Briefings in Bioinformatics Advance Access published September 23, 2013

doi:10.1093/bib/bbt067

New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing

Kai Song, Jie Ren, Gesine Reinert, Minghua Deng, Michael S. Waterman and Fengzhu Sun

Submitted: 28th May 2013; Received (in revised form): 25th July 2013

Abstract

With the development of next-generation sequencing (NGS) technologies, a large amount of short read data has been generated. Assembly of these short reads can be challenging for genomes and metagenomes without template sequences, making alignment-based genome sequence comparison difficult. In addition, sequence reads from NGS can come from different regions of various genomes and they may not be alignable. Sequence signature-based methods for genome comparison based on the frequencies of word patterns in genomes and metagenomes can potentially be useful for the analysis of short reads data from NGS. Here we review the recent development of alignment-free genome and metagenome comparison based on the frequencies of word patterns with emphasis on the dissimilarity measures between sequences, the statistical power of these measures when two sequences are related and the applications of these measures to NGS data.

Keywords: *alignment-free; word patterns; Markov model; genome comparison; statistical power; NGS data*

INTRODUCTION

Sequence comparison continues to play crucial roles in molecular sequence analysis. The dominant approaches for sequence comparison are alignment-based including the Smith–Waterman algorithm [1] and BLAST [2]. Although alignment-based approaches generally yield excellent results when the molecular sequences of interest can be reliably aligned, their applications are limited when the sequences are divergent or come from different regions of various genomes and a reliable alignment cannot be obtained. Another drawback of alignment-based approaches is that they are generally time-consuming and thus, are limited in dealing with large-scale sequence data generated with the new sequencing technologies. The next-generation sequencing (NGS) technologies usually generate relatively short reads that can be difficult to assemble, and alignment-based approaches cannot be applied when the complete sequences are not known. Alignment-free

Corresponding authors. Fengzhu Sun or Michael S. Waterman, Molecular and Computational Biology Program, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, USA. Tel: +1-213-740-2413; Fax: +1-213-740-8431; E-mail: fsun@usc.edu or msw@usc.edu

Kai Song and Jie Ren are graduate students at the Department of Probability and Statistics, Peking University, China. They work in computational biology.

Gesine Reinert, PhD, is a University Lecturer in the Department of Statistics and a fellow of Keble College, Oxford University, UK. Her research interests include applied probability, computational biology, and statistics, in particular, Stein's method, networks and word count statistics.

Minghua Deng, PhD, is a full professor in the Department of Probability and Statistics, Peking University, China. His research includes using probabilistic and statistical methods to solve biological problems.

Michael S. Waterman, PhD, is a full professor of mathematics, computer sciences and biological sciences, University of Southern California, USA. He has contributed significantly to the computational analysis of molecular sequence data. He is a member of the US National Academy of Sciences and National Academy of Engineering.

Fengzhu Sun, PhD, is a full professor of Molecular and Computational Biology Program, University of Southern California, USA. His research includes developing statistical approaches for the analysis of genomics and proteomics data. He is an elected fellow of AAAS.

sequence comparison approaches provide attractive alternatives when alignment-based approaches fail.

Studies of alignment-free sequence comparison began in the mid 80s [3, 4] and have been under continuous development. Two excellent reviews are available on this topic [5, 6]. Here we review new developments on alignment-free sequence comparison with emphasis on methods based on *k*-tuple (*k*-word, *k*-gram) frequencies. For studies on alignment-free sequence comparison based on other ideas such as chaos theory and common substrings, please see [6] for details.

Alignment-free sequence comparison has been successfully applied to many biological problems including (i) the study of evolution of organisms using whole-genome sequences, (ii) the evolution of regulatory sequences such as promoters, enhancers and inhibitors, (iii) the identification of *cis*-regulatory modules (CRM) and (iv) the comparison of metagenomic communities based on the sequence data using NGS technologies.

For alignment-free sequence comparison of two sequences using *k*-tuples, the first step is to count the number of occurrences of every *k*-tuple in both sequences separately and record these in a vector of *k*-tuple frequencies for each sequence; this counting can be carried out in linear time. Second, a measure d of difference between the two sequences **A** and **B** is defined based on the two frequency vectors. If the measure satisfies distance constraints, *i.e.* (a) $d(A,B) \geq 0$ and equality holds if and only if $A = B$, and (b) for any sequences A, B and C, $d(A,C) \leq d(A,B) + d(B,C)$, then the measure is a distance measure. Otherwise, the measure is called a dissimilarity measure. Third, the sequences are then clustered based on the distance or dissimilarity measures and the resulting clusters are finally compared with current biological knowledge about the sequences to evaluate the effectiveness of the measures. Many measures have been developed over the years. Here we present a general review of such measures and their applications to molecular sequence analysis with an emphasis on NGS data.

The article is organized as follows. In Section 1, we review theoretical studies of the approximate distributions of the popular D_2 statistic and of its power. As D_2 may mainly measure background noise in each sequence separately, in Section 2 we describe adjusted similarity measures based on word counts. Section 3 focuses on alignment-free genome and metagenome comparison using NGS data. Section 4 contains a discussion and conclusions.

THEORETICAL STUDIES OF THE APPROXIMATE DISTRIBUTIONS OF D_2 AND ITS STATISTICAL POWER

Alignment-free sequence comparison by the number of word matches: the D_2 statistic

The earliest development of alignment-free sequence comparison can be traced back to the 1980's. Blaisdell [4] used the likelihood ratio statistic or the corresponding Chi-square statistic, to determine the order of Markov chain (MC) of sequences of interest. More specifically, for a given sequence $A = A_1A_2\cdots A_n$ with letters from a finite alphabet $\mathcal{A}$, let $(X_w, w \in \mathcal{A}^k)$ be the number of occurrences of tuple $\boldsymbol{w}$ in **A**. The objective is to test if the sequence **A** can be modelled as a (*k*-2)-th order MC as opposed to a (*k*-1)-th order MC using the likelihood ratio statistic

$$L_k = 2\sum_{a_1,\cdots,a_k\in\mathcal{A}} X_{a_1a_2\cdots a_k}\log\frac{X_{a_1a_2\cdots a_k}}{E_{a_1a_2\cdots a_k}}$$

or the Chi-square statistic

$$S_k = \sum_{a_1,\cdots,a_k\in\mathcal{A}} \frac{(X_{a_1a_2\cdots a_k} - E_{a_1a_2\cdots a_k})^2}{E_{a_1a_2\cdots a_k}}$$

where $E_{a_1a_2\cdots a_k}$ is the expected count of the *k*-tuple $a_1a_2\cdots a_k$ under the (*k*-2)-th order MC, which can be estimated by

$$\hat{E}_{a_1a_2\cdots a_k} = \frac{X_{a_1a_2\cdots a_{k-1}a_k}X_{a_2\cdots a_{k-1}a_k}}{X_{a_2\cdots a_{k-2}a_{k-1}}}. \quad (1)$$

Both L_k and S_k have an approximate χ^2-distribution with $4^{k-2}\times 9$ degrees of freedom. Blaisdell [3] used similar ideas to test if a set of sequences have the same transition matrix, and thus are more likely to be related. These statistics can be considered as the origin of alignment-free sequence comparison statistics.

Torney *et al.* [7] used the number of *k*-tuple matches between two sequences **A** and **B** as a statistic to measure the similarity between them. Let

$$\begin{aligned} D_2 &= \sum_{i,j} I(A_iA_{i+1}\cdots A_{i+k-1} = B_jB_{j+1}\cdots B_{j+k-1}) \\ &= \sum_{w\in\mathcal{A}^k} X_wY_w, \end{aligned} \quad (2)$$

where X_w and Y_w are the number of occurrences of tuple w in sequences **A** and **B**, respectively, and $I(\cdot)$ denotes the logical indicator: $I(E) = 1$ if event E is true, and 0 otherwise. The D_2 statistic has been used in many applications including sequence database searches [8] and clustering of expressed sequence tags [9]. Owing to its wide range of applications, extensive studies on the distributions of D_2 have been carried out.

The distribution of the D_2 statistic under the null model of two independent sequences

Lippert *et al.* [10] studied the limiting distribution of D_2 under the independent identically distributed (i.i.d.) model for both sequences with the same nucleotide frequencies $p_a, a \in \mathcal{A}$, where $\mathcal{A}$ indicates the set of all the possible letters. When $p_a, a \in \mathcal{A}$ are not all equal, it was shown that when $k \geq 2\log(n)$ where n is the length of both sequences, D_2 has an approximate Poisson distribution, and when $k < 1/2\log(n)$, D_2 has an approximate normal distribution. It was further suggested in [10] and explicitly proved in [11] that the variance of D_2 is dominated by the variance of the number of occurrences of each k-tuple in individual sequences. However, when $p_A = p_C = p_G = p_T = 1/4$, D_2 is approximately neither normal nor Poisson. Instead, D_2 tends to the sum of products of normal distributions.

The fundamental results in [10] were further extended to more general models for the sequences of interest [12–14]. Kantorovitz *et al.* [12] showed that D_2 is approximately normal as both k and n tend to infinity when the nucleotide frequencies are the same for both sequences. Burden *et al.* [14] extended the D_2 statistic to allow word matches with up to a certain number of mismatches and again showed that this new statistic is approximately normally distributed. Foret *et al.* [13] compared the empirical and theoretical distributions of D_2 and its variations and found that the approximations are consistent with the empirical distributions.

The power of the D_2 statistic under the alternative model of two related sequences

The results from [10, 12–14] played key roles in estimating the statistical significance of D_2 between two sequences under the null hypothesis that the two sequences are unrelated. In contrast, these studies do not address what kind of relationship the statistic D_2 and its variants can detect, and what the statistical power is when the alternative hypothesis that the two sequences are related holds. To answer these questions, precise definitions of relatedness between the sequences of interest are needed. The alternative hypothesis depends on the scientific questions of interest. One of the key applications of alignment-free sequence comparison is to find CRMs that locate in the upstream regions of genes controlled by the same sets of transcription factors. Sequences in the same CRM tend to have the same transcription factor binding sites and thus share similar sets of k-tuples. Therefore, Reinert *et al.* [11] modelled the relatedness of sequences by the sharing of common *motifs*, that is, word patterns that are significantly enriched in the sequences. They refer to the model as a common motif model. The model for each sequence consists of the following three components:

(1) The background sequences are modelled by an i.i.d. model.
(2) The foreground motifs are modelled by position weight matrices that give the nucleotide probability distribution at each position of the motifs. The foreground motif model can be easily extended to the situation that the nucleotides along the motifs depend on each other. The motifs can also be easily generalized to CRMs consisting of combinations of several motifs.
(3) The occurrences of the motifs are modelled as binomial random variables along the genome sequence, with $1 - \lambda$ denoting the probability that a motif instance starts at a nucleotide position; $1 - \lambda$ is referred as the *motif density*. Once a motif is inserted, the nucleotide positions, which are now covered by the motif, are ignored, and the insertion process resumes at the end of the motif, so that inserted motifs do not overlap.

Intuitively, the relatedness between the two sequences increases with the motif density, and thus the power of a reasonable statistic should also increase with motif density. Simulation studies under the common motif model showed that for small values of k, the power of D_2 can be smaller than the type I error rate and can decrease with sequence length. Thus, D_2 is not an appropriate statistic to test the relationship between sequences under the common motif model for small values of k. When

k is relatively large, the power does increase with sequence length. Thus, for many practical applications with relatively large values of k, D_2 can be useful. These observations were further proved theoretically in [15] based on the hidden Markov model for each sequence developed in [16].

ADJUSTED DISSIMILARITY MEASURES FOR ALIGNMENT-FREE SEQUENCE COMPARISON BASED ON *K*-TUPLE COUNTS

In addition to D_2, many other distance and dissimilarity measures based on k-tuple counts have been proposed for the comparison of molecular sequences. Those include (a) normalization of D_2 by $D2z$, (b) correlation of relative differences of the tuple counts from their expectations, (c) comparison of relative abundance of k-tuples and (d) comparison of k-tuple distributions based on Markov models.

Normalization of D_2 by $D2z$

It was realized that the D_2 statistic defined in Equation (2) depends on the underlying sequence models [12]. Normalizing the D_2 statistic using its mean and standard deviation can potentially improve the power of detecting the relationships between the sequences and remove the biases due to the background models of the sequences. For normalization, the background models for the sequences of interest are needed. Kantorovitz *et al.* [12] modified the D_2 statistic to $D2z$

$$D2z(A,B) = \frac{D_2(A,B) - E(D_2)}{\sqrt{Var(D_2)}} \tag{3}$$

where the expectation and variance of D_2 are calculated based on a Markov model for the sequences.

The authors compared the effectiveness of the $D2z$ statistic with several other distance or dissimilarity measures between the k-tuple vectors including (i) the Euclidean distance [3], (ii) Kullback-Leibler discrepancy (kld) [17], (iii) the linear Pearson correlation coefficient (lcc) between the count vectors, (iv) the cosine of the angle between the two k-tuple count vectors (cos) and (v) two other measures based on the approximate Poisson distribution of word counts [18]. Using four fly and three human CRMs as examples, it was shown that the $D2z$ statistic outperforms the statistics described above for the identification of CRMs [12]. Note though, the Euclidean, kld, lcc and cos measures need only one parameter, the value of k, while $D2z$ needs an additional parameter r, the order of the MC for the sequences. In the seven data sets studied, the combination of $k = 5$ and $r = 0$ yielded the best performance for $D2z$ in five of the data sets, and the combination of $k = 6$ and $r = 2$ has the best performance in two of the seven data sets.

Correlation of relative differences of *k*-tuple counts between two sequences

Instead of normalizing D_2 by $D2z$, Hao and colleagues [19–21] considered the relative difference vector of the number of occurrences of every k-tuple w with its expected count under the (k-2)-th order MC model given in Equation (1) for each sequence. They then used the correlation coefficient between the relative difference vectors corresponding to two sequences to measure their similarity. We use the corresponding author's last name *Hao* as the short form of this measure to simplify the notation.

$$Hao = \frac{1}{2}\left(1 - \frac{\sum_w \left(\frac{X_w - E_w^X}{E_w^X}\right)\left(\frac{Y_w - E_w^Y}{E_w^Y}\right)}{\sqrt{\sum_w \left(\frac{X_w - E_w^X}{E_w^X}\right)^2 \sum_w \left(\frac{Y_w - E_w^Y}{E_w^Y}\right)^2}}\right) \tag{4}$$

where E_w^X and E_w^Y are defined as in Equation (1) based on sequences **A** and **B**, respectively.

The authors applied this dissimilarity measure to whole-genome phylogenetic analysis [21–26] and classification of metagenomic samples [27]. However, Jiang *et al.* [28] recently found that, although the *Hao* statistic performs reasonably well with high-coverage data sets, it is not stable when the data are limited.

Alignment-free sequence comparison based on di-, tri- and tetra-nucleotide relative abundance

Karlin and colleagues observed that the relative dinucleotide frequencies defined by

$$\rho_{ab}(A) = \frac{f_{ab}}{f_a f_b} \tag{5}$$

where f_a is the frequency of nucleotide a and, more generally, f_w is the observed frequency of word pattern w, are relatively stable across different parts of the same genome and differ across different genomes [29–33]. Therefore, they proposed to use the l_1-norm between the relative dinucleotide frequencies between two genome sequences as dissimilarity measure, that is,

$$\delta(A,B) = \sum_{a,b\in\mathcal{A}} |\rho_{ab}(A) - \rho_{ab}(B)|. \tag{6}$$

This distance measure has been used to study the evolutionary relationships among viruses [29], bacteria [30], plasmids, prokaryotes [34] and eukaryotes [30, 32]. The dinucleotide frequencies have been extended to tri- and tetra-nucleotide biases [35, 36] as

$$\gamma_{abc} = \frac{f_{abc} f_a f_b f_c}{f_{ab} f_{bc} f_{aNc}}$$

and

$$\tau_{abcd} = \frac{f_{abcd} f_{ab} f_{aNc} f_{aN_1N_2d} f_{bc} f_{bNd} f_{cd}}{f_{abc} f_{abNd} f_{bcd} f_a f_b f_c f_d}$$

where N, N_1 and N_2 indicate any letters. Under the MC model, the corresponding biases are defined by the ratio of the observed count of a tri- or tetra-nucleotide in the sequence over its expected count under the first and second order Markov models given in Equation (1). Similarly, the difference between two genomes can be calculated by the l_p norm of the 64-dimensional vector γ or the 256-dimensional vector τ.

It can be seen that the *Hao* dissimilarity measure is closely related to the dissimilarity measure developed from Karlin's group when the MC model was used to model the sequences. *Hao*'s group used one minus the correlation between the relative difference vectors while Karlin's group used the l_p-norm between the relative difference vectors to measure the dissimilarity.

Alignment-free sequence comparison based on direct comparison of tuple frequencies with Markov models

Kim and colleagues [37–40] designed an approach based on the Jensen-Shannon entropy as a distance measure to study the evolutionary relationships of prokaryotes, *Escherichia coli/Shigella*, and viruses.

In a series of articles, Wang's group compared two sequences using the k-tuple frequency vectors coupled with MC models of order $r = 0, 1, 2$ [41, 42]. The methods in [42] can be described as follows. Assuming an r-th order MC for each sequence, the transition probability matrix can be estimated using the sequence data. The probability of each k-tuple to occur in the sequence **A** can then be calculated based on the transition probabilities denoted as $P(w|\Lambda_A^r)$, where Λ_A^r is the r-th order MC for sequence **A**. Similarly, the k-tuple probability distribution can be defined based on the second sequence **B**. The statistic $S1.k.r(A, B)$ is defined by the symmetric Jensen-Shannon divergence of the two probability measures $P(w|\Lambda_A^r)$ and $P(w|\Lambda_B^r)$. They also considered a weighted version of $S1.k.r(A, B)$ denoted as $S2.k.r(A, B)$ that was defined by replacing $P(w|\Lambda_S^r)$ with $f_w^S P(w|\Lambda_S^r)$, where f_w^S is the frequency of w in sequence $S = A$ or **B**. It was shown in [41,42] that for the identification of CRMs with appropriate choices of k and r, the dissimilarity measure $S2.k.r$ performs the best compared with the dissimilarity measures discussed above. We also tried $S2$ on the data sets used in [43] and found that $S2$ outperforms other dissimilarity measures by a significant margin for the identification of CRM sequences (see Supplementary Tables S1–S4).

The statistics D_2^S and D_2^* and their statistical power

Two relatively new normalization methods for the tuple counts have recently been proposed [11, 15]. The first statistic is based on the observation by Shepp [44] that for two independent normal random variables X and Y with mean zero, $XY/\sqrt{X^2 + Y^2}$ is also normally distributed. We normalize X_w and Y_w by

$$\widetilde{X}_w = X_w - \overline{n} p_w^X \quad \text{and} \quad \widetilde{Y}_w = Y_w - \overline{n} p_w^Y$$

where $\overline{n} = n - k$ and $\overline{m} = m - k$, and p_w^X and p_w^Y are the probability of k-tuple w under the background model for sequences **A** and **B**, respectively. The statistic D_2^S is defined by

$$D_2^S = \sum_{w \in \mathcal{A}^k} \frac{\widetilde{X}_w \widetilde{Y}_w}{\sqrt{\widetilde{X}_w^2 + \widetilde{Y}_w^2}} \tag{7}$$

where $\mathcal{A}^k$ is the set of all k-tuples. The second statistic, D_2^*, is based on the intuitive idea that the number of occurrences of tuple w is approximately Poisson and thus its mean and variance are approximately the same for relatively long tuples;

$$D_2^* = \sum_{w \in \mathcal{A}^k} \frac{\widetilde{X}_w \widetilde{Y}_w}{\sqrt{\overline{m}\overline{n} p_w^X p_w^Y}}. \tag{8}$$

These statistics were used to test the null hypothesis H_0 that the two sequences are not related versus the alternative hypothesis H_1 that two sequences are related by the common motif model. It was shown by simulations [11] and theoretically [15] that the new statistics D_2^S and D_2^* have the following properties:

(1) The power of both D_2^S and D_2^* is generally higher than that of D_2, and increases with the sequence length.

(2) The statistic D_2^* has the highest power when the length of the tuples, k, equals the length of the inserted motif.
(3) When the sequence length is relatively short, the statistic D_2^* is more powerful than D_2^S, while when the sequence length is long, the power of D_2^S is generally higher than that of D_2^*.

Although D_2^S and D_2^* are powerful statistics for the comparison of genomic sequences, they have the drawback that their magnitudes depend on a variety of factors including sequence length and nucleotide frequencies. To overcome these problems, the statistics D_2^S and D_2^* are further normalized to d_2^S and d_2^*, respectively, with range from 0 to 1, where

$$d_2^S = \frac{1}{2} \times \left(1 - \frac{D_2^S}{\sqrt{\sum_{w\in A^k} \widetilde{X}_w^2/\sqrt{\widetilde{X}_w^2+\widetilde{Y}_w^2}}\sqrt{\sum_{w\in A^k} \widetilde{Y}_w^2/\sqrt{\widetilde{X}_w^2+\widetilde{Y}_w^2}}}\right)$$

and

$$d_2^* = \frac{1}{2}\left(1 - \frac{D_2^*}{\sqrt{\sum_{w\in A^k} \widetilde{X}_w^2/(\overline{n}p_w^X)}\sqrt{\sum_{w\in A^k} \widetilde{Y}_w^2/(\overline{m}p_w^Y)}}\right).$$

When the two sequences are the same, both d_2^S and d_2^* equal to 0, while if they are anti-correlated, these statistics are close to 1. Therefore, they can be used as dissimilarity measures for two sequences and can be used to cluster a group of sequences of interest.

In addition to the alternative hypothesis that the two sequences share common motifs or CRMs, Reinert *et al.* [11] and Wan *et al.* [15] considered another alternative model H_1', called the pattern transfer model, which relates two sequences by randomly transferring patterns from one sequence to the other. Under this model, however, the power of D_2^S and D_2^* first increases with sequence length and then approximates a value less than 1. Therefore, neither is appropriate for testing the alternative hypothesis H_1'. To overcome this problem, Liu *et al.* [45] developed two new statistics, T_{sum}^S and T_{sum}^*, corresponding to D_2^S and D_2^*, respectively. To define T_{sum}^S, the maximum similarity measured by D_2^S between the fragment from i to $i+W-1$ in sequence **A** and any fragment of length W in sequence **B** is calculated and it is denoted as X_i^S. Similarly, let Y_j^S be the maximum similarity measured by D_2^S between the fragment from j to $j+W-1$ in sequence **B** and any fragments of length W in sequence **A**. Then

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

The statistic T_{sum}^* can be similarly defined by replacing D_2^S with D_2^*.

It was shown in [45] that the power of both T_{sum}^S and T_{sum}^* under the alternative model H_1' increases with sequence length and approximates 1 as sequence length tends to infinity.

Consideration of mismatched tuples

In the above studies, exact matches of the tuples are needed. During evolution, mutations can occur and hence it is natural to consider tuple matches allowing a giving number of mismatches. Also because genomic orientations of CRMs are usually unknown, both strands of the sequences need to be taken into account for alignment-free genome comparison. Thus, another line of extension of the statistics discussed above is the consideration of reverse complement and mismatched tuples [43]. For each tuple w, its neighbourhood, $\varsigma(w)$, is defined as the set of tuples with up to a certain number of mismatches with w. Each tuple w' in the neighbourhood $\varsigma(w)$ is associated with a weight $a_{w'}$. The weighted tuple neighbourhood count $X_{\varsigma(w)}$ for every tuple w in sequence **A** is defined by

$$X_{\varsigma(w)} = \sum\nolimits_{w'\in\varsigma(w)} a_{w'} X_{w'}.$$

The corresponding extensions of d_2^S and d_2^* in a mismatch model are given as

$$d_2^S = \frac{1}{2}\left(1 - \frac{\sum_w \frac{\widetilde{X}_{\varsigma(w)}\widetilde{Y}_{\varsigma(w)}}{\sqrt{\widetilde{X}_{\varsigma(w)}^2+\widetilde{Y}_{\varsigma(w)}^2}}}{\sqrt{\sum_w \frac{\widetilde{X}_{\varsigma(w)}^2}{\sqrt{\widetilde{X}_{\varsigma(w)}^2+\widetilde{Y}_{\varsigma(w)}^2}}}\sqrt{\sum_w \frac{\widetilde{Y}_{\varsigma(w)}^2}{\sqrt{\widetilde{X}_{\varsigma(w)}^2+\widetilde{Y}_{\varsigma(w)}^2}}}}\right) \quad (9)$$

and

$$d_2^* = \frac{1}{2}\left(1 - \frac{\sum_w \frac{\widetilde{X}_{\varsigma(w)}\widetilde{Y}_{\varsigma(w)}}{\sqrt{EX_{\varsigma(w)}}\sqrt{EY_{\varsigma(w)}}}}{\sqrt{\sum_w \frac{\widetilde{X}_{\varsigma(w)}^2}{EX_{\varsigma(w)}}}\sqrt{\sum_w \frac{\widetilde{Y}_{\varsigma(w)}^2}{EY_{\varsigma(w)}}}}\right). \quad (10)$$

Here $\widetilde{X}_{\varsigma(w)} = X_{\varsigma(w)} - EX_{\varsigma(w)}$ and $\widetilde{Y}_{\varsigma(w)}$ is defined analogously. Göke *et al.* [43] also proposed another similarity measure N_2 defined as

$$N_2 = \frac{\sum_w \frac{\widetilde{X}_{\varsigma(w)}\widetilde{Y}_{\varsigma(w)}}{\sqrt{\mathrm{var}(X_{\varsigma(w)})}\sqrt{\mathrm{var}(Y_{\varsigma(w)})}}}{\sqrt{\sum_w \frac{\widetilde{X}^2_{\varsigma(w)}}{\mathrm{var}(X_{\varsigma(w)})}}\sqrt{\sum_w \frac{\widetilde{Y}^2_{\varsigma(w)}}{\mathrm{var}(Y_{\varsigma(w)})}}} \quad (11)$$

where $\mathrm{var}(X_{\varsigma(w)})$, the variance of $X_{\varsigma(w)}$, can be obtained as in [43]. The dissimilarity measure corresponding to N_2 is defined by $n_2 = (1 - N_2)/2$.

The idea of using mismatched tuples can be applied to other dissimilarity measures such as the Jensen-Shannon divergence, *Hao*, and the l_p-distance between relative abundance vectors of two different sequences. The details are omitted here.

Empirical comparison of dissimilarity measures with mismatches

Because the consideration of mismatched tuples in alignment-free sequence comparison is relatively new, comprehensive evaluations of various dissimilarity measures allowing mismatched tuples have not been carried out yet. Here we use the comparison of CRM sequences as an example to study the effect of mismatch weight on *Hao*, n_2, d_2^* and d_2^S. We also consider two more statistics: the Jensen-Shannon information (*JS*) and *S2* in [42] discussed above.

In our implementation, we combine both the reverse complement and one-word mismatches in [43], and the neighbourhood of a word w can be defined as

$$\varsigma(w) = \{w', rc(w') | dist_{ham\,\min g}(w, w') \leq 1\},$$

where $rc(w)$ is the reverse complement of w, and $dist_{ham\,\min g}$ indicates the Hamming distance.

We used enhancers active in the forebrain, midbrain, limb and heart tissues of developing mouse embryo [43, 46, 47] to study the effectiveness of the different dissimilarity measures to cluster CRMs. These sequences form the set of *positive* samples. As in [12, 42], sequences of the same length as the positive samples are randomly chosen from the mouse genome, ensuring a maximum of 30% of repetitive sequence. These sequences form the set of *negative* samples. As in [43], we randomly chose 500 sequences in both the positive set and the negative set. Each pair of sequences in the positive set was compared and so was each pair in the negative set using the dissimilarity measures described above. We tested if sequence pairs from the positive set were more similar than sequence pairs from the negative set.

For a given dissimilarity measure with a fixed mismatch weight, we calculated the area under the receiver operating characteristic curve (AUC-ROC) as follows. First, we randomly chose 500 positive sequences from the CRM sequences and 500 random sequences from the mouse genome as negative sequences. Second, we calculated the dissimilarity measures for all pairs of positive sequences and all pairs of negative sequences. Third, all the dissimilarity scores for the positive pairs and negative pairs were mixed together. If the dissimilarity score of a pair was lower than a given threshold, the pair was predicted as from the positive samples and otherwise, the pair was predicted as from the negative samples. Fourth, the predictions were compared with the real data to find the false positive rate and the true positive rate. By changing the threshold, the ROC curve can be plotted and thus the AUC score can be calculated. We repeated these four steps 25 times. We let the tuple size to be 4, 5 and 6, the order of MC to be 0, 1 and 2 and the mismatch weight to be 0, 0.001, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75 and 1.00 as in [43]. Supplementary Tables S1–S4 give the average AUC-ROC scores for all the six dissimilarity measures for the four sets of CRM sequences.

We make the following three observations. First, the performances of n_2, d_2^* and d_2^S are close and are better than that of JS and *Hao* in general. Second, for $k = 4$, the performances of n_2, d_2^* and d_2^S are the best when the mismatch weight is around 0.05, but the differences with respect to different mismatch weights are negligible. For $k = 5$ or 6, the optimal performances are observed for the mismatch weight close to 1. These observations are consistent with that in [43]. Third, the statistic *S2* of Dai *et al.* [42] performs surprisingly well when $k = 6$ with independent identically distributed model for the background sequences. One potential explanation for the good performance of *S2* is that it does not consider k-tuples that are not present in the sequence. The number of such 6-tuples is large for CRM sequences of length around 1 kb.

ALIGNMENT-FREE GENOME AND METAGENOME COMPARISON USING NGS DATA

With the development of NGS technologies, many short sequence reads can be easily generated resulting

in a huge amount of available sequencing data. Yet, the assembly of these sequence reads can be challenging owing to their short length. Sequence-signature–based methods for molecular sequence data analysis have the advantage over alignment-based methods in that they can be directly applied to NGS data. The sequence reads are homogeneously or heterogeneously sampled from the original genome. Here *homogeneous* means that the sequence reads start at each position of the genome with equal probability, while *heterogeneous* means that some regions may be preferentially sequenced.

To overcome these challenges, Song *et al.* [48] extended D_2, D_2^S and D_2^* to develop new statistics to be applicable for NGS data and studied their corresponding power by both simulations and theoretical studies. Jiang *et al.* [28] used sequence signature methods to cluster microbial communities. The theoretical studies of the power of D_2, D_2^S and D_2^* are based on the limiting distributions of tuple counts in NGS data [49]. It was shown that the qualitative relative performances of these statistics on a large set of short reads are the same as that for long sequences. However, owing to the additional randomness involved in the sampling of the reads from the genomes during NGS, the power of these statistics is lower than when the complete genomic sequences are known. When the sampling of reads is homogeneously distributed, it was shown that, as the sequencing depth increases, the power of these statistics for NGS data approximates that when the genome sequences are known. In addition, heterogeneous sampling of the reads along the genome can further decrease the power of D_2^S and D_2^*. On the other hand, NGS read length does not significantly affect their power. Further, the dissimilarity measures d_2, d_2^S and d_2^* were then extended to NGS data and were applied to cluster different tree species with unknown complete genome sequences [48, 50]. It was shown that among these three dissimilarity measures the resulting clustering tree based on the dissimilarity measure d_2^S is the most consistent with the characteristics of the trees [48].

The three dissimilarity measures d_2, d_2^S and d_2^* were also used to compare complex microbial communities consisting of hundreds to thousands of species [28]; each microbial community was treated as a pan-genome. These dissimilarity measures were also compared with several other dissimilarity measures including the *Hao* dissimilarity measure defined in Equation (4), measures based on relative di-, tri- and tetra-nucleotide frequencies as in Equation (6) [36], and the standard l_p-measures between the frequency vectors.

Simulation studies were first used to see if these dissimilarity measures can recover the relationships among microbial communities using metagenomic short read data. Two types of relationships among the microbial communities were studied. First, they can be related through group relationships such as when the communities are divided into several groups where communities in each group have similar microbial species compositions. For any given dissimilarity measure, the dissimilarity between any pair of microbial communities can be calculated to form a dissimilarity matrix. Hierarchical clustering with average linkage was then used to cluster the microbial communities based on the dissimilarity matrix. The resulting clusters were compared with the simulated group relationships of the microbial communities. It was shown that among the three dissimilarity measures tested, the clustering tree derived based on d_2^S is the most similar to the simulated group relationships among the microbial communities. Second, the microbial communities can be related through a gradient relationship for the abundance levels of microbial organisms. Principal coordinate analysis was then applied to the dissimilarity matrix. The correlation between the principal coordinate and the gradient can be calculated; high correlation indicates better performance of a dissimilarity measure. It was shown that when sequencing depth is low to moderate (1000–10 000 reads per community), the correlation is highest when based on d_2^* and d_2^S. When the sequencing depth is high, the correlations based on *Hao*, d_2^* and d_2^S are similar and are higher than those based on other dissimilarity measures.

These dissimilarity measures were also used to analyse 39 fecal samples from 33 mammalian host species [51], 56 marine samples across the world [52] and 13 fecal samples from human individuals [53]. Using the d_2^S dissimilarity measure, the fecal samples from carnivores can be separated from that of the herbivores. Even within the herbivores, the fecal samples from the hindgut-fermenting herbivores can be separated from that of the foregut-fermenting herbivores. In contrast, the fecal samples from the omnivore samples are diverse and are mixed together with the carnivore and herbivore samples. For the marine data, metagenomic samples from the same region tend to cluster together. For

the human gut samples, the metagenomic samples from the adult can be separated from that of unweaned infants. These studies showed the importance of using alignment-free methods for the comparison of metagenomic samples.

In [28], the dissimilarity measure *S2* of Dai *et al.* [42] was not evaluated with respect to the classification of microbial communities. Here we carried out the same analysis as in [28] with the three real data sets described above using *S*2, and the parsimony scores of the resulting trees are given in Supplementary Table S5 for the fecal samples of mammalian host species, Tables S6 and S7 for the open and coastal water samples, respectively, and Table S8 for the human fecal samples in the Supplementary Material, where the parsimony score is the minimum number of changes of the community labels needed to explain the clusters. The lower the parsimony score, the better the dissimilarity score is. In all these tables, only the results related to S2 are new and the results related to other statistics were presented in [28]. We found that for the comparison of microbial communities, d_2^S outperforms *S*2, and d_2^S has the best performance among all the dissimilarity measures we studied so far.

The programs for calculating most of the statistics described in this review are available online and the corresponding websites are given in Table 1.

DISCUSSION AND CONCLUSIONS

Alignment-based approaches for sequence comparison will continue to play dominant roles in genomic studies when the sequences of interest can be reliably aligned. On the other hand, alignment-free sequence comparison approaches have been shown to provide important information on the evolution of gene regulatory regions and the comparison of genomes and metagenomes in situations where reliable alignments are not available. Alignment-free approaches based on *k*-tuple counts are especially powerful for the analysis of NGS short read data from unknown genomes and metagenomes.

In this review, we summarized recent developments of alignment-free sequence comparison concentrating on approaches based on *k*-tuple count vectors. We emphasized the dissimilarity measures used for genome and metagenome comparisons, the statistical distributions of the measures and the power of these statistics when genomes of interest are related. Both theoretical studies and real data analysis showed that the newly developed statistics D_2^S and D_2^* are generally more powerful than the original statistic D_2. The introduction of mismatched tuples in these statistics can further increase their power. For the comparison of relatively short (e.g. 1 kb) CRM sequences, the statistic *S2* seems to perform well with appropriate choices of tuple length and order of MC for the sequences. For the comparison of long genome sequences and metagenomic communities, the d_2^S dissimilarity measure seems to yield the best results.

Many other alignment-free sequence comparison approaches are available including those based on chaos game representation [54], common substrings between sequences [55], longest common words [56], the minimal words [57], the difference between the longest common and shortest absent words [58] and sequence representation based on natural vectors [59]. Also, there are recent results from the area of machine learning applied to alignment-free sequence comparison, see for example [60]. Because these approaches are based on different philosophies of sequence comparison from the *k*-tuple–based methods and it is not clear whether they can be applied to NGS data, we did not review them here. Further studies are needed to understand the advantages and disadvantages of the various alignment-free sequence comparison methods.

Table 1: The statistics reviewed in this article and the websites for the software to calculate the corresponding statistics

Statistical measure	Website
D2, D2S, D2Star	http://www-rcf.usc.edu/~fsun/Programs/D2/d2-all.html
d2, d2S, d2Star	http://www-rcf.usc.edu/~fsun/Programs/D2NGS/D2NGSmain.html
D2z	http://veda.cs.uiuc.edu/cgi-bin/d2z/download.pl
Hao	http://tlife.fudan.edu.cn/cvtree/
S2	http://math.dlut.edu.cn/daiqi/mplusd.html
N2	http://www.seqan.de/projects/alf/

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- Alignment-free sequence comparison is essential for the comparison of genomes based on NGS reads even if the reads are not from homologous regions.
- Metagenomes can be effectively clustered based on NGS reads using sequence signatures.
- Normalization of pattern counts by centralizing around their means can significantly increase the power of alignment-free genome comparison.
- The newly developed dissimilarity measures d_2^* and d_2^S outperform others for genome and metagenome comparison.

Acknowledgements

We thank Dr Xuegong Zhang and Mr Bai Jiang for collaboration on the use of sequence signatures to compare microbial communities.

FUNDING

This work was supported by the US National Institutes of Health [P50HG002790, R21HG006199] and NSF [DMS-1043075, OCE 1136818]. M.D., K.S. and J.R. are supported by the National Natural Science Foundation of China [31171262, 11021463] and the National Key Basic Research Project of China [2009CB918503].

References

1. Smith TF, Waterman MS. Comparison of biosequences. *Adv Appl Math* 1981;**2**:482–9.
2. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
3. Blaisdell BE. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc Natl Acad Sci USA* 1986;**83**:5155–9.
4. Blaisdell BE. Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding. *J Mol Evol* 1985;**21**:278–88.
5. Vinga S, Almeida J. Alignment-free sequence comparison - a review. *Bioinformatics* 2003;**19**:513–23.
6. Vinga S. Biological sequence analysis by vector-valued functions: revisiting alignment-free methodologies for DNA and protein classification. In: *Advanced Computational Methods for Biocomputing and Bioimaging*. New York: Nova Science Publishers, 2007;71–107.
7. Torney DC, Burks C, Davison D, *et al*. Computation of d2: a measure of sequence dissimilarity. In: Bell GI, Marr TG (eds). Computers and DNA1990:109–125: the Proceedings of the Interface between Computation Science and Nucleic Acid Sequencing Workshop, December 12–16, 1988 in Santa Fe, New Mexico, USA.
8. Hide W, Burke J, Davison DB. Biological evaluation of d2, an algorithm for high-performance sequence comparison. *J Comput Biol* 1994;**1**:199–215.
9. Miller RT, Christoffels AG, Gopalakrishnan C, *et al*. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res* 1999;**9**:1143–55.
10. Lippert RA, Huang HY, Waterman MS. Distributional regimes for the number of k-word matches between two random sequences. *Proc Natl Acad Sci USA* 2002;**99**:13980–9.
11. Reinert G, Chew D, Sun F, *et al*. Alignment-free sequence comparison (I): statistics and power. *J Comput Biol* 2009;**16**:1615–34.
12. Kantorovitz MR, Robinson GE, Sinha S. A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics* 2007;**23**:I249–55.
13. Foret S, Wilson SR, Burden CJ. Empirical distribution of k-word matches in biological sequences. *Pattern Recognit* 2009;**42**:539–548.
14. Burden CJ, Kantorovitz MR, Wilson SR. Approximate word matches between two random sequences. *Ann Appl Probab* 2008;**18**:1–21.
15. Wan L, Reinert G, Sun F, *et al*. Alignment-free sequence comparison (II): theoretical power of comparison statistics. *J Comput Biol* 2010;**17**:1467–90.
16. Zhai Z, Ku S-Y, Luan Y, *et al*. The power of detecting enriched patterns: an HMM approach. *J Comput Biol* 2010;**17**:581–92.
17. Wu TJ, Hsieh YC, Li LA. Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. *Biometrics* 2001;**57**:441–8.
18. Van Helden J. Metrics for comparing regulatory sequences on the basis of pattern counts. *Bioinformatics* 2004;**20**:399–406.
19. Xu Z, Hao BL. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic Acids Res* 2009;**37**:W174–8.
20. Qi J, Luo H, Hao BL. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res* 2004;**32**:W45–7.
21. Qi J, Wang B, Hao BL. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J Mol Evol* 2004;**58**:1–11.
22. Gao L, Qi J. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. *BMC Evol Biol* 2007;**7**(1):41.
23. Wang H, Xu Z, Gao L, *et al*. A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC Evol Biol* 2009;**9**(1):195.
24. Qi J, Luo H, Hao BL. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res* 2004;**32**:W45–7.
25. Foret S, Wilson SR, Burden CJ. Characterizing the D2 statistic: word matches in biological sequences. *Stat Appl Genet Mol Biol* 2009;**8**(1):1–21.
26. Li Q, Xu Z, Hao BL. Composition vector approach to whole-genome-based prokaryotic phylogeny: success and foundations. *J Biotechnol* 2010;**149**:115–19.

27. Hua WY, Xu Z, Zhang MH, *et al*. The application of CVTree in structural analysis of microbial communities by 454 pyrosequencing. *Chin J Microecol* 2010;**22**(4): 312–16.
28. Jiang B, Song K, Ren J, *et al*. Comparison of metagenomic samples using sequence signatures. *BMC Genomics* 2012; **13**(1):730.
29. Mrázek J, Karlin S. Distinctive features of large complex virus genomes and proteomes. *Proc Natl Acad Sci USA* 2007;**104**:5127–32.
30. Karlin S, Mrazek J. Compositional differences within and between eukaryotic genomes. *Proc Natl Acad Sci USA* 1997; **94**:10227–32.
31. Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* 1995;**11**:283–90.
32. Gentles AJ, Karlin S. Genome-scale compositional comparisons in eukaryotes. *Genome Res* 2001;**11**:540–6.
33. Burge C, Campbell AM, Karlin S. Over-and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci* 1992;**89**:1358–62.
34. Campbell A, Mrazek J, Karlin S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc Natl Acad Sci USA* 1999;**96**:9184–89.
35. Karlin S, Mrazek J, Campbell AM. Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* 1997;**179**:3899–913.
36. Willner D, Thurber RV, Rohwer F. Metagenomic signatures of 86 microbial and viral metagenomes. *Environ Microbiol* 2009;**11**:1752–66.
37. Jun SR, Sims GE, Wu GA, *et al*. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: an alignment-free method with optimal feature resolution. *Proc Natl Acad Sci USA* 2010;**107**:133–8.
38. Wu GA, Jun SR, Sims GE, *et al*. Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. *Proc Natl Acad Sci USA* 2009;**106**:12826–31.
39. Sims GE, Jun SR, Wu GA, *et al*. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc Natl Acad Sci USA* 2009;**106**: 2677–82.
40. Sims GE, Kim SH. Whole-genome phylogeny of *Escherichia coli*/Shigella group by feature frequency profiles (FFPs). *Proc Natl Acad Sci USA* 2011;**108**:8329–34.
41. Dai Q, Wang T. Comparison study on k-word statistical measures for protein: from sequence to 'sequence space'. *BMC Bioinformatics* 2008;**9**(1):394.
42. Dai Q, Yang Y, Wang T. Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. *Bioinformatics* 2008;**24**: 2296–302.
43. Göke J, Schulz MH, Lasserre J, *et al*. Estimation of pairwise sequence similarity of mammalian enhancers with word neighbourhood counts. *Bioinformatics* 2012;**28**:656–63.
44. Shepp L. Normal functions of normal random variables. *SIAM Rev* 1964;**6**(4):459–460.
45. Liu X, Wan L, Li J, *et al*. New powerful statistics for alignment-free sequence comparison under a pattern transfer model. *Journal of Theoretical Biology* 2011;**284**:106–16.
46. Blow MJ, McCulley DJ, Li Z, *et al*. ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* 2010;**42**: 806–10.
47. Visel A, Blow MJ, Li Z, *et al*. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 2009;**457**:854–8.
48. Song K, Ren J, Zhai Z, *et al*. Alignment-free sequence comparison based on next-generation sequencing reads. *J Comput Biol* 2013;**20**:64–79.
49. Zhai Z, Reinert G, Song K, *et al*. Normal and compound Poisson approximations for pattern occurrences in NGS reads. *J Comput Biol* 2012;**19**:839–854.
50. Cannon CH, Kua CS, Zhang D, *et al*. Assembly free comparative genomics of short-read sequence data discovers the needles in the haystack. *Mol Ecol* 2010;**19**:147–61.
51. Muegge BD, Kuczynski J, Knights D, *et al*. Diet Drives Convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* 2011;**332**:970–4.
52. Rusch DB, Halpern AL, Sutton G, *et al*. The Sorcerer II Global ocean sampling expedition: northwest atlantic through eastern tropical pacific. *PloS Biol* 2007;**5**:398–431.
53. Kurokawa K, Itoh T, Kuwahara T, *et al*. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* 2007;**14**:169–81.
54. Jeffrey HJ. Chaos game representation of gene structure. *Nucleic Acids Res* 1990;**18**:2163–70.
55. Ulitsky I, Burstein D, Tuller T, *et al*. The average common substring approach to phylogenomic reconstruction. *J Comput Biol* 2006;**13**:336–50.
56. Haubold B, Pierstorff N, Möller F, *et al*. Genome comparison without alignment using shortest unique substrings. *BMC Bioinformatics* 2005;**6**:123.
57. Pinho AJ, Ferreira PJ, Garcia SP, *et al*. On finding minimal absent words. *BMC Bioinformatics* 2009;**10**:137.
58. Yang L, Zhang X, Wang T, *et al*. Large local analysis of the unaligned genome and its application. *J Comput Biol* 2013; **20**:19–29.
59. Zhao B, He RL, Yau SST. A new distribution vector and its application in genome clustering. *Mol Phylogenet Evol* 2011; **59**:438–443.
60. Didier G, Corel E, Laprevotte I, *et al*. Variable length local decoding and alignment-free sequence comparison. *Theor Comput Sci* 2012;**462**:1–11.