_0+1$. When $Z_i=1$, the k_0-word pattern $A_iA_{i+1}\ldots A_{i+k_0-1}$ in sequence $\mathbf{A}$ is chosen and it then *replaces* $B_iB_{i+1}\ldots B_{i+k_0-1}$ in sequence $\mathbf{B}^{(0)}$. The values $Z_{i+1},\ldots,Z_{i+k_0-1}$ are then ignored and for $j>i+k_0-1$, if $Z_j=1$, the k_0-word pattern occurring at position j in sequence $\mathbf{A}$ again replaces the k_0-word pattern occurring at position j in sequence $B^{(0)}$, and so on. The resulting second sequence is denoted as $\mathbf{B}=B_1B_2\cdots B_n$. We refer the two sequences $\mathbf{A}$ and $\mathbf{B}$ as related through the pattern transfer model. In the case $\lambda=1$ no pattern transfer takes place; this case is our null model.

Throughout this paper, we consider the pattern transfer model and refer to this model as the alternative model when $\lambda<1$.

2.2. New statistics for comparing two sequences related by the pattern transfer model

In Reinert et al. (2009) and Wan et al. (2010), we first showed by simulations and then theoretically proved that the power of all the three statistics D_2, D_2^*, and D_2^s is low and approaches a limit less than 1 when the sequence length tends to infinity. The

primary reason for the relatively low power of D_2^* and D_2^s for detecting the relationships between two sequences under the pattern transfer model is that the means of $X_{\mathbf{w}}$ and $Y_{\mathbf{w}}$ are equal, namely $p_{\mathbf{w}}$, even under the alternative model, where $p_{\mathbf{w}}$ is the probability of word pattern $\mathbf{w}$ under the null model. It was shown in Wan et al. (2010) that D_2^* and $D_2^s/\sqrt{n}$ converge under both the null and the alternative models. Denote

$$\lim_{n\to\infty} D_2^* = \tilde{Z}_\lambda^*, \quad \lim_{n\to\infty} D_2^s/\sqrt{n} = \tilde{Z}_\lambda^s.$$

Note that although D_2^* and D_2^s depend only on the sequences to be compared, their distributions and in turn their limit distributions depend on the models for the sequences to be compared. In our case, their limit distributions depend on λ and we index their limit distributions by λ. In Theorem 3.3 of Wan et al. (2010), it was shown that the asymptotic power of D_2^* and D_2^s under the alternative model when $\lambda<1$ is $P\{\tilde{Z}_\lambda^* \geq \tilde{z}_\alpha^*\}$ and $P\{\tilde{Z}_\lambda^s \geq \tilde{z}_\alpha^s\}$, where $\tilde{z}_\alpha^*$ and $\tilde{z}_\alpha^s$ are the upper α quantile of $\tilde{Z}_1^*$ and $\tilde{Z}_1^s$,respectively. As the limiting random variables $\tilde{Z}_1^*$ and $\tilde{Z}_1^s$ are non-trivial, these results showed that the power of D_2^* and D_2^s approaches a limit that is generally less than 1 when sequence length tends to infinity.

In order to derive new statistics to compare sequences related through the pattern transfer model, we note that both $E(\tilde{Z}_\lambda^*)>0$ and $E(\tilde{Z}_\lambda^s)>0$ for $\lambda<1$ while $E(\tilde{Z}_1^*)=E(\tilde{Z}_1^s)=0$ (Wan et al., 2010). So for $\lambda<1$,

$$E(\tilde{Z}_\lambda^*-\tilde{Z}_1^*)>0, \quad E(\tilde{Z}_\lambda^s-\tilde{Z}_1^s)>0.$$

Thus, when the sequence length is relatively large, we should have $E(D_2^*)>0$ and $E(D_2^s)>0$ under the alternative model $\lambda<1$. On the other hand, under the null model $\lambda=1$, $E(D_2^*)=E(D_2^s)=0$. Thus, to test the null hypothesis $H_0:\lambda=1$ versus the alternative hypothesis $H_1:\lambda<1$, we can test if $E(D_2^*)>0$ or $E(D_2^s)>0$. Based on approximating the mean by a sample mean, the idea of the new statistic is to partition the long sequence of length n into consecutive $d=\lfloor n/W\rfloor$ non-overlapping (discrete) subintervals of length W, calculate D_2^* or D_2^s in each subinterval, and denote the corresponding values in the i-th subinterval by $D_2^*(i)$ or $D_2^s(i)$, respectively. We introduce new statistics T^* and T^s as

$$T^* = \sum_{i=1}^{d} D_2^*(i),$$

and

$$T^s = \sum_{i=1}^{d} D_2^s(i).$$

We reject the null hypothesis in favor of the alternative when T^* (or T^s) is large. If we fix the window length W, the power of the statistic will tend to 1 as the sequence length n tends to infinity.

The following two considerations prompt us to further improve the test statistics T^* and T^s defined above. First, in the pattern transfer model, it is assumed that the patterns transferred from one sequence to the other have the same location along the two sequences. This assumption may not be realistic in real data. We refer to the sequence $\mathbf{A}$ from which the patterns originally come as the donor sequence and to the sequence $\mathbf{B}$ to which the patterns are transferred as the acceptor sequence. The transferred patterns may lie anywhere in the acceptor sequence. Thus, rather than comparing at the same location in both sequences, for each subinterval in either sequence, we should compare with all the subintervals along the other sequence. The second consideration is that we should find segments in the other sequence that are most similar to the segment of interest. Thus, we take the maximum of the test statistics across all the subintervals in the other sequence. Based on these two considerations, we describe the final test statistics as follows.

Consider two sequences of length n, $\mathbf{A}=A_1A_2\ldots A_n$ and $\mathbf{B}=B_1B_2\ldots B_n$. We compare the subintervals of length W from one sequence to that in the other sequence. For each pair of positions $i,j\in[1,n-W+1]$, we compare the subinterval of length W starting at i in sequence $\mathbf{A}$ and the subinterval starting at j in sequence $\mathbf{B}$, $\mathbf{A}[i,i+W-1]=A_iA_{i+1}\cdots A_{i+W-1}$ and $\mathbf{B}[j,j+W-1]=B_jB_{j+1}\cdots B_{j+W-1}$, using D_2^*. Let

$$M^*[i,j,W]=D_2^*(\mathbf{A}[i,i+W-1],\mathbf{B}[j,j+W-1]), \tag{1}$$

and

$$X_i^* = \max_{1\leq j\leq n-W+1} M^*[i,j,W], \quad Y_j^* = \max_{1\leq i\leq n-W+1} M^*[i,j,W].$$

The final statistic we will use to detect the relatedness of the two sequences is

$$T_{\text{sum}}^* = \sum_{i=1}^{n-W+1} X_i^* + \sum_{j=1}^{n-W+1} Y_j^*.$$

The statistic T_{sum}^s can be similarly defined as T_{sum}^* by replacing all the $*$'s in the superscript with s.

Due to the definition of X_i^* and Y_j^* (X_i^s and Y_j^s), they all depend on each other. Theoretical studies of the power of T_{sum}^* and T_{sum}^s are difficult. Thus, we resort to simulations to study their power.

2.3. Computational issues related to the calculation of T_{sum}^* and T_{sum}^s

It is computationally expensive to calculate T_{sum}^* and T_{sum}^s. There are $(n-W+1)^2$ choices of pairs of subintervals of length W. In each pair of subintervals, the number of occurrences of each word $\mathbf{w}$ in each subinterval needs to be counted, and 4^k multiplications and 4^k-1 summations are needed to calculate $M^*[i,j,W]$ and $M^s[i,j,W]$. Then the values of $M^*[i,j,W]$ and $M^s[i,j,W]$ need to be sorted with respect to i and j, respectively. Thus, the number of operations can be huge even for moderate values of sequence length n.

To overcome the computational problems, two procedures were implemented to reduce the computational time. First, we do not compare all pairs of subintervals. Instead, for interval $[i,i+W-1]$ in the donor sequence, we only compare with subintervals $[k\times S+1,k\times S+W], k=0,1,2,\ldots,\lfloor(n-W+1)/S\rfloor$ in the acceptor sequence. Similarly, we only compare subinterval $[j,j+W-1]$ in the acceptor sequence with subintervals $[k\times S+1,k\times S+W]$, $k=0,1,2,\ldots,\lfloor(n-W+1)/S\rfloor$ in the donor sequence. We refer to S as the shift size because we shift the comparison window by size S. For simplicity, we still use X_i^* (X_i^s) and Y_i^* (Y_i^s) to denote the maximum over these subsets of intervals. Second, we calculate the vector consisting of the number of occurrences for each pattern across all the intervals using the following recursive formula,

$$N_{\mathbf{w}}[i+1,i+W]=N_{\mathbf{w}}[i,i+W-1]-I(A_i\cdots A_i+k-1=\mathbf{w}) +I(A_{i+W-k+1}\cdots A_{i+W}=\mathbf{w}),$$

where $I(\cdot)$ is the indicator function with $I(E)=1$ if event E happens and $I(E)=0$, otherwise. Using this recursive formula, we are able to calculate the pattern occurrence vector for any interval of length W linearly in terms of sequence length.

We note that the values of the statistics T_{sum}^* and its counterpart T_{sum}^s depend on the lengths and the nucleotide frequencies of the sequences to be compared. The range of the statistics can be very large when the sequences are long. Thus, their values cannot be used to indicate how closely two sequences are related. We suggest using the corresponding Monte-Carlo p-values to indicate the strength of relatedness between two sequences. However, computing the Monte-Carlo p-values can be time consuming because many random sequence pairs need to be generated. Another difficulty when calculating the Monte-Carlo p-values is

the choice of appropriate random sequence models. An alternative strategy is to re-normalize the D_2^* and D_2^s to C_2^* and C_2^s in the definition of T_{sum}^* and its counterpart T_{sum}^s, respectively, where

$$C_2^* = \frac{(n-k+1)D_2^*}{\sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k}\tilde{X}_{\mathbf{w}}^2/p_{\mathbf{w}}}\sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k}\tilde{Y}_{\mathbf{w}}^2/p_{\mathbf{w}}}},$$

and

$$C_2^s = \frac{D_2^s}{\sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k}\tilde{X}_{\mathbf{w}}^2/\sqrt{\tilde{X}_{\mathbf{w}}^2+\tilde{Y}_{\mathbf{w}}^2}}\sqrt{\sum_{\mathbf{w}\in\mathcal{A}^k}\tilde{Y}_{\mathbf{w}}^2/\sqrt{\tilde{X}_{\mathbf{w}}^2+\tilde{Y}_{\mathbf{w}}^2}}}.$$

Note that the ranges of both C_2^* and C_2^s are from -1 to 1. When the values of C_2^* and C_2^s are close to 1, the sequences are closely related. Similar to the procedures for defining T_{sum}^*, we can define a new corresponding statistic $\widehat{R}_{sum}^*$ by changing D's to C's in the definition. Since $\widehat{R}_{sum}^*$ has a range $[-2\lfloor(n-W+1)/S\rfloor, 2\lfloor(n-W+1)/S\rfloor]$, where W is the window size and S is the shift size, we can normalize $\widehat{R}_{sum}^*$ by $2\lfloor(n-W+1)/S\rfloor$ so that

$$R_{sum}^* = \frac{\widehat{R}_{sum}^*}{2\lfloor(n-W+1)/S\rfloor},$$

which has a range between -1 and 1. Thus, R_{sum}^* can be used to measure the strengths of relatedness through the pattern transfer model. We can similarly define R_{sum}^s. These new statistics can potentially be used to study the relationships among groups of sequences without calculating the p-values resulting in significant saving of computational time. For simplicity, for the power studies we focus on the T-statistics. When analyzing real data, we shall study the R-statistics instead, to assess their relationship with sequence alignment similarity.

2.4. Simulation studies

To compare the power of the new statistics T_{sum}^* and T_{sum}^s with the corresponding global statistics D_2^* and D_2^s, we resort to simulations.

2.4.1. The effect of tuple length, window size, and shift size on the power of the statistics

We first use simulations to study the effects of tuple length k, window size W, and shift size S on the power of T_{sum}^* and T_{sum}^s. The simulation is similar to that in Reinert et al. (2009) and only a brief description of the simulation approaches is given below. Two nucleotide frequencies are considered: the uniform model with $p_a = 1/4, a = A,C,G,T$, and the GC-rich model with $p_C = p_G = 1/3$, $p_A = p_T = 1/6$. The following three steps are used to find the threshold values for the corresponding statistics T_{sum}^* and T_{sum}^s for a given type I error α. First, ten thousand pairs of independent identically distributed (i.i.d) sequences of length $n = 2^j \times 10^2$, $j = 1,2,\ldots,7$ for each fixed j, are generated. Second, for each combination of tuple length $k=2$, 3, 4, 5, 6, window length $W=400$, 800, 1600, and shift size $S=400$, 800, 1600 ($W \geq S$), the statistics T_{sum}^* and T_{sum}^s are calculated for each pair of sequences and their values are sorted in descending order. Third, for a given type I error α, the top 100α quantile is identified and denoted as $t_{sum}^*(\alpha,n,k,W,S)$ and $t_{sum}^s(\alpha,n,k,W,S)$, respectively.

We next approximate the power of T_{sum}^* and T_{sum}^s using simulations. In our simulation study, we set the length of pattern to be transferred as $k_0=5$ and the probability that a transferred pattern starts at any position $1-\lambda = 0.05$. These parameters are the same as in Reinert et al. (2009) for the pattern transfer model. For these parameters, the power of both D_2^* and D_2^s when $k=5$ tends to a limit less than 0.6. The following three steps are used to simulate the power. First, one thousand pairs of related sequences are generated for different sequence lengths $n = 2^j \times 10^2$, $j = 1,\ldots,7$ through the pattern transfer model described in Section 2.1. Second, the values of T_{sum}^* and T_{sum}^s for different combinations of (n,k,W,S) are calculated for each pair of sequences. Third, the power p is approximated by the fraction of times $\hat{p}$ that $T_{sum}^* \geq t_{sum}^*(\alpha,n,k,W,S)$ (or $T_{sum}^s \geq t_{sum}^s(\alpha,n,k,W,S)$). The standard deviation of the estimated power is $\sqrt{p(1-p)/1000} \leq 0.016$. Thus, the 95% confidence interval of p is at most $(\hat{p}-0.03, \hat{p}+0.03)$.

2.4.2. The effect of evolutionary time after pattern transfer on the power of the statistics

In the above simulations, we study the power of the statistics to detect the relationship between two sequences immediately after the pattern transfer. In many situations, the two sequences evolve after the pattern transfer. It is important to understand how evolutionary time affects the power of the different statistics. The following simulation strategy is used to answer this question. First, two sequences are generated as in Section 2.4.1 using the GC-rich model and $1-\lambda = 0.05$. Second, we evolve one of the sequences using the HKY model (Hasegawa et al., 1985) with transition to transversion ratio equal to 2.0. Let $\theta = t\mu$ where t is the evolutionary time and μ is the mutation rate. Note that t and μ are confounded in the HKY model and the evolved sequence depends only on θ. The values of the statistics for the original sequence and the evolved sequences are calculated with $W=S=400$ and $k=5$. Third, we repeat the first and the second steps 1000 times and the power is approximated by the fraction of times that the value of the statistic is equal to or larger than the corresponding threshold values found in Section 2.4.1. We repeat the above steps for different values of $\theta = 0.01-0.10$, and sequence lengths 2000 and 5600.

2.4.3. The power of the statistics based on Drosophila intergenic sequences

The above two simulation strategies generate i.i.d sequences through pattern transfer with/without evolution. For many genomic sequences, the i.i.d model may not fit the sequence data well and the Markov models may fit the sequence better. Under this scenario, the statistics D_2^*, D_2^s, T_{sum}^*, and T_{sum}^s need to be slightly modified by estimating $p_{\mathbf{w}}$ using the Markov model instead of the i.i.d model. Here we use the first order Markov model for the analysis. Higher order Markov models do not increase the performance of the statistics for our data set (results not shown). To see the power of these statistics for comparing sequences related through pattern transfer when the original sequences are from genome sequences, we use the following simulations. First, we download all the intergenic sequences (dmel-all-intergenic-r5.35.fasta) of the Drosophila genome from FlyBase (http://flybase.org/). Second, we randomly select 5000 sequence pairs of the same length L from dmel-all-intergenic-r5.35.fasta and calculate the corresponding statistics. These values are used to approximate the background distribution of the statistics and corresponding threshold values for type I error $\alpha = 0.05$ are obtained. To study the power of the statistics, we choose another set of 5000 sequence pairs as above and transfer segments of one sequence to the other as in our pattern transfer model with $k=5$ and $1-\lambda = 0.05$. The power is approximated by the fraction of times that the value of the statistic is equal to or higher than the corresponding threshold value.

2.4.4. Applications of the statistics to the analysis of HIV-1 sequences

We use the statistics C_2^*, C_2^s, R_{sum}^*, and R_{sum}^s to analyze 42 pure type HIV-1 sequences from Wu et al. (2007). The lengths of the HIV-1 strain sequences are in the range of 9–10 kbp. Our objective is to see which of the four statistics are most appropriate to analyze the data under what conditions. To achieve

this objective, we study the correlation of the values of the various statistics with the sequence similarity based on sequence alignment, for different ranges of sequence similarity. The sequence similarity is defined as the percentage of matches between the two sequences over the reported aligned region (including any gaps in the length). The similarity scores of two sequences are directly calculated by the "needle" program. The needle program is a pairwise sequence global alignment tool based on the Needleman–Wunsch method from the EMBOSS (version 6.3.1) software suite (Rice et al., 2009).

3. Results

For two given sequences, the statistics T^*_{sum} and T^s_{sum} depend on the following parameters: the length of the tuple k, the window size W, and the size of the shift S. We first investigate whether there are parameter regions where the new statistics can be more powerful than the original global statistics. We then study how the power of the statistics depends on these parameters. As in our previous studies, we consider two models: the uniform model with $p_A = p_C = p_G = p_T = 1/4$ and a GC-rich model with $p_A = p_T = 1/6$, $p_G = p_C = 1/3$. We present the results for the uniform and the GC-rich models separately.

Fig. 1(a) and (b) compare the power of the new statistic T^*_{sum} with the original global D^*_2 statistic for the (a) uniform and (b) GC-rich models, respectively, when $k=5$ for different values of window size (W)=shift (S)=(400, 800, 1600) as a function of sequence length n. Fig. 1(c) and (d) show the corresponding figures for T^s_{sum} and D^s_2. It can be seen from these figures that the power of the new statistics T^*_{sum} and T^s_{sum} is much higher than that of the corresponding global statistics D^*_2 and D^s_2, respectively. The power of both D^*_2 and D^s_2 increases slowly when sequence length n is relatively short (<2 kb) and stabilizes at their corresponding limits less than 0.60 when sequence length is above 2000. The power of the new statistics T^*_{sum} and T^s_{sum} for all the three situations increases to 1 as sequence length tends to infinity and is much higher than 0.60. We note also that throughout T^*_{sum} is more powerful than T^s_{sum}, and the statistics are more powerful in the uniform setting than in the GC-rich setting. Given the promising results, we next study the effects of tuple length k, window length W, and shift length S on the power of T^*_{sum} and T^s_{sum}.

3.1. The effects of tuple size k on the power of T^*_{sum} and T^s_{sum}

In our simulations, the length of the transferred pattern k_0 is 5. Intuitively, if a statistic is reasonable, we would expect that the tuple size k achieving the highest power should also be 5. Our previous studies (Reinert et al., 2009; Wan et al., 2010) showed that the power of global D^*_2 increases with tuple size k (up to $k=10$). Here we study the power of T^*_{sum} and T^s_{sum} as a function of k for different combinations of window sizes and shift sizes. Fig. 2(a) and (b) show the power of T^*_{sum} as a function of tuple size when window size=shift size=400 for the (a) uniform and (b) GC-rich models, respectively. Fig. 2(c) and (d) give the corresponding figures for T^s_{sum}. The same set of figures for other combinations of window size (W) and shift size (S) are given in the supplementary material. It can be seen from Fig. 2 and supplementary Figs. 1–5 that the power increases with tuple

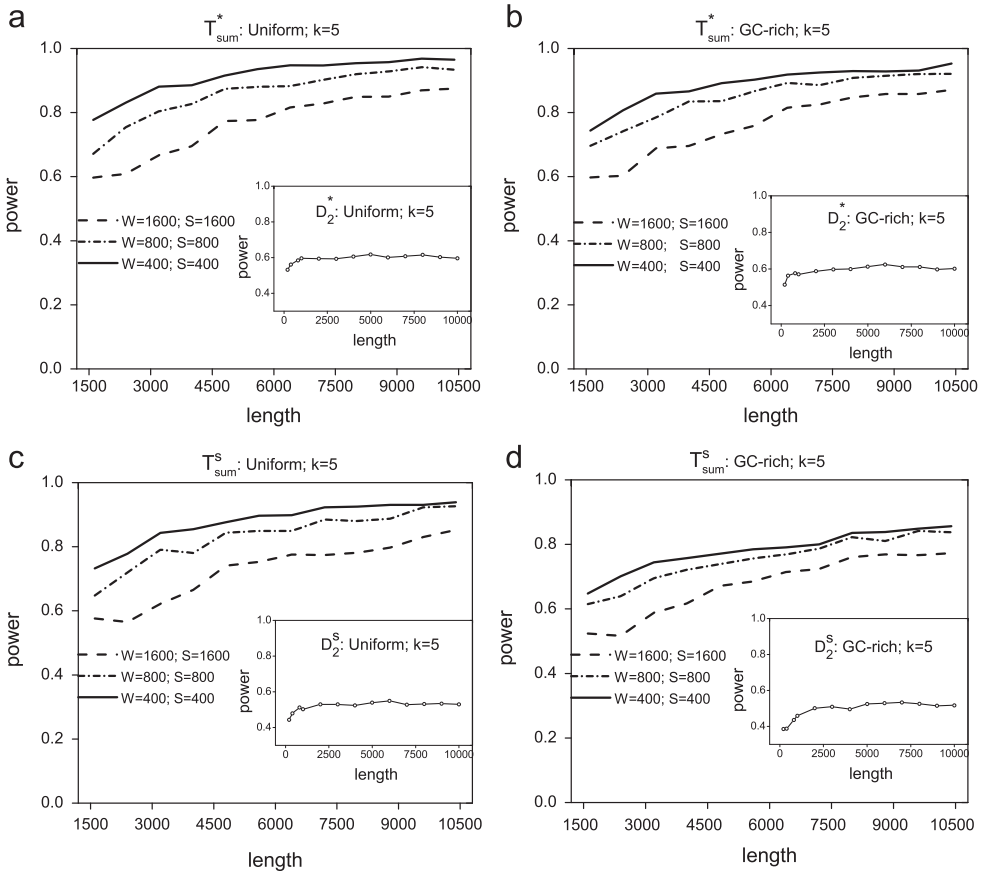

Fig. 1. The power of T^*_{sum} under the (a) uniform and (b) GC-rich models for tuple size $k=5$ and different values of window size (W)=shift size (S)=400, 800, 1600. The same figures for T^s_{sum} are given in (c) and (d), respectively.

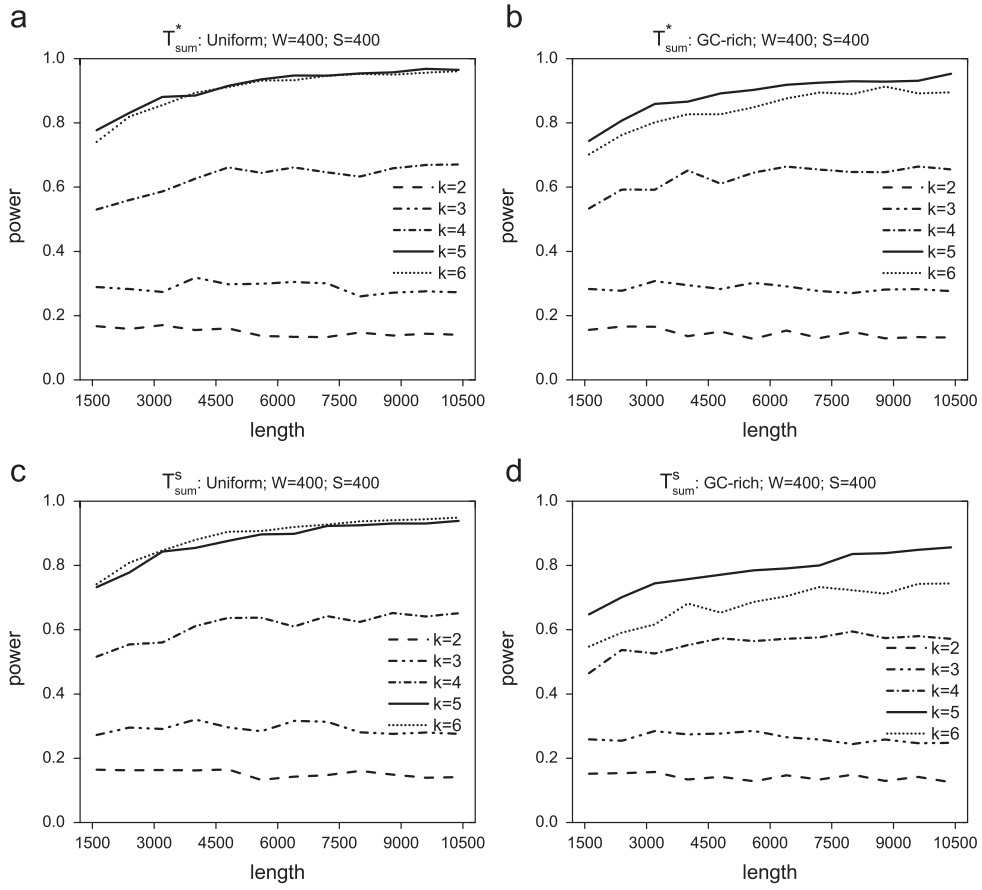

Fig. 2. The effect of tuple size k on the power of T^*_{sum} under the (a) uniform and (b) GC-rich models when window size (W)=shift size (S)=400. The same figures for T^s_{sum} are given in (c) and (d), respectively.

length k when $k \leq 5$ and the power for $k=6$ is slightly smaller than the power for $k=5$ for both T^*_{sum} and T^s_{sum}. Thus, a slight increase of the tuple size above the optimal size will not have a significant effect on the power of the statistics. However, using a smaller tuple size may greatly reduce the power.

3.2. The effects of shift size S on the power of T^*_{sum} and T^s_{sum}

As shown above, the optimal tuple size k which gives the highest power is related to the size of pattern to be transferred. In our simulations, we let the size of the transferred patterns be 5, and we fix the tuple size $k=5$ when we study the effect of window size and shift size. Based on the definitions of T^*_{sum} and T^s_{sum}, we let the shift size (S) be less than or equal to the window size (W) so that the whole sequence is covered. We expect that, for fixed window size, the power of the test is a decreasing function of shift size. The smaller the shift size (S), the higher the power of the statistics is. Fig. 3(a) and (b) show the power of T^*_{sum} for shift size $S=400$, 800, 1600 with tuple size $k=5$ and window size $W=1600$ for the (a) uniform and (b) GC-rich models, respectively. Fig. 3(c) and (d) show the corresponding results for T^*_{sum}. The results when $k=5$ and $W=800$ with shift size $S=400$, 800 are given in the supplementary material. Fig. 3 and supplementary Fig. 6 confirm our original hypothesis that the smaller the shift size, the higher the power of the test. It should also be noted that the power with shift size $S=800$ is similar to that with shift size $S=400$. This may be caused by the overlaps between the shifted intervals.

3.3. The effects of window size W on the power of T^*_{sum} and T^s_{sum}

We next study the effect of window size W. To study the effect of window size on the power of the tests, we fix the shift size $S=400$ and tuple size $k=5$. We let the window size be $W=400$, 800, 1600. Fig. 4(a) and (b) show the power of T^*_{sum} for different window sizes for the (a) uniform and (b) GC-rich models, respectively. Fig. 4(c) and (d) show the corresponding figures for T^s_{sum}. The results with $k=5$ and shift $S=800$ for window size $W=800$, 1600 are given in the supplementary material. In the range of window sizes considered, the tendency is that the smaller the window size, the higher the power of the test statistics is. We note, however, that particularly when the sequence is short, the difference between window size 400 and 800 is not very pronounced.

3.4. The effect of evolutionary time after pattern transfer on the power of T^*_{sum} and T^s_{sum}

To study the effect of evolutionary time after pattern transfer on the power of the test statistics, we let $W=S=400$ and $k=5$. Different sequence lengths and relative evolutionary rate $\theta=t\mu$ are studied. Fig. 5 shows the change of power with respect to θ when (a) $L=2000$ and (b) $L=5600$, respectively. We continue to see that T^*_{sum} and T^s_{sum} are more powerful than D^*_2 and D^s_2, respectively, across all values of θ. In the simulated situations, the power of T^*_{sum} is generally higher than the power of T^s_{sum}. As expected, the power for all the test statistics decreases as a function of θ. Taking the average human mutation rate of about $\mu=10^{-8}$ per base per generation as an example, when $t=10^6$

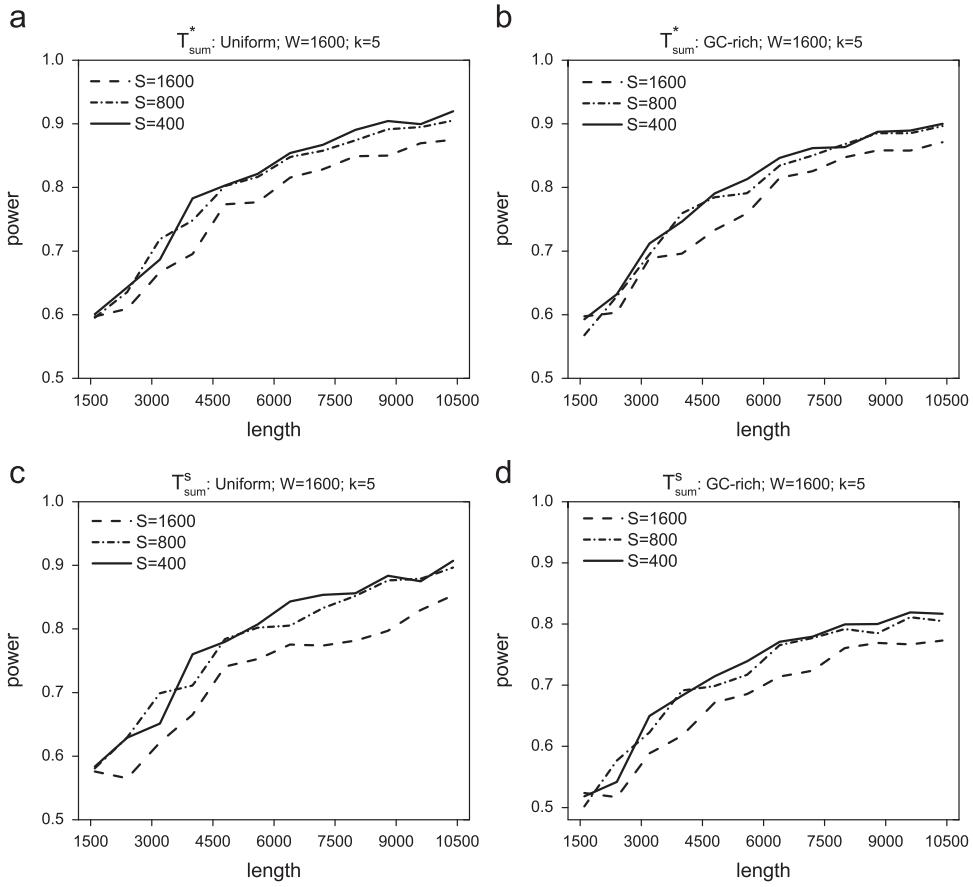

Fig. 3. The effects of shift size on the power of T^*_{sum} under the (a) uniform and (b) GC-rich models with tuple size $k=5$ and window size $W=1600$. The same figures for T^s_{sum} are given in (c) and (d), respectively.

generations, the power of T^*_{sum} decreases from 0.79 to 0.76. On the other hand, with $t=10^7$ generations, the power of T^*_{sum} decreases to 0.41, which is low.

3.5. The power of T^*_{sum} and T^s_{sum} based on real sequence pairs

Fig. 6 shows the power of T^*_{sum} and T^s_{sum} for different window and shift sizes when the probability of transfer $1-\lambda=0.05$ and type I error $\alpha=0.05$. It can be seen from the figure that the power of T^*_{sum} increases when the window size and shift size decrease and is higher than D^*_2. The same conclusions hold for T^s_{sum} and D^s_2. Comparing Fig. 6 with Fig. 1, we can see that the power of the statistics for comparing the real sequences related through pattern transfer is lower than that for the simulated i.i.d background sequences and that the power decreases with sequence length. We also found that the power of all statistics does not increase with the sequence length. A potential explanation for these observations is that the intergenic sequences are heterogeneous and a common Markov model may not fit the sequences well, in particular for long sequences, resulting in low power of our statistics for detecting their relationships. For our real sequences, the power of T^*_{sum} and T^s_{sum} are mostly similar and sometimes the power of T^s_{sum} is higher than the power of T^*_{sum} indicating that T^s_{sum} is more robust to the mis-specification of sequence models.

3.6. The correlation between the various statistics with sequence alignment similarity score

The relationships between the values of the statistics, C^*, C^s, R^*_{sum}, and R^s_{sum}, and sequence alignment similarity scores are shown in Fig. 7. From the figure, it is clear that there is not a linear relationship between sequence alignment similarity and any of the four statistics in the whole range of the similarity score from 58% to 97%. The figure also indicates that neither of the four statistics are highly correlated with sequence alignment similarity when the similarity is less than 78%. Thus, we consider only the region with similarity greater than 78%. The region is divided into three intervals: [78%,83%), [83%,88%), and [88%,1). The Pearson correlation coefficients (together with their 95% confidence intervals) between C^*, C^s, R^*_{sum}, and R^*_{sum} and sequence alignment similarity are given in Table 1. It can be seen that when the sequence alignment similarity is between 78% and 83%, R^*_{sum} and R^s_{sum} are positively correlated with alignment similarity while C^* and C^s do not show a significant association with alignment similarity, indicating the lack of distinguishing power for sequences in this range. When the alignment similarity is between 83% and 88%, the correlation of C^* (C^s) with alignment similarity seems to be higher than the corresponding correlations between R^*_{sum} (R^s_{sum}), however, their 95% confidence intervals overlap indicating that the difference is not statistically significant. The results indicate that, in this region, there are indications that C^* and C^s have better distinguishing power than the corresponding statistics R^*_{sum} and R^s_{sum}, respectively. However, the correlations are only moderate with a range between 44% and 51%. When the sequence alignment similarity score is higher than 88%, all four statistics are highly correlated with sequence alignment similarity with correlation above 0.70 and their 95% confidence intervals overlap. Thus, all of the statistics have similar and high distinguishing power.

We also found in this analysis that although the calculation of R^*_{sum} and R^s_{sum} (window size $W=800$, and shift size $S=800$) is slower than the calculation of C^* and C^s, it is faster than the

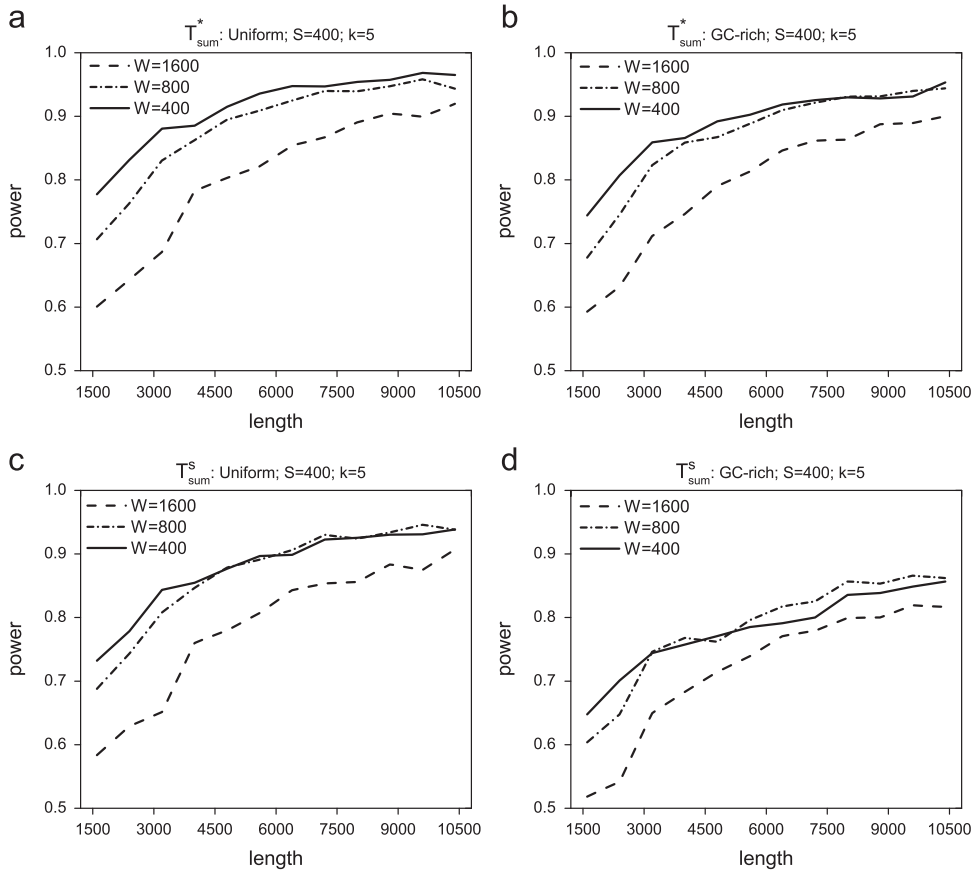

Fig. 4. The effects of window size on the power of T^*_{sum} under the (a) uniform and b) GC-rich models when tuple size $k=5$ and shift size $S=400$. The same figures for T^s_{sum} are given in (c) and (d), respectively.

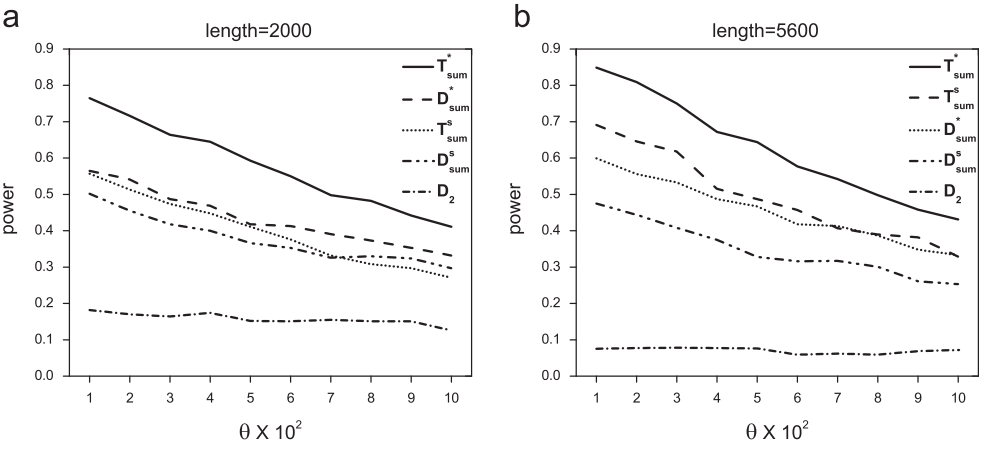

Fig. 5. The effects of evolutionary time on the power of T^*_{sum}, T^s_{sum}, and D^*_2, D^s_2, D_2 when (a) $L=2000$ and (b) $L=5600$ under the GC-rich model with tuple size $k=5$, window size $W=400$, and shift size $S=400$. The pattern transfer probability $1-\lambda=0.05$ and type I error $\alpha=0.05$.

pairwise sequence alignment (the calculation of R^*_{sum} and R^s_{sum} uses less than 1/3 computation time of the needle program).

4. Discussion and conclusions

In this paper, we develop two new statistics T^*_{sum} and T^s_{sum}, plus re-normalized versions, for alignment-free sequence comparison. These statistics depend on three parameters: tuple size k, window size W and shift size S. We show through simulations that the new statistics can be much more powerful than the previous global statistics D^*_2 and D^s_2, respectively, under a pattern transfer model when appropriate parameter values are used. Although under this alternative model the power of D^*_2 and D^s_2 approaches limits that are less than 1 as sequence length tends to infinity, the power of the new statistics T^*_{sum} and T^s_{sum} increases with sequence length and tends to 1 as the sequence length tends to infinity, when choosing a large enough tuple size. Thus, these new statistics are appropriate for detecting relatedness of two sequences under the pattern transfer model.

We then study the effect of the parameters on the power of the statistics T^*_{sum} and T^s_{sum}. We show that the power is highest when

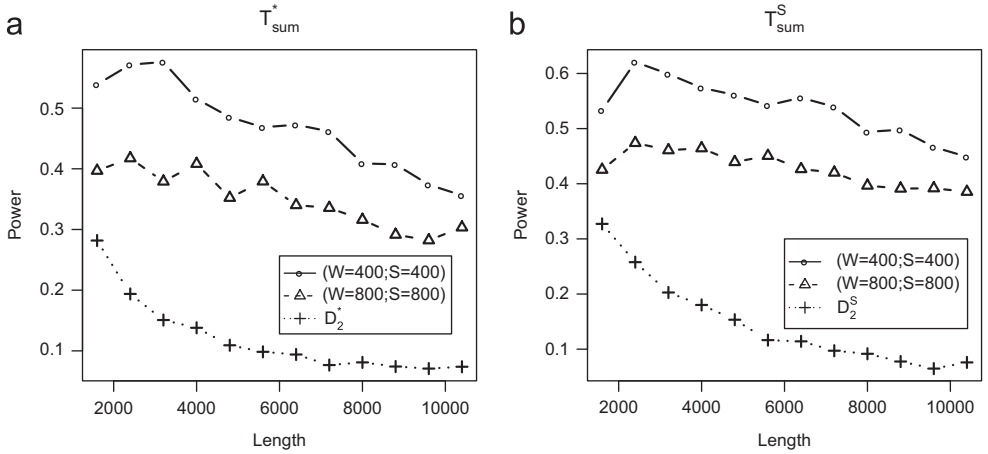

Fig. 6. The power of the statistics: (a) T^*_{sum} and D^*_2, and (b) T^s_{sum} and D^s_2, when comparing two Drosophila intergenic sequences related through pattern transfer for different window and shift sizes, transfer probability $1-\lambda=0.05$, and type I error $\alpha=0.05$.

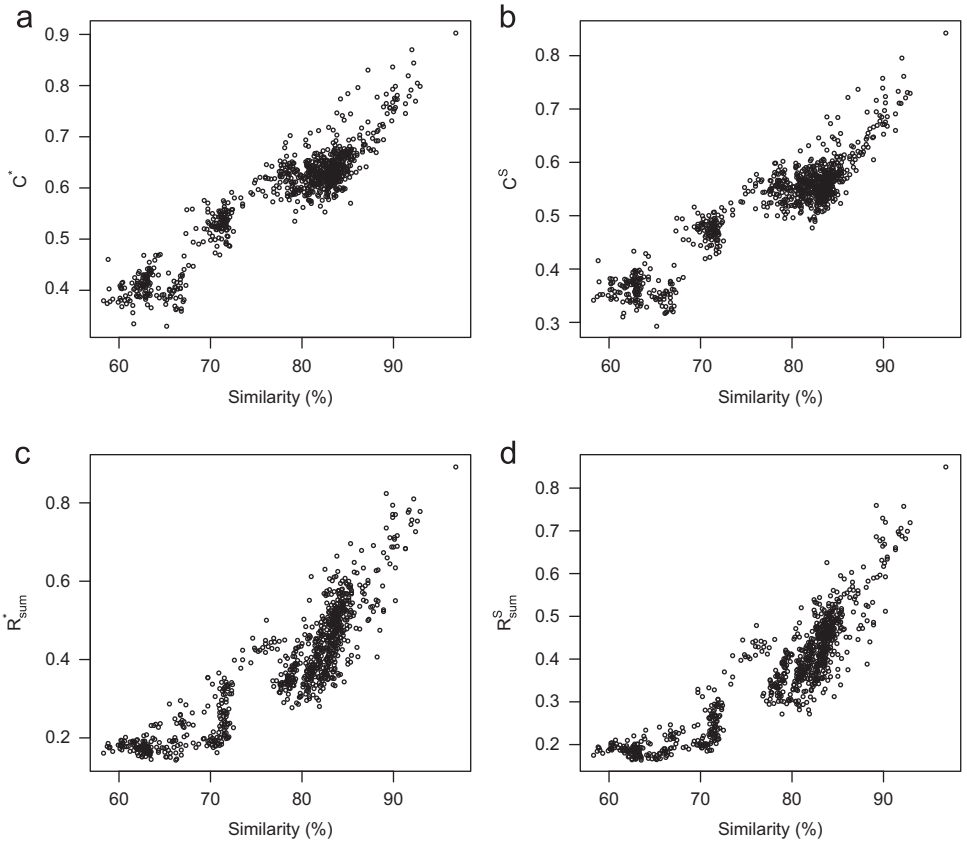

Fig. 7. The relationship between sequence alignment similarity score with (a) C^*, (b) C^s, (c) R^*_{sum}, and (d) R^s_{sum}.

the tuple length used in the statistics is the same as the length of the patterns being transferred and decreases with shift size S. We conjecture that when the window size is too small, the power of T^*_{sum} and T^s_{sum} can be low because D^*_2 or D^s_2 will not be able to identify the similarity between pair of subintervals of length W resulting in low power. The optimal window size should depend on the value of λ and there should be an inverse relationship between the optimal window size and λ.

There are several limitations of our study. First, the pattern transfer model is too simplistic to model the relatedness of two sequences that are related through the transfer of genetic material. The rate of pattern transfer $1-\lambda$ is most likely to vary along the genome sequences to be compared and may be sequence dependent. Some regions are more likely to be transferred than other regions. Thus, it is more appropriate to model λ as a random variable instead of a constant. Second, the transferred genomic regions may have some specific characteristics compared to the other regions. Most current studies on horizontal gene transfer consider the transfer of coding regions and we are not aware of studies on the exchange of genetic material for noncoding regions or gene regulatory regions; there may be considerable differences. Third, we assume that the regions being transferred have the same length which clearly is an over-simplification of exchange of genetic material between genomes. Despite these limitations, our study provides evidence that our new statistics can potentially be used to study relationships between genome sequences under the pattern transfer model.

Table 1
The Pearson correlation coefficients (PCC) with their 95% confidence intervals (CI) between the sequence alignment similarity and statistics C^*, C^s, R^*_{sum}, and R^s_{sum} in different intervals of sequence alignment similarity. The PCCs are calculated using the *R* function "cor", and the CIs of PCC are calculated using the *R* function "cor.test".

Sample size (n)	Alignment similarity interval		
Statistic	[78,83)% 258 PCC (CI)	[83,88)% 273 PCC (CI)	[88,100)% 39 PCC (CI)
C^*	0.06 (−0.06, 0.18)	0.51 (0.41, 0.59)	0.79 (0.63,0.88)
C^s	0 (−0.12, 0.13)	0.54 (0.45, 0.62)	0.79 (0.62, 0.88)
R^*_{sum}	0.40 (0.30, 0.50)	0.44 (0.34, 0.53)	0.70 (0.49, 0.83)
R^s_{sum}	0.40 (0.30, 0.50)	0.49 (0.39, 0.57)	0.74 (0.56, 0.86)

Table 2
The power (× 100) of T^*_{sum} and T^s_{sum} for different values of $W=S$ and the corresponding global statistics D^*_2 and D^s_2 under the common motif model for the GC-rich scenario.

Test	$W=S$	Sequence length					
		800	1600	3200	4000	4800	5600
T^*_{sum}	400	52	82	98	100	100	100
T^*_{sum}	800	65	93	98	100	100	100
T^*_{sum}	1600		94	98	100	100	100
D^*_2		65	95	99	100	100	100
T^s_{sum}	400	24	36	49	73	81	92
T^s_{sum}	800	28	43	54	75	83	93
T^s_{sum}	1600		53	59	82	82	95
D^s_2		28	53	69	81	88	96

Throughout the paper, we assume the pattern transfer model. A natural question is whether the new statistics T^*_{sum} and T^s_{sum} are more powerful than the corresponding global statistics D^*_2 and D^s_2, respectively, under the common motif model described in detail in Reinert et al. (2009) and Wan et al. (2010). We carried out a relatively simple simulation study to answer this question. As in Reinert et al. (2009), the inserted pattern is "AGCCA" and the probability that the pattern is inserted at a position is 0.01. Both uniform and GC-rich background models are simulated. Table 2 shows the power of the new statistics T^*_{sum} and T^s_{sum} for different values of $W=S=400$, 800, 1600 and the corresponding global statistics D^*_2 and D^s_2 under the GC-rich model. The results for the uniform model are given in the supplementary material. It can be seen from this table that the global statistics D^*_2 and D^s_2 are generally more powerful than the statistics T^*_{sum} and T^s_{sum}, respectively, under the common motif model. There are significant differences between the common motif model and the pattern transfer model. Under the common motif model, the expected number of word occurrences differs from that under the null model, thus making the power of global statistics close to 1 when sequence length tends to infinity. The local statistics such as T^*_{sum} and T^s_{sum} did not make full use of that information making them less powerful compared to the global statistics under the common motif models. Thus, depending on the real underlying models for the relationships between the sequences, different statistics should be used.

Acknowledgments

We sincerely thank two anonymous reviewers for suggestions that lead to significant improvement of the paper. The research is supported by the National Natural Science Foundation of China 11071146, 60928007 and 60805010 (FZS), US NIH P50 HG 002790 (LW), R21AG032743 and R21HG006199 (FZS), and the Institute for Mathematical Sciences of the National University of Singapore (GR).

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2011.06.020.

References

Blaisdell, B.E., 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. Proceedings of the National Academy of Sciences of the United States of America 83 (14), 5155–5159.
Burden, C.J., Kantorovitz, M.R., Wilson, S.R., 2008. Approximate word matches between two random sequences. Annals of Applied Probability 18 (1), 1–21.
Dai, Q., Wang, T., 2008. Comparison study on k-word statistical measures for protein: from sequence to sequence space. BMC Bioinformatics 9 (1), 394–395.
Dalevi, D., Dubhashi, D., Hermansson, M., 2006. Bayesian classifiers for detecting HGT using fixed and variable order Markov models of genomic signatures. Bioinformatics 22 (5), 517–522.
Dufraigne, C., Fertil, B., Lespinats, S., Giron, A., Deschavanne, P., 2005. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Research 33 (1), e6.
Ehnfors, J., Kost-Alimova, M., Persson, N.L., Bergsmedh, A., Castro, J., et al., 2009. Horizontal transfer of tumor DNA to endothelial cells in vivo. Cell Death and Differentiation 16 (5), 749–757.
Forêt, S., Kantorovitz, M.R., Burden, C.J., 2006. Asymptotic behaviour and optimal word size for exact and approximate word matches between random sequences. BMC Bioinformatics 7 (Suppl 5), S21.
Forêt, S., Wilson, S.R., Burden, C.J., 2009a. Characterizing the D2 statistic: word matches in biological sequences. Statistical Applications in Genetics and Molecular Biology 8 (1), 43.
Forêt, S., Wilson, S.R., Burden, C.J., 2009b. Empirical distribution of k-word matches in biological sequences. Pattern Recognition 42 (4), 539–548.
Gao, L., Qi, J., 2007. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. BMC Evolutionary Biology 7 (1), 41.
Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22 (2), 160–174.
Hooper, S., Mavromatis, K., Kyrpides, N., 2009. Microbial co-habitation and lateral gene transfer: what transposases can tell us. Genome Biology 10 (4), R45.
Hooper, S., Raes, J., Foerstner, K., Harrington, E., Dalevi, D., Bork, P., 2008. A molecular study of microbe transfer between distant environments. PLoS One 3 (7), e2607.
Ivan, A., Halfon, M.S., Sinha, S., 2008. Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs. Genome Biology 9 (1), R22.
Jun, S., Sims, G., Wu, G., Kim, S., 2010. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: an alignment-free method with optimal feature resolution. Proceedings of the National Academy of Sciences of the United States of America 107 (1), 133–138.
Kantorovitz, M.R., Booth, H.S., Burden, C.J., Wilson, S.R., 2007a. Asymptotic behavior of k-word matches between two uniformly distributed sequences. Journal of Applied Probability 44 (3), 788–805.
Kantorovitz, M.R., Robinson, G.E., Sinha, S., 2007b. A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics 23 (13), i249–i255.
Koohy, H., Dyer, N., Reid, J., Koentges, G., Ott, S., 2010. An alignment-free model for comparison of regulatory sequences. Bioinformatics 26 (19), 2391–2397.
Leung, G., Eisen, M., 2009. Identifying cis-regulatory sequences by word profile similarity. PLoS One 4, e6901.
Lippert, R.A., Huang, H.Y., Waterman, M.S., 2002. Distributional regimes for the number of k-word matches between two random sequences. Proceedings of the National Academy of Sciences of the United States of America 100 (13), 13980–13989.
Qi, J., Luo, H., Hao, B., 2004. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. Nucleic Acids Research 32 (Web Server Issue), W45.
Reinert, G., Chew, D., Sun, F., Waterman, M., 2009. Alignment-free sequence comparison (I): statistics and power. Journal of Computational Biology 16 (12), 1615–1634.
Rice, P., Longden, I., Bleasby, A., 2009. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16 (6), 276–277.
Sandberg, R., Winberg, G., Bränden, C., Kaske, A., Ernberg, I., Cöster, J., 2001. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. Genome Research 11 (8), 1404–1409.
Sims, G., Jun, S., Wu, G., Kim, S., 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proceedings of the National Academy of Sciences of the United States of America 106 (8), 2677–2682.

Suzuki, H., Sota, M., Brown, C., Top, E., 2008. Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes. Nucleic Acids Research 36 (22), e147.
Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison—a review. Bioinformatics 19 (4), 513–523.
Wan, L., Reinert, G., Sun, F., Waterman, M.S., 2010. Alignment-free sequence comparison (II): theoretical power of comparison statistics. Journal of Computational Biology 17 (11), 1467–1490.
Wang, H., Xu, Z., Gao, L., Hao, B., 2009. A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evolutionary Biology 9 (1), 195.
Waterhouse, M., Themeli, M., Bertz, H., Zoumbos, N., Finke, J., Spyridonidis, A., 2011. Horizontal DNA transfer from donor to host cells as an alternative mechanism of epithelial chimerism after allogeneic hematopoietic cell transplantation. Biology of Blood and Marrow Transplantation 17 (3), 319–329.
Wu, G., Jun, S., Sims, G., Kim, S., 2009. Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. Proceedings of the National Academy of Sciences of the United States of America 106 (31), 12826–12831.
Wu, T., Burke, J., Davison, D., 1997. A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. Biometrics 53 (4), 1431–1439.
Wu, T., Huang, Y., Li, L., 2005. Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. Bioinformatics 21 (22), 4125–4132.
Wu, X., Cai, Z., Wan, X., Hoang, T., Goebel, R., Lin, G., 2007. Nucleotide composition string selection in HIV-1 subtyping using whole genomes. Bioinformatics 23 (14), 1744–1752.
Yang, K., Zhang, L., 2008. Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction. Nucleic Acids Research 36 (5), e33.