BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 21 2013, pages 2690–2698
doi:10.1093/bioinformatics/btt462

Sequence analysis

Advance Access publication August 29, 2013

Multiple alignment-free sequence comparison

Jie Ren[1], Kai Song[1], Fengzhu Sun[2,3], Minghua Deng[1] and Gesine Reinert[4,*]

[1]School of Mathematics, Peking University, Beijing 100871, PR China, [2]Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089-2910, USA, [3]MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, PR China and [4]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Recently, a range of new statistics have become available for the alignment-free comparison of two sequences based on *k*-tuple word content. Here, we extend these statistics to the simultaneous comparison of more than two sequences. Our suite of statistics contains, first, C_l^* and C_l^S, extensions of statistics for pairwise comparison of the joint *k*-tuple content of all the sequences, and second, $\overline{C_2^*}$, $\overline{C_2^S}$ and $\overline{C_2^{geo}}$, averages of sums of pairwise comparison statistics. The two tasks we consider are, first, to identify sequences that are similar to a set of target sequences, and, second, to measure the similarity within a set of sequences.

Results: Our investigation uses both simulated data as well as *cis*-regulatory module data where the task is to identify *cis*-regulatory modules with similar transcription factor binding sites. We find that although for real data, all of our statistics show a similar performance, on simulated data the Shepp-type statistics are in some instances outperformed by star-type statistics. The multiple alignment-free statistics are more sensitive to contamination in the data than the pairwise average statistics.

Availability: Our implementation of the five statistics is available as R package named 'multiAlignFree' at be http://www-rcf.usc.edu/~fsun/Programs/multiAlignFree/multiAlignFreemain.html.

Contact: reinert@stats.ox.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 18, 2013; revised on July 12, 2013; accepted on August 5, 2013

1 INTRODUCTION

A *cis*-regulatory module (CRM) is a stretch of DNA in charge of regulating expression of some nearby genes (Davidson, 2006). CRMs typically consist of a few hundred base pairs, and they contain multiple transcription factor binding sites (TFBS). As CRMs with similar TFBSs may show similar regulatory patterns, a pertinent question is how to classify CRM sequences and how to measure the similarity within a set of CRM sequences.

Usually, the similarity between two CRM sequences is assessed by either global (Needleman *et al.*, 1970) or local alignment (Smith and Waterman, 1981). A main issue with applying alignment-based methods for this problem is that TFBSs are normally much shorter than the CRM, and hence the alignment score of two CRMs may be dominated by the content of their background sequences. Therefore, a high alignment score between two CRMs may arise without them containing similar TFBSs regulatory patterns. Moreover, especially in distantly related species, the CRM sequences may simply not be alignable (Arunachalam *et al.*, 2010; Wolff *et al.*, 1999). Finally, alignment-based methods are time-consuming. Hence, CRMs provide a challenging test case for sequence comparison; see also Hardison and Taylor (2012).

In contrast, alignment-free methods, first proposed by Blaisdell (1986), have been developed at a fast pace during the past two decades. Here, we focus on alignment-free sequence comparison based on the joint *k*-tuple content of sequences. For two sequences $A = A_1A_2 \dots A_{n_1}$ and $B = B_1B_2 \dots B_{n_2}$ of letters from a finite alphabet $\mathcal{A}$ of size d, and for a *k*-tuple word $w \in \mathcal{A}^k$, let $X_w^{(1)}$ denote the frequency of w in sequence A, and similarly $X_w^{(2)}$ denote its frequency in sequence B. We centralize the word counts based on a probabilistic model. Let $P_w^{(i)}$ denote the probability of occurrence of the word $w = w_1, w_2, \dots, w_k$ in the i^{th} sequence, $i = 1, 2$. For example, when the letters in the underlying sequence are assumed to be drawn *independently* at random from the *identical distribution* (the i.i.d. model), then $P_w^{(i)} = \prod_{j=1}^k p_{w_j}^{(i)}$, where $p_{w_j}^{(i)}$ is the probability of letter w_j in the ith sequence; in practice, $p_{w_j}^{(i)}$ is estimated by the frequency of letter w_j in the ith sequence.

For a word w, the centralized count variables

$$\tilde{X}_w^{(i)} = X_w^{(i)} - (n_i - k + 1)P_w^{(i)}, \quad i = 1, 2, \tag{1}$$

and their approximate standardization,

$$\tilde{X}_w^{*,(i)} = \frac{X_w^{(i)} - (n_i - k + 1)P_w^{(i)}}{\sqrt{(n_i - k + 1)P_i^{(1)}}}, \quad i = 1, 2,$$

measure the deviation of the word frequencies in the sequence from the word occurrence expectation in the so-called background sequences. The statistics D_2^* and D_2^S for alignment-free comparison of two sequences are now defined by

$$D_2^*(A, B) = \sum_{w \in \mathcal{A}^k} \tilde{X}_w^{*,(1)} \tilde{X}_w^{*,(2)}$$

and

$$D_2^S(A, B) = \sum_{w \in \mathcal{A}^k} \frac{\tilde{X}_w^{(1)} \tilde{X}_w^{(2)}}{\sqrt{(\tilde{X}_w^{(1)})^2 + (\tilde{X}_w^{(2)})^2}},$$

see Reinert *et al.* (2009) and Wan *et al.* (2010). In Reinert *et al.* (2009), it has been shown that these statistics are powerful in

*To whom correspondence should be addressed.

distinguishing two sequences related by a common motif. We refer to the first statistic as *star-type* and to the second statistic as *Shepp-type*; according to Shepp (1964), if the counts were independent mean zero normal variables, then the self-standardized statistic D_2^S would again be normally distributed.

To eliminate the effect of the length of the sequences, two renormalized versions, C_2^* and C_2^S, of D_2^* and D_2^S are introduced in Liu *et al.* (2011) by applying the Cauchy–Schwarz inequality;

$$C_2^*(A,B) = \frac{D_2^*(A,B)}{\sqrt{\sum_{w\in\mathcal{A}^k}(\tilde{X}_w^{*,(1)})^2}\sqrt{\sum_{w\in\mathcal{A}^k}(\tilde{X}_w^{*,(2)})^2}}$$

and, with

$$\tilde{X}_w^{S,(i)} = \frac{\tilde{X}_w^{(i)}}{\left((\tilde{X}_w^{(1)})^2 + (\tilde{X}_w^{(2)})^2\right)^{\frac{1}{4}}}$$

for $i = 1, 2$,

$$C_2^S(A,B) = \frac{D_2^S(A,B)}{\sqrt{\sum_{w\in\mathcal{A}^k}(\tilde{X}_w^{S,(1)})^2}\sqrt{\sum_{w\in\mathcal{A}^k}(\tilde{X}_w^{S,(2)})^2}}.$$

Both C_2^* and C_2^S range from -1 to 1, with a value close to 1 when the sequences are closely related. For sequences that do not have the same length, the total number of k-tuple words could have a large influence on the value of D-type statistics, and hence for such sequences, we use the C-type statistics to measure the similarity. As CRM sequences vary by a few hundred base pairs in length, we only use the C-type statistics in our applications.

Variants of our D-type statistics have been considered in the literature; a main issue with them is that they may be dominated by single-sequence noise. In Lippert *et al.* (2002), it has been shown that the straightforward non-centred statistic $D_2(A,B) = \sum_{w\in\mathcal{A}^k} X_w^{(1)} X_w^{(2)}$ may be dominated either by noise in the single sequences or by difference in sequence backgrounds. Thus, it has weak power in detecting relationship between two sequences; see also Kantorovitz *et al.* (2007); Reinert *et al.* (2009); Wan *et al.* (2010). In Kantorovitz *et al.* (2007), a standardized statistics $D2z = \frac{D_2 - \mathrm{E}(D_2)}{sd(D_2)}$ was proposed, where $E(D_2)$ and $sd(D_2)$ are the mean and standard deviation of D_2 under a given model of the sequence. It was shown that $D2z$ outperforms D_2. Yet, the noise in the single sequence may still dominate the statistics $D2z$ when the background sequence is not uniform (Reinert *et al.*, 2009). Standardizing the count vectors themselves, as in D_2^* and D_2^S, resolves this issue.

In the work of Göke *et al.* (2012), the background sequence model is generalized from *i.i.d* to a first-order Markov chain model, showing that a first-order Markov chain model improves the ability of detecting similarity between regulatory sequences compared with the *i.i.d* model. Higher order Markov chain models could be considered, but owing to the limited length of CRM sequences, estimating the higher order Markov chain model parameters leads to an overfitting problem and results in poor performance (Göke *et al.*, 2012). Göke *et al.* (2012) also showed that including the reverse complement of a word could increase the prediction accuracy.

Although the statistics C_2^* and C_2^S perform well for pairwise alignment-free sequence comparison, in applications, the similarity within a set of more than two sequences may be of interest. There are two main types of problems to tackle. The first problem is an *identification problem*. Suppose that we have already classified $\ell - 1$ sequences according to whether they are CRMs of a particular type, and we are presented with a new sequence. How should we classify this sequence? The second problem is how to *assess the similarity within a set of sequences*. Both problems are motivated by the task of predicting sets of related CRMs, given a set of co-regulated genes.

In Kantorovitz *et al.* (2007), the overall similarity within a set of CRM sequences is addressed as follows. A set of CRMs, known to regulate expression in the same tissue and/or development stage, is taken as the so-called *positive set*. A set of equally many randomly chosen sequences, with lengths matching the CRMs, is taken as the so-called *negative set*. The positive set is randomly partitioned into two disjoint subsets of equal size, and each subset of sequences is concatenated to produce one long sequence. The similarity between these sequences can then be assessed using an alignment-free statistic for pairwise sequence comparison; Kantorovitz *et al.* (2007) recommend $D2z$. As this method requires many repeats of the random partitioning, it is time-intensive, and hence alternative methods are of interest.

Here, we extend the renormalized pairwise alignment-free sequence comparison statistics C_2^* and C_2^S to two families of multiple statistics, denoted by C_l^* and C_l^S. We also introduce three families of average pairwise statistics for the identification problem, called $\overline{C_2^*}$, $\overline{C_2^S}$ and $\overline{C_2^{geo}}$, and their versions for measuring similarity within a set of sequences, called $\overline{\overline{C_2^*}}$, $\overline{\overline{C_2^S}}$ and $\overline{\overline{C_2^{geo}}}$.

We evaluate these families of statistics by testing them on the identification problem described earlier in the text, both through a simulation study and through analysing real data, by using each of our statistics as a score; the higher the score, the more similar the sequences should be. In the simulation study, we follow the common motif model in Reinert *et al.* (2009) and Wan *et al.* (2010) but under a first-order Markov chain model for the background sequence. Then, we apply our statistics to the data from Blow *et al.* (2010) of tissue-specific enhancer CRMs identified *in vivo* in mouse embryos. We also evaluate these families for the task of assessing the similarity within sets of sequences, again using both the simulated sequences as mentioned previously and the CRM data from Blow *et al.* (2010).

Our evaluation is based on 90% confidence intervals (CI) for the *area under the receiver operational characteristic* (ROC) *curve (AUC)*, which is implemented by the R package ROCR in Sing *et al.* (2005). In Hardison and Taylor (2012), it is emphasized that it is often the number of false positives that make CRM prediction not very successful. Hence, we also assess the performance of our statistics through 90% CIs for the false-positive rate at 20% sensitivity. We say that a statistic outperforms the other if the 90% CIs of the former statistic is smaller than the lower bound of the 90% CI of the latter statistic.

We find that although for real data all of our statistics show a similar performance, on simulated data, the Shepp-type statistics are in some instances outperformed by star-type statistics. The multiple alignment-free statistics are more sensitive to contamination in the data than the pairwise average statistics.

The article is organized as follows. In Section 2, we introduce two families of multiple alignment-free statistics, called C_l^* and C_l^S, for $l > 2$. Moreover, three families of average pairwise

alignment-free statistics, called $\overline{C_2^*}$, $\overline{C_2^S}$ and $\overline{C_2^{geo}}$, are introduced with the identification problem in mind, as well as their generalizations $\overline{\overline{C_2^*}}$, $\overline{\overline{C_2^S}}$ and $\overline{\overline{C_2^{geo}}}$ for measuring similarity within a set of sequences. Section 3 contains the evaluation and comparison of these statistics for the CRM identification problem. Section 4 addresses the problem of measuring the similarity within a set of sequences. In Section 5, we assess the effect of contamination in the data. We summarize our results in Section 6.

2 METHODS

2.1 Multiple alignment-free statistics

Let $\{A^{(1)}, A^{(2)}, \ldots, A^{(l)}\}$ be a set of sequences, where $A^{(k)}$ has length n_k. Assume that the sequences follow a first order time-homogeneous irreducible aperiodic Markov chain with transition matrix T and stationary distribution π. For a word $w = w_1, w_2, \ldots, w_k$, the probability of word w in the ith sequence is then $P_w^{(i)} = \pi^{(i)}(w_1)\prod_{j=2}^{k} T^{(i)}_{w_{j-1},w_j}$. We extend w to include its reverse complement $\bar{w}$. Then in our definition,

$$\tilde{X}_w^{(i)} = (X_w^{(i)} + X_{\bar{w}}^{(i)}) - (n_i - k + 1)(P_w^{(i)} + P_{\bar{w}}^{(i)}),$$

where $X_w^{(i)}$ is the number of occurrence of word w in the sequence $A^{(i)}$.

Our proposed statistics use the absolute centralized word count variable $|\tilde{X}_w^{(i)}|$; we call these the *absolute* statistics. We define our *absolute* star-type multiple alignment-free sequence comparison statistic by

$$C_l^*(A^{(1)}, \ldots, A^{(l)}) = \frac{\sum_{w\in\mathcal{A}^k}\prod_{i=1}^{l}|\tilde{X}_w^{(i)}|}{\prod_{i=1}^{l}\left\{\sum_{w\in\mathcal{A}^k}\left(|\tilde{X}_w^{(i)}|\right)^l\right\}^{\frac{1}{l}}}. \quad (2)$$

Similarly, following Quine (1994), our *absolute* Shepp-type multiple alignment-free sequence comparison statistic is defined as

$$C_l^S(A^{(1)}, \ldots, A^{(l)}) = \frac{\sum_{w\in\mathcal{A}^k}\prod_{i=1}^{l}\tilde{X}_w^{S,(i)}}{\prod_{i=1}^{l}\left\{\sum_{w\in\mathcal{A}^k}\left(\tilde{X}_w^{S,(i)}\right)^l\right\}^{\frac{1}{l}}}, \quad (3)$$

where $\tilde{X}_w^{S,(i)} = \frac{|\tilde{X}_w^{(i)}|}{\left\{\sum_{j=1}^{l}(\tilde{X}_w^{(-j)})^2\right\}^{\frac{1}{2l}}}$ and

$$\tilde{X}_w^{(-i)} = \tilde{X}_w^{(1)}\ldots\tilde{X}_w^{(i-1)}\tilde{X}_w^{(i+1)}\ldots\tilde{X}_w^{(l)}.$$

For measuring the similarity within a set of sequences, we directly apply these statistics to a set of sequences as a measurement of similarity. Larger values indicate higher similarity among the sequences, and thus higher conservation of regulatory patterns for these sequences. We also use these multiple alignment-free statistics for the CRM identification problem, by measuring the similarity within a set of sequences containing one candidate sequence and the known CRM sets. Higher values of the statistics indicate higher similarity among the candidate sequence and the known CRM sequences, and a higher likelihood that the candidate sequence has a similar regulatory pattern as the given CRMs.

Although it would be tempting to use as extension of D_2^* and D_2^S, the statistics based on the products of the centred counts $\tilde{X}_w^{(i)}$, these extensions do not work well when the number of sequences is odd. Indeed, for pairwise D_2 type statistics, for closely related sequences, the centralized counts should often have the same signs so that the product $\tilde{X}_w^{(1)}\tilde{X}_w^{(2)}$ would often be positive. However, for the multiple alignment-free statistics with an odd number of sequences, when all the centralized counts have a negative sign, their product will give a negative contribution to the similarity score. Therefore, instead of using $\tilde{X}_w^{(i)}$, we use the absolute centralized word count variable $|\tilde{X}_w^{(i)}|$; in the supplementary material, we also report on the statistics corresponding to (2) and (3) using the original centralized word count variables $\tilde{X}_w^{(i)}$ and the squared centralized word count variables $\check{X}_w^{(i)} = (\tilde{X}_w^{(i)})^2$.

2.2 Average pairwise alignment-free statistics

2.2.1 Identification of CRMs Here, we consider what we call the CRM identification problem: Given a set of CRM sequences with similar TFBSs regulatory pattern, how can we identify new CRMs in the genome that belong to a similar regulatory pattern as the given CRMs? Suppose that we have a set of identified CRM sequences $S_0 = \{s_1, s_2, \ldots, s_l\}$ with a similar regulatory pattern, and we have a set $S_1 = \{c_1, c_2, \ldots, c_k\}$ of candidate sequences. We want to pick CRMs from the candidate set that have a similar regulatory pattern as sequences in S_0.

There are two possible approaches to this problem. On the one hand, we can use $C_{l+1}^*(s_1, \ldots, s_l; c_i)$ and $C_{l+1}^S(s_1, \ldots, s_l; c_i)$ as the similarity score. On the other hand, we can measure the similarity by averaging over the pairwise alignment-free statistics D_2^* or D_2^S, respectively; the length-adjusted versions are

$$\overline{C_2^*(s_1, \ldots, s_l; c_i)} = \frac{1}{l}\sum_{j=1}^{l} C_2^*(s_j, c_i) \quad \text{and} \quad (4)$$

$$\overline{C_2^S(s_1, \ldots, s_l; c_i)} = \frac{1}{l}\sum_{j=1}^{l} C_2^S(s_j, c_i). \quad (5)$$

A closer look reveals that $\overline{C_2^*(s_1, \ldots, s_l; c_i)}$ is the inner product between the standardized and normalized k-tuple vector of sequence c_i and the *arithmetic* mean of the standardized and normalized k-tuple vector of sequence $s_1, \ldots, s_l$. By taking the *geometric* mean of the standardized and normalized k-tuple vectors of sequences $s_1, \ldots, s_l$ instead of arithmetic mean, we create another length-adjusted statistic $\overline{C_2^{geo}}(s_1, \ldots, s_l; c_i)$ based on $|\tilde{X}_w^{*,(i)}|$,

$$\overline{C_2^{geo}(s_1, \ldots, s_l; c_i)} = \sum_{w\in\mathcal{A}^k}\frac{|\tilde{X}_w^{*,(c_i)}|}{\left(\sum_{w\in\mathcal{A}^k}\left(|\tilde{X}_w^{*,(c_i)}|\right)^{\frac{l}{l-1}}\right)^{\frac{l-1}{l}}} \times \sum_{w\in\mathcal{A}^k}\left(\prod_{j=1}^{l}\frac{|\tilde{X}_w^{*,(s_j)}|}{\left(\sum_{w\in\mathcal{A}^k}\left(|\tilde{X}_w^{*,(s_j)}|\right)^l\right)^{1/l}}\right)^{\frac{1}{l}}. \quad (6)$$

Instead of taking the geometric mean based on the absolute centralized counts, we could have based it on the squared centralized counts. In simulations, we find that calculating the squared statistics requires more computations than the absolute version without considerable gain in precision, see Supplementary Materials 1.2.

2.2.2 Measuring similarity within sets of sequences Now, we consider the task of assessing the similarity within sets of sequences. For a set $S = \{s_1, s_2, \ldots, s_l\}$, the multiple alignment-free statistics (2)–(3) can be used to measure the similarity within a set of sequences, but the average pairwise statistics need to be extended. Let $S_{-i} = \{s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_l\}$. We define three average pairwise alignment-free statistics

$$\overline{\overline{C_2^*(S)}} = \frac{1}{l}\sum_{i=1}^{l}\overline{C_2^*(S_{-i}; s_i)}, \quad (7)$$

$$\overline{\overline{C_2^S(S)}} = \frac{1}{l}\sum_{i=1}^{l}\overline{C_2^S(S_{-i}; s_i)}, \quad \text{and} \quad (8)$$

$$\overline{\overline{C_2^{geo}(S)}} = \frac{1}{l}\sum_{i=1}^{l}\overline{C_2^{geo}(S_{-i}; s_i)}. \quad (9)$$

2.3 A simple example

We illustrate the methods with a simple example of the following five sequences, (i) 'ATGCATAT', (ii) 'ATATATGC', (iii) 'ATGCATGT', (iv) 'ATATGCAT', (v) 'CTCGGAGA', where the second sequence is the first one with the first and second half transposed; the third sequence is the first one with the second-to-last position changed from 'A' to 'G'; the fourth sequence is the complement of the first one; and the last one has no 2-tuples in common with the other sequences. With $k=2$, and all letters equally likely and independent (so $P_w = 1/16$), the 5 by 5 matrices calculated based on C_2^* and C_2^S are

$$C_2^* = \begin{bmatrix} 1.00 & 0.93 & 0.84 & 1.00 & -0.44 \\ 0.93 & 1.00 & 0.61 & 0.93 & -0.41 \\ 0.84 & 0.61 & 1.00 & 0.84 & -0.52 \\ 1.00 & 0.93 & 0.84 & 1.00 & -0.44 \\ -0.44 & -0.41 & -0.52 & -0.44 & 1.00 \end{bmatrix}$$

and

$$C_2^S = \begin{bmatrix} 1.00 & 0.89 & 0.76 & 1.00 & -0.39 \\ 0.89 & 1.00 & 0.59 & 0.89 & -0.30 \\ 0.76 & 0.59 & 1.00 & 0.76 & -0.43 \\ 1.00 & 0.89 & 0.76 & 1.00 & -0.39 \\ -0.39 & -0.30 & -0.43 & -0.39 & 1.00 \end{bmatrix}.$$

As our statistics include the reverse complement, sequences 1 and 4 have a pairwise similarity of 1. Both of these pairwise statistics consider transposition as closer to the original sequence than the change of one letter; C_2^* gives a higher score than C_2^S for comparisons between the first four sequences, which are similar, and a lower score to comparison with the last sequence, which is different from the other sequences.

We also calculate the multiple statistics for comparing all 10 combinations of three sequences, see Table 1. Note that the values of the statistics cannot be compared across methods, and C_l^*, C_l^S and $\overline{\overline{C_2^{geo}}}$ range between 0 and 1, whereas $\overline{\overline{C_2^*}}$ and $\overline{\overline{C_2^S}}$ and take values in the range between -1 and $+1$. For pairs of triplets that only differ in whether sequence 1 or sequence 4 is included (such as $\{1,2,3\}$ and $\{2,3,4\}$), the similarity scores are the same; for such pairs, we only report the results when 1 is included. The scores for triplets involving both sequence 1 and 4 differ from the pairwise scores; adding the sequence 4 increases all scores except the C_l^S score for $\{1,2,4\}$.

We would expect that the statistics pick up that sequence 5 is dissimilar to sequences 1–4. The pairwise similarity score between the sequences 1–4 and sequence 5 are all negative, indicating the strong dissimilarity between them. Similarly, the triplets including the sequence 5 have significantly smaller values of the similarity score than the others. C_l^* and the three pairwise average statistics give the highest score for $\{1,2,4\}$, whereas C_l^S gives its highest score for $\{1,3,4\}$, by considering the transposition as a more severe change than the change of one letter. All five statistics give the lowest score to $\{2,3,5\}$. The multiple statistics as well as $\overline{\overline{C_2^{geo}}}$ assign the second-lowest scores to $\{1,2,5\}$, whereas the averaging statistics $\overline{\overline{C_2^*}}$ and $\overline{\overline{C_2^S}}$ assign their second-lowest scores to $\{1,3,5\}$.

When replacing sequence 3 to be 'ATACATAT', that is, we replace the third letter 'G' by 'A' instead of replacing the second-to-last letter 'A' by 'G', then all five statistics assign the lowest score to {1,3,5} (data not shown).

Table 1. The multiple statistics for comparing three sequences from the five sequences, (i) 'ATGCATAT', (ii) 'ATATATGC', (iii) 'ATGCATGT', (iv) 'ATATGCAT', (v) 'CTCGGAGA'

Triplets	C_l^*	C_l^S	$\overline{\overline{C_2^*}}$	$\overline{\overline{C_2^S}}$	$\overline{\overline{C_2^{geo}}}$
{1, 2, 3}	0.82	0.60	0.79	0.75	0.83
{1, 2, 4}	0.95	0.80	0.95	0.93	0.94
{1, 2, 5}	0.40	0.49	0.03	0.07	0.65
{1, 3, 4}	0.89	0.81	0.89	0.84	0.92
{1, 3, 5}	0.44	0.52	−0.04	−0.02	0.70
{1, 4, 5}	0.42	0.62	0.04	0.07	0.70
{2, 3, 5}	0.38	0.38	−0.11	−0.05	0.61

3 PERFORMANCE ON CRM IDENTIFICATION

First, we evaluate and compare the proposed statistics (2)–(6) on the CRM identification problem. Given a set of known CRM sequences $S_0 = \{s_1, s_2, \ldots, s_l\}$ that have a similar TFBS regulatory pattern, and a set of candidate sequences $S_1 = \{c_1, c_2, \ldots, c_k\}$, which contains both CRMs and background sequences, we group S_0 with each candidate sequence c_i separately, where $c_i \in S_1$. Then, we find the similarity score for each set of sequences $\{S_0; c_i\}$ and predict that groups with higher score have the higher similarity.

3.1 Simulation study

To simulate the data for the CRM identification problem, we generate two types of sequences: the first type are to resemble CRM sequences, with common binding motifs inserted, and the second type are background sequences, without any motifs. For simplicity, we refer to the first type as *CRM sequences* in this subsection. We generate the simulated sequences following the 'common motif' model proposed in Reinert *et al.* (2009) but under the first-order Markov chain sequence background model instead of an i.i.d model.

Let $\mathcal{A} = \{A, C, G, T\}$ denote the nucleotide alphabet. For a background sequence $A_1, \ldots, A_n$ of length n consisting of letters from $\mathcal{A}$, each letter A_k is generated sequentially under an irreducible aperiodic first-order Markov chain model, parameterized by a transition probability matrix T, where $T_{i,j} = Prob(A_{k+1} = j | A_k = i)$, $i, j \in \mathcal{A}$. Based on the transition matrix, we derive the stationary distribution π. Then, the letter A_1 of the first position is randomly chosen according to π, and subsequent letters A_{k+1} are randomly generated depending on the previous letter A_k following the A_k^{th} row of the transition matrix T. Here, we use the Markov transition matrix estimated from the real CRM sequences in mouse forebrain.

To generate CRM sequences, we insert the motif word $m = AGCCA$ of length $L = 5$ randomly into the background sequence at rate $1 - \lambda$. That is, for a background sequence $A_1, \ldots, A_n$ with length n, where $A_k \in \mathcal{A}, k = 1, \ldots, n$, we generate Bernoulli random variables $Z_1, Z_2, \ldots, Z_{n-L+1}$ with $Prob(Z_k = 1) = 1 - \lambda$. If for position k, $Z_k = 1$, and no motif has been inserted which overlaps position k, then we replace $A_k A_{k+1} \ldots A_{k+L-1}$ by the word m. Motif insertions are not allowed to overlap.

This toy model is clearly overly simplistic; CRMs are often defined by their functional output rather than their motif content. Hence, the shared motif content of similar CRMs can be rather complicated. Yet, the model picks up on one mechanism that may relate CRMs, see Hardison and Taylor (2012), and it is useful as a model to assess the performance of the five statistics.

Table 2. (a) The 90% CI of AUC scores and (b) the 90% CI of false-positive rate at 20% sensitivity for the CRM identification problem on simulated data over 100 repeats under a first-order Markov chain, $N = 1000$, $k = 5$, $|S_0| = 10$, $|S_{CRM}| = 100$, $|S_{bkg}| = 100$; $1 - \lambda$ is the motif density

λ	C_l^*	C_l^S	$\overline{C_2^*}$	$\overline{C_2^S}$	$\overline{C_2^{geo}}$
(a) The 90% CI of AUC scores					
0.95	[1.00,1.00]	[0.65,0.83]	[1.00,1.00]	[1.00,1.00]	[1.00,1.00]
0.97	[1.00,1.00]	[0.57,0.75]	[1.00,1.00]	[1.00,1.00]	[1.00,1.00]
0.99	[0.99,1.00]	[0.48,0.63]	[0.98,1.00]	[0.93,0.98]	[0.98,1.00]
0.991	[0.98,1.00]	[0.46,0.62]	[0.97,0.99]	[0.91,0.97]	[0.97,1.00]
0.993	[0.96,0.99]	[0.46,0.62]	[0.91,0.97]	[0.82,0.92]	[0.92,0.98]
0.995	[0.88,0.95]	[0.45,0.59]	[0.78,0.89]	[0.69,0.82]	[0.75,0.89]
(b) the 90% CI of the false positive rate at 20% sensitivity					
0.95	[0.00,0.00]	[0.00,0.06]	[0.00,0.00]	[0.00,0.00]	[0.00,0.00]
0.97	[0.00,0.00]	[0.01,0.12]	[0.00,0.00]	[0.00,0.00]	[0.00,0.00]
0.99	[0.00,0.00]	[0.07,0.24]	[0.00,0.00]	[0.00,0.00]	[0.00,0.00]
0.991	[0.00,0.00]	[0.07,0.26]	[0.00,0.00]	[0.00,0.01]	[0.00,0.00]
0.993	[0.00,0.00]	[0.08,0.25]	[0.00,0.00]	[0.00,0.02]	[0.00,0.00]
0.995	[0.00,0.00]	[0.10,0.27]	[0.00,0.03]	[0.00,0.08]	[0.00,0.02]

Moreover, we can relate the results to previous studies such as Reinert *et al.* (2009), which used the same toy model.

We set the length of sequences in both the positive and the negative sets to $N = 1000$ base pairs. We consider $\lambda \in \{0.95, 0.97, 0.99, 0.991, 0.993, 0.995\}$. For each value of λ, we first generate 10 CRMs as the known CRM set S_0. Then, for the candidate set S_1, we generate 200 sequences in total—a set S_{CRM} of 100 CRMs with the same motif word m and the same insertion rate λ as for S_0, and a set S_{bkg} of 100 background sequences. For each of the following experiments, we simulate 100 repeats.

Table 2 shows (a) the 90% CI of AUC score and (b) the 90% CI of false-positive rate at 20% sensitivity of the five similarity measures (2)–(6) over 100 repeats under the first-order Markov chain sequence model, when $N = 1000$, $k = 5$, $|S_0| = 10$, $|S_{CRM}| = 100$, $|S_{bkg}| = 100$, $\lambda \in \{0.95, 0.97, 0.99, 0.991, 0.993, 0.995\}$.

For the 90% CI of AUC scores, most of the statistics show a good performance and, based on a 90% CI, are not statistically significantly different, with the exception of C_l^S, which is outperformed by the other statistics. As Reinert *et al.* (2009) and Song *et al.* (2013) find that Shepp-type statistics are worse than star-type statistics in their simulation studies, the poor perfomance of C_l^S is not a surprise.

When λ is larger than 0.995, the values of the AUC score for our statistics become close to 0.5 (data not shown), which is the AUC expected from random guesses. The reason behind this deterioration is simply that the higher λ is, the fewer motifs are inserted in the generated CRM sequences. For example, when $\lambda = 0.997$, there are 30% of CRM sequences having at most one instance of the motif inserted. When $\lambda = 0.999$, then 65% of CRM sequences have at most one instance of the motif inserted. The 5% precision values give similar results, with much larger standard deviations due to smaller numbers; they can be found in the Supplementary Table S1.

The false-positive rates at 20% sensitivity are fairly low with the exception of C_l^S, which is outperformed by all other statistics for large λ.

3.2 Mouse tissue data

Although the results from the simulation study are encouraging, we need to test the statistics on real data, and we use CRM data as our test case. Tissue-specific enhancers in mouse embryos have been identified in Blow *et al.* (2010) using ChIP-Seq technology with the enhancer-associated protein p300. There are four tissues under study, namely, forebrain, midbrain, limb and heart. We take the four datasets one by one as our experimental data.

We filter out the CRM sequences that are too long or too short because our statistics may not work well when there is considerable heterogeneity in sequence length in conjunction with a possible model mis-specification. Therefore, we take sequences with length between 1000 and 1100 base pairs and ensure a maximum of 30% of repetitive region, as our CRM sequences. There are 106 of 2759 sequences meeting the criteria in forebrain dataset, 142 of 2786 in midbrain, 102 of 3839 in limb and 78 of 3597 in heart. Supplementary Table S4 of the supplementary materials shows the results for the full dataset; they are qualitatively similar.

As an indication of the strength of the regulatory patterns in CRM, we use the QuEST Scores as provided in the supplementary table of Blow *et al.* (2010). We use the CRMs with the 15 highest QuEST scores as the candidate sequences for S_0 and denote them as S_0^{pool}; the set S_{CRM}^{pool} of the remaining CRMs is taken as set of candidate sequences for S_{CRM}. For the set S_{bkg}, we randomly select a set S_{bkg}^{pool} of sequences with the same length as the sequences in S_{CRM}^{pool} from the mouse genome.

We randomly choose 10 sequences from S_0^{pool} to form the set S_0, and 30 sequences from S_{CRM}^{pool} as well as 30 sequences from S_{bkg}^{pool} to form the candidate sequences set S_1. Assuming that the sequences come from a first-order Markov chain model, we calculate the five statistics (2)–(6) to score the similarity within each set $\{S_0; c_i\}$, where $c_i \in S_1$, and evaluate their prediction accuracy by the AUC. We repeat the experiment 30 times. Table 3 shows (a) the 90% CI of AUC scores and (b) the 90% CI of false-positive rate at 20% sensitivity of each of the five statistics in mouse forebrain, midbrain, limb and heart tissue enhancers

Table 3. (a) The 90% CI of AUC scores and (b) the 90% CI of false-positive rate at 20% sensitivity for the CRM identification on four mouse tissue-specific enhancer datasets, under a first-order Markov chain; $N = 1000 - 1100$ bp, $k = 5$, $|S_0| = 10$, $|S_{CRM}| = 30$, $|S_{bkg}| = 30$

Tissue	C_l^*	C_l^S	$\overline{C_2^*}$	$\overline{C_2^S}$	$\overline{C_2^{geo}}$
(a) The 90% CI of AUC scores					
Forebrain	[0.52,0.69]	[0.67,0.85]	[0.58,0.77]	[0.55,0.74]	[0.51,0.71]
Midbrain	[0.51,0.71]	[0.55,0.71]	[0.64,0.81]	[0.61,0.79]	[0.53,0.74]
Limb	[0.68,0.82]	[0.58,0.83]	[0.69,0.86]	[0.64,0.81]	[0.62,0.83]
Heart	[0.41,0.62]	[0.52,0.77]	[0.57,0.79]	[0.58,0.78]	[0.51,0.67]
(b) The 90% CI of the false-positive rate at 20% sensitivity					
Forebrain	[0.00,0.19]	[0.00,0.07]	[0.00,0.14]	[0.02,0.19]	[0.02,0.17]
Midbrain	[0.07,0.25]	[0.00,0.20]	[0.00,0.15]	[0.00,0.15]	[0.07,0.27]
Limb	[0.00,0.08]	[0.00,0.10]	[0.00,0.05]	[0.00,0.13]	[0.00,0.12]
Heart	[0.07,0.27]	[0.00,0.17]	[0.00,0.22]	[0.03,0.17]	[0.07,0.29]

datasets, under the first-order Markov chain model and $k = 5$. The average values are located fairly in the middle of the intervals (data not shown).

For the AUC scores, all of the 90% CIs for each tissue type overlap, indicating that the statistics perform similarly. It is notable that the 90% AUC CIs for the limb dataset are on average higher than those in the other three datasets. This finding is consistent with the fact that the sequences in limb dataset have relatively high QuEST scores, indicating that the limb dataset has a strong regulatory signal. Most of the 90% CIs for the false-positive rate at 20% sensitivity contain 0, and they all overlap, indicating again that the statistics perform similarly.

Summarizing our results from the simulation study and the mouse tissue data, while all the statistics perform similarly on the real data, in the simulation study the combined Shepp-type statistic C_l^S is outperformed by the other statistics for large λ.

4 PERFORMANCE ON MEASURING SIMILARITY

To assess whether a set of regulatory sequences such as the set of CRMs are scored higher by our proposed statistics (2)–(3) and (7)–(9) than a set of unrelated sequences randomly chosen from the genome, we construct the evaluation model as follows. We make up two sets of sequences: the so-called *positive set* containing CRM sequences with a similar regulatory pattern and the so-called *negative set* containing unrelated sequences randomly chosen from the genome. The two sets are constructed to contain the same number of sequences and have the same lengths. Each triplet of three sequences from the positive set is measured for their similarity by the proposed five statistics, and so is each triplet in the negative set. We assess the statistics by first ranking all triplets from both positive and negative sets by their scores. Triplets from the positive set are treated as '+' and triplets from the negative set are treated as '−'. For a given threshold, if the value of a statistic for a triplet is above the threshold, the triplet is predicted as '+'; otherwise the triplet is predicted as '−'.

4.1 Simulation study

Following the simulation methods detailed in Section 3.1, we generate 50 CRM-like sequences for the positive set, and 50 background sequences for the negative set, all of length $N = 1000$ base pairs. We carry out the experiment for motif insertion rate $\lambda \in \{0.95, 0.97, 0.99, 0.991, 0.993, 0.995\}$. For each λ, we repeat the experiment 30 times and calculate the 90% CI of AUCs together with that of the false-positive rate at 20% sensitivity. Table 4 shows (a) the 90% CI of AUC score and (b) the 90% CI of false-positive rate at 20% sensitivity of the five similarity measures on simulated data, under the first-order Markov chain sequence model.

In this simulation study, the Shepp-type statistics C_l^S and $\overline{\overline{C_2^S}}$ are outperformed by the other statistics when λ is large, and $\overline{\overline{C_2^{geo}}}$ is outperformed by the star-type statistics for $\lambda = 0.993$. This conclusion is consistent with the simulation results for the CRM identification problem.

4.2 Mouse tissue data

In this section, we use the tissue-specific enhancers in mouse embryos (Blow *et al.*, 2010) again. Four tissues, forebrain, midbrain, limb and heart, are studied as separate datasets. Again, we take the CRM sequences with 1000–1100 bp as our CRM sequences pool. For the background sequences pool, we randomly select sequences from the mouse genome, of the same length as that of the CRM sequences.

For each tissue, we randomly choose 20 CRMs from the CRM pool as the positive set, and we choose 20 background sequences from the background sequences pool as the negative set. Then we calculate statistics (2)–(3) and (7)–(9) for the triplets from the positive set and from the negative set, and we compute the AUC; again we run 30 repetitions. Table 5 presents (a) the 90% CI of AUC scores and (b) the 90% CI of false-positive rate at 20% sensitivity for the five statistics on each of mouse four tissue-specific enhancer datasets under the first-order Markov chain model, and $k = 5$; again the average AUC scores lie in the middle of the intervals, and the 5% precision scores can be found in the appendix.

For each tissue type, all of the 90% CIs overlap, showing that the statistics perform similarly.

We repeated the experiment on the full dataset, without restricting the sequence length, giving similar results, see

Table 4. (a) The 90% CI of AUC scores and (b) the 90% CI of false-positive rate at 20% sensitivity for measuring similarity within a set of sequences on simulated data, under the first-order Markov chain sequence model, $N = 1000$, $k = 5$, |positive set| = 50, |negative set| = 50; $1 - \lambda$ is the motif density

λ	C_l^*	C_l^S	$\overline{C_2^*}$	$\overline{C_2^S}$	$\overline{C_2^{geo}}$
(a) The 90% CI of AUC scores					
0.95	[1.00,1.00]	[0.99,1.00]	[1.00,1.00]	[1.00,1.00]	[1.00,1.00]
0.97	[1.00,1.00]	[0.88,0.95]	[1.00,1.00]	[1.00,1.00]	[1.00,1.00]
0.99	[0.89,0.99]	[0.56,0.69]	[0.95,0.99]	[0.85,0.91]	[0.77,0.91]
0.991	[0.83,0.98]	[0.51,0.70]	[0.93,0.97]	[0.82,0.88]	[0.69,0.89]
0.993	[0.67,0.88]	[0.48,0.65]	[0.83,0.91]	[0.72,0.81]	[0.57,0.78]
0.995	[0.53,0.76]	[0.48,0.70]	[0.70,0.81]	[0.63,0.73]	[0.51,0.72]
(b) The 90% CI of false-positive rate at 20% sensitivity					
0.95	[0.00,0.00]	[0.00,0.00]	[0.00,0.00]	[0.00,0.00]	[0.00,0.00]
0.97	[0.00,0.00]	[0.00,0.00]	[0.00,0.00]	[0.00,0.00]	[0.00,0.00]
0.99	[0.00,0.00]	[0.05,0.14]	[0.00,0.00]	[0.00,0.01]	[0.00,0.01]
0.991	[0.00,0.00]	[0.04,0.19]	[0.00,0.00]	[0.01,0.01]	[0.00,0.03]
0.993	[0.00,0.01]	[0.07,0.23]	[0.00,0.01]	[0.01,0.05]	[0.01,0.12]
0.995	[0.01,0.16]	[0.05,0.23]	[0.01,0.05]	[0.04,0.09]	[0.03,0.21]

Table 5. (a) The 90% CI of AUC scores and (b) the 90% CI of false-positive rate at 20% sensitivity for measuring similarity within a set of sequences based on four mouse tissue-specific enhancer datasets, under first-order Markov chain background sequences, $N = 1000 - 1100$ bp, $k = 5$, |positive set| = 20, |negative set| = 20

Tissues	C_l^*	C_l^S	$\overline{C_2^*}$	$\overline{C_2^S}$	$\overline{C_2^{geo}}$
(a) The 90% CI of AUC scores					
Forebrain	[0.62,0.96]	[0.60,0.87]	[0.57,0.80]	[0.51,0.79]	[0.69,0.97]
Midbrain	[0.65,0.93]	[0.54,0.86]	[0.58,0.83]	[0.52,0.80]	[0.68,0.94]
Limb	[0.75,0.97]	[0.51,0.83]	[0.64,0.89]	[0.55,0.83]	[0.77,0.99]
Heart	[0.49,0.87]	[0.47,0.74]	[0.58,0.78]	[0.50,0.72]	[0.61,0.88]
(b) The 90% CI of the false-positive rate at 20% sensitivity					
Forebrain	[0.00,0.10]	[0.00,0.13]	[0.02,0.13]	[0.03,0.22]	[0.00,0.06]
Midbrain	[0.00,0.11]	[0.01,0.21]	[0.01,0.14]	[0.02,0.22]	[0.00,0.10]
Limb	[0.00,0.07]	[0.01,0.16]	[0.00,0.09]	[0.01,0.15]	[0.00,0.05]
Heart	[0.01,0.16]	[0.04,0.25]	[0.02,0.12]	[0.05,0.20]	[0.00,0.12]

Supplementary Tables S11 and S12. Again the AUC scores have relatively large deviations, and the differences between the AUC scores of the different statistics are not statistically significant.

5 THE EFFECT OF CONTAMINATION

To study the sensitivities of the statistics to the contamination of S_0 with random sequences, we carry out a simulation experiment; we concentrate on the CRM identification problem. For the simulation with $|S_0| = 10$, $|S_1| = 100$, $|S_2| = 100$, $k = 5$ and $N = 1000$, we randomly replace 10, 20 and 30% of sequences in S_0 with background sequences, to model the contamination of sequences in S_0. Figure 1 shows the results of this simulation. We find that both multiple alignment-free statistics C_l^* and C_l^S are relatively more sensitive to the contamination because for $\lambda = 0.95$, the AUC scores of C_l^S begins to drop at 10% random sequences mixed into S_0, and the AUC scores of C_l^* drops at the 20% contamination of S_0, whereas the other three pairwise average statistics $\overline{C_2^*}$, $\overline{C_2^S}$ and $\overline{C_2^{geo}}$ keep their AUC scores to be 1. When $\lambda = 0.993$, the statistics C_l^*, C_l^S and $\overline{C_2^{geo}}$ have the lowest lower bounds, whereas the performance of $\overline{C_2^*}$ and $\overline{C_2^S}$ is relatively stable. As the AUC scores of C_l^S have already been ~0.5 (near to random guess), they do not decrease further.

The simulation study indicates that the multiple alignment-free statistics perform well when the sequences in S_0 have a strong regulatory pattern, but the mixture of random sequences in S_0 could largely reduce the performance of the multiple alignment-free statistics, compared with the pairwise average ones. In real data, our results in Table 3 show the similar trend. Blow *et al.* (2010) states that the forebrain CRM dataset is the most conserved in the four datasets, indicating most of the forebrain CRM sequences are real, and the fraction of random sequences is low. In our results, C_l^S gives the best performance in forebrain dataset at the average values. The fact is consistent with the findings in Blow *et al.* (2010).

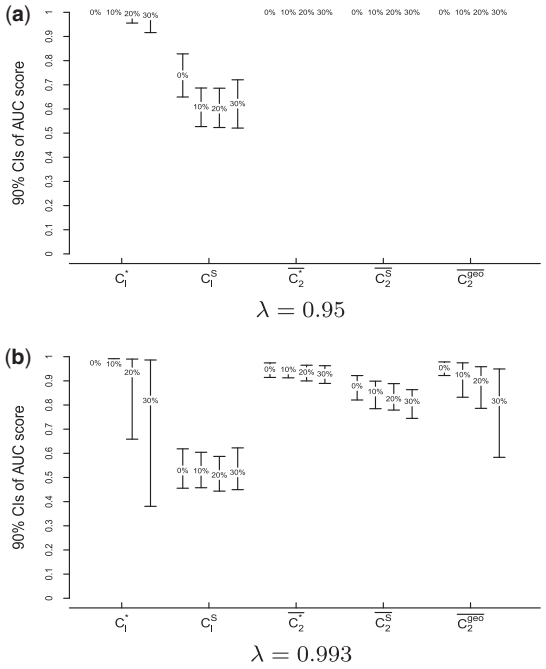

Fig. 1. The 90% CI of AUC score (the bar) and average AUC scores (the position of the number) of the five statistics for the CRM identification on 0, 10, 20 and 30% contaminated S_0 sequences, under a first-order Markov chain, $N = 1000$, $k = 5$, $|S_0| = 10$, $|S_{CRM}| = 100$, $|S_{bkg}| = 100$; $1 - \lambda$ is the motif density, and 0% refers to the results in Table 2, for comparison with the case of no contamination of S_0. The parameter values are (a) $\lambda = 0.95$ and (b) $\lambda = 0.993$.

Further, to assess whether less contaminated data can improve the performance of the multiple alignment-free statistics, we use a more restricted criterion to construct S_0. In every experiment of the 30 repeats, we randomly pick up five sequences from the CRMs with the 10 highest QuEST scores to construct S_0, 30 sequences from the remaining CRMs with 50 highest scores to construct S_{CRM} and 30 sequences from the background sequences generated in the same way as before. In this setting the average values of AUC score of the multiple alignment-free statistics increase more than those of the pairwise average statistics, although their CIs overlap.

6 DISCUSSION

The success of the alignment-free sequence comparison statistics D_2^* and D_2^S as studied in Reinert *et al.* (2009) and Wan *et al.* (2010) inspired us to define natural extensions for multiple alignment-free sequence comparison. We propose two families of multiple alignment-free statistics, C_l^* and C_l^S, which involve products of more than two vectors, and we compare these families to three families of average pairwise alignment-free statistics, $\overline{C_2^*}$, $\overline{C_2^S}$ and $\overline{C_2^{geo}}$, with generalizations $\overline{\overline{C_2^*}}$, $\overline{\overline{C_2^S}}$ and $\overline{\overline{C_2^{geo}}}$.

Our statistics are tested on two problems, namely, the CRM identification problem (given a set of CRMs, identify new CRMs belonging to similar regulatory patterns as the given CRMs) and measuring the similarity within a set of CRM sequences. We evaluate the statistics both on simulated data and on mouse tissue data.

For both the CRM identification and measurement of the similarity within a set of CRM sequences, in simulations, the Shepp-type statistics are outperformed by other statistics when λ is large, whereas on real data, all of the 90% CIs overlap. The false-positive rate at 20% sensitivity is encouragingly small not only in simulations but also in real data. Contamination has a stronger effect on the multiple statistics than on the pairwise average statistics.

While Song *et al.* (2013) and Jiang *et al.* (2012) observed that the Shepp-type statistics outperform the star-type statistics for the comparison of whole-genome sequences and metagenomic communities, respectively, based on NGS reads data, here our data do not confirm this finding. Our data are consistent with the simulation results in Reinert *et al.* (2009); Wan *et al.* (2010), which show that for short sequences the star-type statistics perform better, whereas for long sequences, the Shepp-type statistics perform better; the CRM sequences in this article are of length of order 1000 bp and can hence be considered as short.

Our results on the real dataset when restricted to sequences of length 1000–1100 bp do not differ significantly from those on the full real dataset, indicating that our statistics control for sequence length.

Although on the CRM data we use, there is no significant difference between all five statistics, we conjecture that on high-quality data, the multiple statistics may outperform the average pairwise statistics. With the impending arrival of more high-quality data, our suite of statistics provides a toolbox ready to be used.

Pairwise alignment-free statistics have been generalized to allow for k-tuple word mismatches, see Burden *et al.* (2008) and Göke *et al.* (2012). It would be straightforward to extend our multiple alignment-free statistics to allow for word mismatches; the choice of mismatch weights is however not obvious, see Göke *et al.* (2012).

ACKNOWLEDGEMENTS

The authors would like to thank the referees for very helpful comments, which improved the article.

Funding: G.R. was supported in part by the Oxford Martin School. F.S. is partially supported by US NIH R21HG006199; and NSF DMS-1043075 and OCE 1136818. M.D. is supported by the National Natural Science Foundation of China (31171262, 11021463), and the National Key Basic Research Project of China (2009CB918503).

Conflict of Interest: none declared.

REFERENCES

Arunachalam,M. *et al.* (2010) An alignment-free method to identify candidate orthologous enhancers in multiple drosophila genomes. *Bioinformatics*, **26**, 2109–2115.

Blaisdell,B. (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Natl Acad. Sci. USA*, **83**, 5155–5159.

Blow,M. *et al.* (2010) Chip-seq identification of weakly conserved heart enhancers. *Nat. Genet.*, **42**, 806–810.

Burden,C. *et al*. (2008) Approximate word matches between two random sequences. *Ann. Appl. Probab*., **18**, 1–21.
Davidson,E. (2006) *The Regulatory Genome: Gene Regulatory Networks In Development and Evolution*. Elsevier, Burlington, MA.
Göke,J. *et al*. (2012) Estimation of pairwise sequence similarity of mammalian enhancers with word neighbourhood counts. *Bioinformatics*, **28**, 656–663.
Hardison,R.C. and Taylor,J. (2012) Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet*., **13**, 469–483.
Jiang,B. *et al*. (2012) Comparison of metagenomic samples using sequence signatures. *BMC Genomics*, **13**, 730.
Kantorovitz,M. *et al*. (2007) A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, **23**, i249–i255.
Lippert,R. *et al*. (2002) Distributional regimes for the number of k-word matches between two random sequences. *Proc. Natl Acad. Sci. USA*, **99**, 13980–13989.
Liu,X. *et al*. (2011) New powerful statistics for alignment-free sequence comparison under a pattern transfer model. *J. Theor. Biol*., **284** (1), 106–116.
Needleman,S. *et al*. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol*., **48**, 443–453.
Quine,M. (1994) A result of Shepp. *Appl. Math. Lett*., **7**, 27–32.
Reinert,G. *et al*. (2009) Alignment-free sequence comparison (i): statistics and power. *J. Comput. Biol*., **16**, 1615–1634.
Shepp,L. (1964) Normal functions of normal random variables. *SIAM Rev*., **6**, 459–460.
Sing,T. *et al*. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.
Smith,T. and Waterman,M. (1981) Identification of common molecular subsequences. *J. Mol. Biol*., **147**, 195–197.
Song,K. *et al*. (2013) Alignment-free sequence comparison based on next generation sequencing reads. *J. Comput. Biol*., **20**, 64–79.
Wan,L. *et al*. (2010) Alignment-free sequence comparison (ii): theoretical power of comparison statistics. *J. Comput. Biol*., **17** (11), 1467–1490.
Wolff,C. *et al*. (1999) Structure and evolution of a pair-rule interaction element: runt regulatory sequences in D. melanogaster and D. virilis. *Mech. Dev*., **80**, 87–99.