The number of k-mer matches between two *DNA* sequences as a function of k

Sophie Röhling[1], Thomas Dencker[1], and Burkhard Morgenstern[1,2]

[1]University of Göttingen, Department of Bioinformatics, Goldschmidtstr. 1, 37077 Göttingen, Germany
[2]Göttingen Center of Molecular Biosciences (GZMB), Justus-von-Liebig-Weg 11, 37077 Göttingen, Germany

April 30, 2019

Abstract

We study the number N_k of length-k word matches between pairs of evolutionarily related DNA sequences depending on k. We show that the number of substitutions per site that occurred since two sequences evolved from their last common ancestor, can be estimated from the slope of a certain function of N_k. This approach can be generalized from contiguous word matches to so-called *spaced-word* matches, where mismatches are allowed at certain pre-defined positions. Based on these theoretical results, we implemented a software program for alignment-free sequence comparison called *Slope-SpaM*. Test runs on simulated sequence data show that *Slope-SpaM* can accurately estimate phylogenetic distances for distance values up to around 0.5 substitutions per position. The statistical stability of our results is improved if spaced words are used instead of contiguous words. Unlike previous methods that are based on the number of (spaced) word matches, *Slope-SpaM* produces accurate results, even if sequences share only local homologies.
Availability: `https://github.com/burkhard-morgenstern/Slope-SpaM`
Contact: bmorgen@gwdg.de

1 Introduction

Phylogeny reconstruction is a fundamental task in computational biology [15]. Here, a basic step is to estimate pairwise evolutionary distances between protein or nucleic-acid sequences. Under the *Jukes-Cantor* model of evolution [25], the distance between two evolutionarily related *DNA* sequences can be defined as the (estimated) number of nucleotide substitutions per site that have occurred since the two sequences have evolved from their last common ancestor. Traditionally, phylogenetic distances are inferred from pairwise or multiple sequence alignments. For the huge amounts of sequence data that are now available, however, sequence alignment has become far too slow. Therefore, considerable efforts have been made in recent years, to develop fast *alignment-free* approaches that can estimate phylogenetic distances without the need to calculate full alignments of the input sequences, see [20, 49, 56, 5, 26] for recent review articles.

A number of alignment-free approaches are able to use unassembled short sequencing reads as input [53, 2, 14, 40, 34, 27, 4, 46]. Alignment-free approaches are not only used in phylogeny reconstruction, but are also important in metagenomics [11, 40, 33] and in medical applications, for example to identify drug-resistant bacteria [6] or to classify viruses [54, 3]. In all these applications, it is crucial to rapidly estimate the degree of similarity or dissimilarity between large sets of sequence data.

Some alignment-free approaches are based on word frequencies [41, 47] or on the length of common substrings [51, 29, 23]. Other methods use variants of the D_2 distance which is defined as the number of word matches of a pre-defined length between two sequences [42, 52, 50, 3]; a review focusing on these methods is given in [43]. *kWIP* [38] is a further development of this concept that uses information-theoretical weighting. Most of these approaches calculate heuristic measures of sequence (dis-)similarity that may not be easy to interpret. At the same time, alignment-free methods have been proposed that can accurately estimate phylogenetic distances between sequences based on stochastic models of DNA or protein evolution, using the length of common substrings [22, 36] or so-called *micro alignments* [53, 21, 31, 30].

Several authors have proposed to estimate phylogenetic distances from the number of k-mer matches between two sequences. The tools *Cnidaria* [2] and *AAF* [14] use the *Jaccard index* between sets of k-mers from two – assembled or unassembled – genomes to estimate the

distance between them. In *Mash* [40], the *MinHash* [7] technique is used to reduce the input sequences to small 'sketches' which can be used to rapidly approximate the *Jaccard index*. In a previous article, we proposed another way to infer evolutionary distances between DNA sequences based on the number of word matches between them, and we generalized this to so-called *spaced-word* matches [37]. This distance function is now used by default in the program *Spaced* [28].

A *spaced-word match* is a pair of words from two sequences that are identical at certain positions, specified by a pre-defined binary pattern of *match* and *don't-care* positions. Theoretically, the distance measure proposed in [37] is based on a simple model of molecular evolution without insertions or deletions. In particular, we assumed in this previous paper that the compared sequences are homologous to each other over their entire length. In practice, the derived distance values are still reasonably accurate if a limited number of insertions and deletions is allowed, and phylogenetic trees could be obtained from these distance values that are similar to trees obtained with more traditional approaches. Like other methods that are based on the number of common k-mers, however, our previous approach can no longer produce accurate results for sequences that share only local regions of homology.

Recently, Bromberg *et al.* published an interesting new approach to alignment-free protein sequence comparison that they called *Slope Tree* [8]. They defined a distance measure using the decay of the number of k-mer matches between two sequences, as a function of k. Trees reconstructed with this distance measure were in accordance to known phylogenetic trees for various sets of prokaryotes. *Slope Tree* can also correct for horizontal gene transfer and can deal with composition variation and low complexity sequences. From a theoretical point-of-view, however, it is not clear if the distance measure used in *Slope Tree* is an accurate estimator of evolutionary distances.

In the present paper, we study the number N_k of word or spaced-word matches between two DNA sequences, where k is the word length or the number of *match positions* of the underlying pattern for spaced words, respectively. Inspired by Bromberg's *Slope Tree* approach, we study the decay of N_k as a function of k. More precisely, we define a function $F(k)$ that depends on N_k and that can be approximated – under a simple probabilistic model of DNA evolution, and for a certain range of k – by an affine-linear function of k. The number of substitutions per site can be estimated from the slope of F, we therefore call

our implementation *Slope-SpaM*, where *SpaM* stands for *Spaced-Word Matches*. Using simulated DNA sequences, we show that *Slope-SpaM* can accurately estimate phylogenetic distances. In contrast to other methods that are based on the number of (spaced) word matches, *Slope-SpaM* produces still accurate results if the compared sequences share only local homologies. We also applied *Slope-SpaM* to infer phylogenetic trees based on genome sequences from the benchmarking project *AFproject* [1].

2 The number of k-mer matches as a function of k

We are using standard notation from stringology as used, for example, in [18]. For a string or sequence S over some alphabet $\mathcal{A}$, $|S|$ denotes the length of S, and $S(i)$ is the i-th character of S, $1 \leq i \leq |S|$. $S[i..j]$ is the (contiguous) substring of S from i to j. We consider a pair of DNA sequences S_1 and S_2 that have evolved under the *Jukes-Cantor* substitution model [25] from some unknown ancestral sequence. That is, we assume that substitution rates are equal for all nucleotides and sequence positions, and that substitution events at different positions are independent of each other. For simplicity, we first assume that there are no insertions and deletions (indels). We call a pair of positions or k-mers from S_1 and S_2, respectively, *homologous* if they go back to the same position or k-mer in the ancestral sequence. In our model, we have a nucleotide match probability p for homologous positions and a background match probability q for non-homologous nucleotides; the probability of two homologous k-mers to match exactly is p^k.

In our indel-free model, S_1 and S_2 must have the same length $L = |S_1| = |S_2|$, and positions i_1 and i_2 in S_1 and S_2, respectively, are homologous if and only if $i_1 = i_2$. Note that, under this model, a pair of k-mers from S_1 and S_2 is either homologous or completely non-homologous, in the sense that *none* of the corresponding pairs of positions is homologous, and for a pair of non-homologous k-mers from S_1 and S_2, the probability of an exact match is q^k. Let the random variable X_k be defined as the number of k-mer matches between S_1 and S_2. More precisely, X_k is defined as the number of pairs (i_1, i_2) for which

$$S_1[i_1..i_1 + k - 1] = S_2[i_2..i_2 + k - 1]$$

holds. There are $(L-k+1)$ possible homologous and $(L-k+1)\cdot(L-k)$ possible background k-mer matches, so the expected total number of k-mer matches is

$$E(X_k) = (L-k+1)\cdot p^k + (L-k+1)\cdot(L-k)\cdot q^k \qquad (1)$$

In [37], we used this expression directly to estimate the match probability p for two observed sequences using a *moment-based* approach, by replacing the expected number $E(X_k)$ by the empirical number N_k of word matches or spaced-word matches, respectively. Although in equation (1), an indel-free model is assumed, we could show in our previous paper, that this approach gives still reasonable estimates of p for sequences with insertions and deletions, as long as the sequences are globally related, *i.e.* as long as the insertions and deletions are small compared to the length of the sequences. It is clear, however, that this estimate will become inaccurate in the presence of large insertions and deletions, *i.e.* if sequences are only locally related.

Herein, we propose a different approach to estimate evolutionary distances from the number N_k of k-mer matches, by considering the *decay* of N_k if k increases. For simplicity, we first consider an indel-free model as above. From equation (1), we obtain

$$\ln p \cdot k + \ln(L-k+1) = \ln\Big(E(X_k) - (L-k+1)\cdot(L-k)\cdot q^k\Big) \qquad (2)$$

which – for a suitable range of k – is an approximately affine-linear function of k with slope $\ln p$. Substituting the expected value $E(X_k)$ in the right-hand side of (2) with the corresponding *empirical* number N_k of k-mer matches for two observed sequences, we define

$$F(k) = \ln\Big(N_k - (L-k+1)\cdot(L-k)\cdot q^k\Big)$$

In principle, we can now estimate p as the exponential of the slope of F. Note that, in practice, we will have to restrict ourselves to a certain range of k. If k is too small, N_k will be dominated by background word matches, if k is too large, no word matches will be found at all.

If we want there to be at least as many homologous as background word matches, we have to require $N\cdot p^k \geq N^2\cdot q^k$. We therefore obtain a lower bound for k as

$$\frac{\ln N}{\ln \frac{p}{q}} \leq k$$

If, on the other hand, we want to have at least one expected word match, we have $N \cdot p^k \geq 1$, so k would be upper-bounded by

$$k \leq -\frac{\ln N}{\ln p}$$

Therfore, if we want to have a range of length K between the lower and the upper bound for k, we must have

$$\frac{\ln N}{\ln \frac{p}{q}} + \frac{\ln N}{\ln p} \leq -K$$

It follows that the match probability p must be large enough for our approach to work. As an example, with $q = 1/4$ and a range of, say, $K = 15$, we would need $p > 0.52$, corresponding to a *Jukes-Cantor* distance of 0.53 substitutions per position.

The above considerations can be generalized to a model of *DNA* evolution with insertions and deletions. Let us consider two *DNA* sequences S_1 and S_2 of different lengths L_1 and L_2, respectively, that have evolved from a common ancestor under the *Jukes-Cantor* model, this time with insertions and deletions. Note that, unlike with the indel-free model, it is now possible that a k-mer match involves homologous as well as background nucleotide matches. Instead of deriving exact equations like (1) and (2), we therefore make some simplifications and approximations.

We can decompose S_1 and S_2 into indel-free pairs of 'homologous' substrings that are separated by non-homologous segments of the sequences. Let L_H be the total length of the homologous substring pairs in each sequence. As above, we define the random variable X_k as the number of k-mer matches between the two sequences, and p and q are, again, the homologous and background nucleotide match probabilities. We then use

$$E(X_k) \approx L_H \cdot p^k + L_1 \cdot L_2 \cdot q^k \tag{3}$$

as a rough approximation, and we obtain

$$\ln p \cdot k + \ln L_H \approx \ln \left(E(X_k) - L_1 \cdot L_2 \cdot q^k \right) \tag{4}$$

Similar as in the indel-free case, we define

$$F(k) = \ln \left(N_k - L_1 \cdot L_2 \cdot q^k \right) \tag{5}$$

S_1 :		T	T	A	T	G	A	C	C	A	C	T	C
S_2 :	A	C	T	A	C	G	A	T	C	G	A		
P :			1	1	0	0	1	0	1				

Figure 1: Spaced-word match between two DNA sequences S_1 and S_2 at (2,3) with respect to a pattern $P = 1100101$ representing *match positions* ('1') and *don't-care positions* ('0'). The same spaced word $TA**A*C$ occurs at position 2 in S_1 and at position 3 in S_2.

where N_k is, again, the empirical number of k-mer matches. As above, we can estimate p as the exponential of the slope of F. Note that this estimation can still be applied if L_H is small compared to L_2 and L_2, since L_H appears only in an additive constant on the left-hand side of (4). Thus, while the absolute values $F(k)$ do depend on the extent L_H of the homology between S_1 and S_2, the *slope* of F only depends on p, q, L_1 and L_2, but not on L_H.

3 The number of spaced-word matches

In many fields of biological sequence analysis, k-mers and k-mer matches have been replaced by *spaced words* or *spaced-word matches*, respectively. Let us consider a fixed word P over $\{0, 1\}$ representing *match positions* ('1') and *don't-care* positions ('0'). We call such a word a *pattern*; the number of *match positions* in P is called its *weight*. In most applications, the first and the last symbol of a pattern P are assumed to be match positions. A *spaced word* with respect to P is defined as a string W over the alphabet $\{A, C, G, T, *\}$ of the same length as P with $W(i) = *$ if and only if $P(i) = 0$, i.e. if and only if i is a *don't-care position* of P. Here, '$*$' is interpreted as a *wildcard* symbol. We say that a spaced word W w.r.t P occurs in a sequence S at some position i if $W(m) = S(i + m - 1)$ for all match positions m of P.

We say that there is a *spaced-word match* between sequences S_1 and S_2 at (i, j) if the same spaced word occurs at position i in S_1 and at position j in S_2, see Fig. 1 for an example. A spaced-word match can therefore be seen as a gap-free alignment of length $|P|$ with matching characters at the *match positions* and possible mismatches a the *don't-care positions*. *Spaced-word matches* or *spaced seeds* have

been introduced in database searching as an alternative to exact k-mer matches [9, 35, 32]. The main advantage of spaced words compared to contiguous k-mers is the fact that results based on spaced words are statistically more stable than results based on k-mers [24, 13, 11, 10, 37, 39].

Quite obviously, approximations (3) and (4) remain valid if we define X_k to be the number of spaced-word matches for a given pattern P of weight k, and we can generalize the definition of $F(k)$ accordingly: if we consider a maximum pattern weight $k_{\max}$ and a given set of patterns $\{P_k, 1 \leq k \leq k_{\max}\}$ where k is the weight of pattern P_k, then we can define N_k as the empirical number of spaced-word matches with respect to pattern P_k between two observed DNA sequences. $F(k)$ can then be defined exactly as in (5), and we can estimate the nucleotide match probability p as the exponential of the slope of F.

4 Implementation

It is well known that the set of all k-mer matches between two sequences can be calculated efficiently by lexicographically *sorting* the list of all k-mers from both sequences, such that identical k-mers appear as contiguous *blocks* in this list. Note that, to find k-mer matches for all $k \leq k_{\max}$, it is sufficient to sort the list of $k_{\max}$-mers since this sorted list also contains the sorted lists of k-mers for $k < k_{\max}$.

This standard procedure can be directly generalized to spaced-word matches. In order to calculate the numbers N_k efficiently, we start by generating a suitable set of patterns $\{P_k, 1 \leq k \leq k_{\max}\}$. We first specify a pattern $P = P_{k_{\max}}$ of weight $k_{\max}$, for example by using the tool *rasbhari* [19]. As usual, we require the first and the last position of P to be match positions, i.e. we have $P(1) = P(k_{\max}) = 1$. Let ℓ_k be the k-th *match position* in the pattern P. We then define the k-th pattern as

$$P_k = P[1..\ell_k]$$

In other words, for $k < k_{\max}$, the pattern P_k is obtained by cutting off P after the k-th match position, so each pattern P_k has a weight of k. To find all spaced-word matches with respect to all patterns $P_k, 1 \leq i \leq k_{\max}$, it suffices to lexicographically sort the set of spaced words with respect to P in S_1 and S_2.

To estimate the slope of F, one has to take into account that the values $F(k)$ are statistically stable only for a certain range of k. If k is

too small, N_k and $F(k)$ are dominated by the number of background (spaced) word matches. If k is too large, on the other hand, the number of homologous (spaced) word matches becomes small, and for $N_k < L_1 \cdot L_2 \cdot q^k$, the value $F(k)$ is not even defined. For two input sequences, we therefore need to identify a range $k_1, \ldots, k_2$ in which F is approximately affine linear. To this end, we consider the differences

$$\delta_k = (F(k) - F(k-1))$$

and we choose a maximal range $k_1 \ldots k_2$ where the difference $|\delta_k - \delta_{k-1}|$ is smaller than some threshold T for all k with $k_1 < k \leq k_2$. In our implementation, we used a value of $T = 0.2$. We then estimate the slope of F as the *average* value of δ_k in this range, so we estimate the match probability as

$$\hat{p} = \exp\left(\frac{F(k_2) - F(k_1)}{k_2 - k_1}\right) \tag{6}$$

Finally, we apply the usual *Jukes-Cantor* correction to estimate the *Jukes-Cantor* distance between the sequences, e.g. the number of substitutions per position that have occurred since they diverged from their last common ancestor. Examples are given in Fig. 2, 3 and 4.

5 Test results

5.1 Distances calculated from simulated DNA sequences

We first simulated pairs of *DNA* sequences of length $L = 10,000$ without indels, and with different phylogenetic distances. As a first example, we generated a pair of sequences with a *Jukes-Cantor* distance of 0.2, i.e with 0.2 substitutions per position. According to the *Jukes-Cantor* correction formula, this corresponds to a match probability of $p = 0.82$. In Fig. 2 (top), $F(k)$ is plotted against k for these sequences based on exact k-mer matches; the difference $F(k) - F(k-1)$ is stable for $k_1 = 5 < k \leq k_2 = 20$. With equation (6), we therefore estimate the slope of F from the numbers N_k of k-mer matches as

$$\frac{F(k_2) - F(k_1)}{k_2 - k_1} = \frac{5.24 - 8.12}{20 - 5} = -0.19$$

This way, we obtain an estimated match probability of $\hat{p} = e^{-0.19} = 0.827$ which is rather close to the correct value $p = 0.82$. A similar

value was obtained when we used spaced-word matches instead of contiguous k-mer matches (Fig. 2, bottom). Results for indel-free sequences of the same length, with a *Jukes-Cantor* distance of 0.5 are shown in Fig. 3. Here, the slope of F was 0.449 when k-mer matches were used an 0.446 with spaced-word matches, leading to an estimated match probability of $\hat{p} = 0.638$ which corresponds to a *Jukes-Cantor* distance of 0.494, again very close to the true value.

Fig. 4 shows the corresponding results for a pair of simulated sequences with insertions and deletions and with only local sequence homology. In a region of length $L_H = 10,000$, the sequences are homologous to each other, with a *Jukes-Cantor* distance of 0.2 substitutions per position, as in our first example (Fig. 2) – but this time, we included 20 insertions and deletions of length 50 each at random positions. In addition, the homology region was flanked on both sides and in both sequences by non-related random sequences of length 10,000 each. So both sequences had a length of around 30,000 where one third of the sequences were homologous to each other. Here, *Slope-SpaM* estimated a nucleotide match frequency of $\hat{p} = 0.838$ corresponding to a *Jukes-Cantor* distance of 0.18, somewhat smaller than the true value of 0.2.

For a more systematic evaluation, we simulated pairs of *DNA* sequences with *Jukes-Cantor* distances d between 0.05 and 1.0. For each value of d, we generated 20,000 sequence pairs of length 10,000 each and with distance d, and we estimated the distances with *Slope-SpaM*. In Fig. 5, the average estimated distances are plotted against the real distances. Our estimates are fairly accurate for distances up to around 0.5 substitutions per site, but become statistically less stable for larger distances.

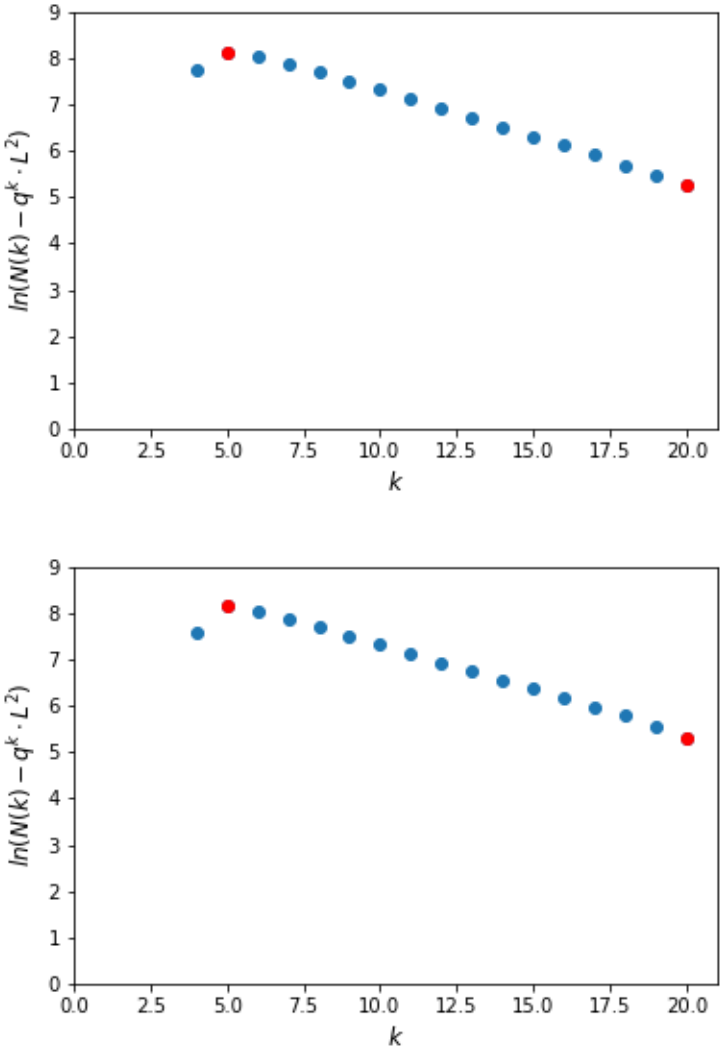

Figure 2: Test runs on simulated indel-free *DNA* sequences of length $L = 10,000$, with a *Jukes-Cantor* distance of 0.2. *Top:* $F(k) = \ln\left(N_k - L^2 \cdot q^k\right)$ plotted against k, where N_k is the number of k-mer matches. *Bottom:* for $P_{20} = 1100111011100111011100010101011011$, patterns P_k were generated for $k = 1, \ldots, 19$ as explained in section 4. $F(k)$ is plotted against k, based on the number N_k of spaced-word matches with respect to P_k. In both cases, $\delta_k = F(k) - F(k-1)$ is stable for $k_1 = 5 < k \leq k_2 = 20$ (red dots).

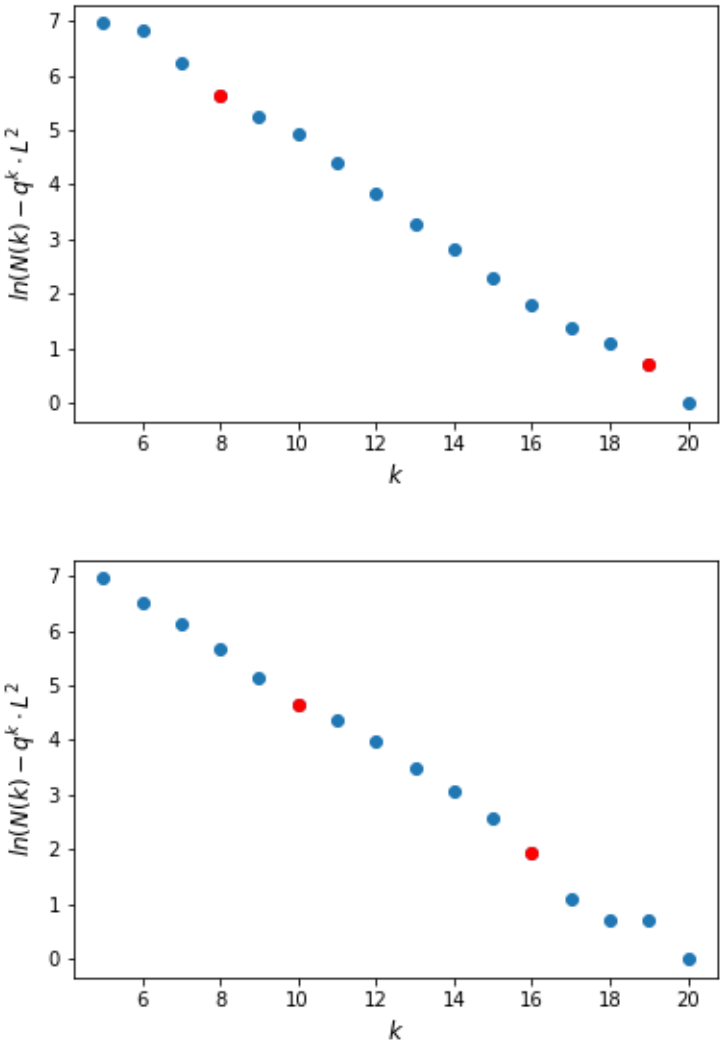

Figure 3: $F(k)$ plotted against k as in Fig. 2 and for simulated indel-free *DNA* sequences of length $L = 10,000$ with a *Jukes-Cantor* distance of 0.5 for contiguous k-mers (top) and for spaced-word matches (bottom).

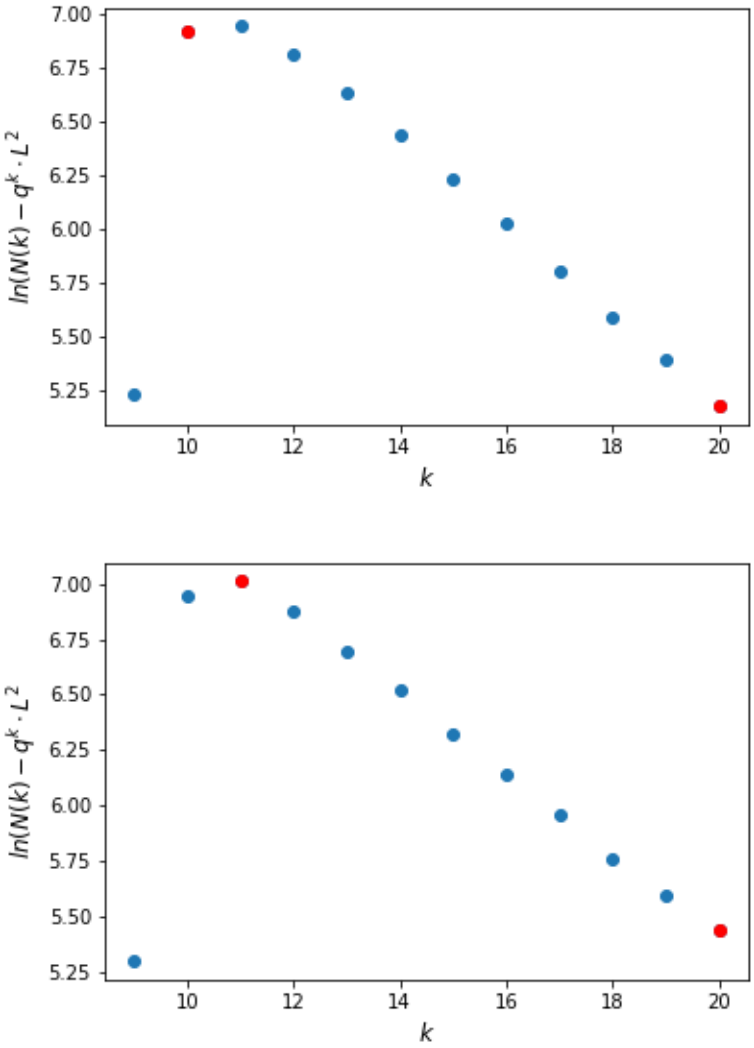

Figure 4: $F(k)$ plotted against k as in Fig. 2 and 3, for simulated indel-free *DNA* sequences of length $L = 30,000$. In a region of length $L_H = 10,000$, the two sequences are related to each other with a *Jukes-Cantor* distance of 0.2; this region contains around 20 indels at random positions. The 'homologous' region is flanked on both sides and in both sequences by unrelated random sequences of length 10,000 each. $F(k)$ is plotted for contiguous k-mers (top) and for spaced-word matches (bottom).

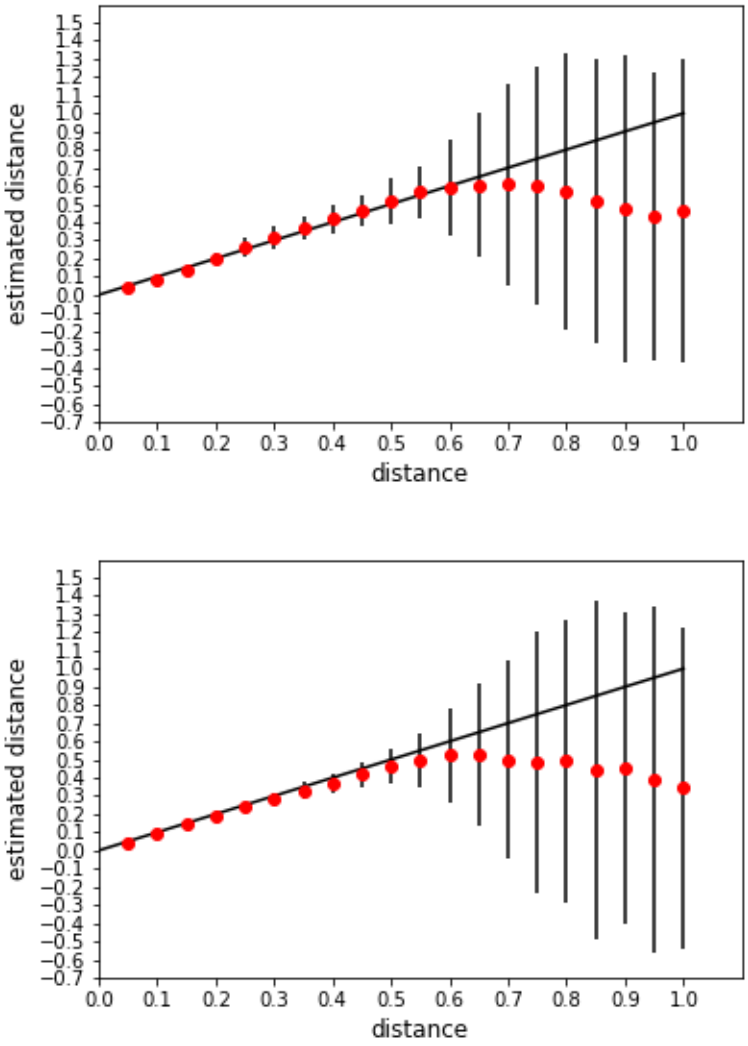

Figure 5: For *Jukes-Cantor* distances d between 0.05 and 1.0, we generated pairs of simulated *DNA* sequences of length $L = 10,000$; for each value of d, 20,000 sequences pairs with a distance of d were generated. For each pair, we estimated the distance d with *Slope-SpaM*, using *exact* k-mers (top) and spaced-word matches (bottom). Average estimated distances are plotted against the real distances; standard deviations are shown as error bars.

5.2 Distance estimation for sequences with local homologies

As shown theoretically in Section 2, our distance estimator should still work if two sequences share only local regions of homology, in contrast to previous approaches that also estimate distances from the number of word or spaced-word matches. To verify this empirically, we generated semi-artificial sequences consisting of local homologies, concatenated with non-homologous sequence regions of varying length. We then estimated phylogenetic distances between these sequences with *Slope-SpaM*, as well as with *Mash* and *Spaced*, two other alignment-free methods that are also based on the number of word matches. To see how these tools are affected by the presence of non-homologous sequence regions, we compared the distance values estimated from the semi-artificial sequences to the distances obtained from the original sequences.

To generate these sequence sets, we started with a set of 19 homologous gene sequences from different strains of the bacterium *Wolbachia*, taken from Gerth *et al.* [17]; the length of these sequences varied between $165kb$ and $175kb$. We then generated 9 additional data sets by concatenating these gene sequences with unrelated *i.i.d.* random sequences of varying length. Here, the proportion of the original homologous sequences within the concatenated sequences was 1.0 in the first set – the original sequences – , 0.9 in the second set, 0.8, in the third set, ..., and 0.1 in the tenth set. Within each of these 10 sequence sets, we estimated all $\binom{19}{2} = 171$ pairwise distances, and we calculated the ratio between each estimated distance value and the distance estimated the respective original sequence pair. For each data set – *i.e.* for each proportion of non-homologous random sequences –, we then took the *average* ratio between the distance of the semi-artificial and the original sequences.

The results of these test runs are plotted Fig. 6. As can be seen, distances calculated with *Mash* are heavily affected by including random sequences to the original homologous sequences. For the tenth data set, where the homologous gene sequences are only 10% of the sequences and 90% are unrelated random sequences, the distances calculated by *Mash* are, on average, 40 times higher than for the original gene sequences that are homologous over their entire length. *Spaced* is also affected if non-related random sequences are added, though to a lesser extent. Still, distances calculated by *Spaced* for the tenth data

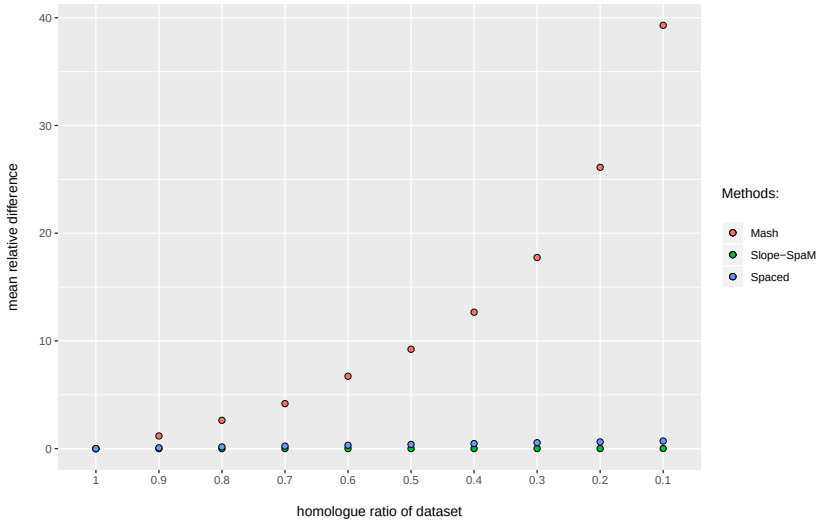

Figure 6: Comparison of distances estimated between *homologous* gene sequences and between *semi-artificial* sequences, obtained from the same homologous sequences by adding non-related random sequences of increasing length. The x-axis is the *fraction* of the homologous sequences within the semi-artificial sequences, the y-axis is the *ratio* between the distance estimated from the semi-artificial sequences and the distances estimated from the original gene sequences, see the main text for more details. Distances were estimated with three different alignment-free programs.

set are 1.7 times larger than the distances estimated for the original sequences. By contrast, *Slope-SpaM* was hardly affected at all by the presence of non-homologous random sequences.

5.3 Phylogenetic tree reconstruction

In addition to the above artificial and semi-artificial sequence pairs, we used sets of real genome sequences with known reference phylogenies from the *AFproject* web page [1] as benchmark data for phylogeny reconstruction. *AFproject* is a collaboration to systematically benchmark and evaluate software tools for alignment-free sequence comparison in different application scenarios [55]. The web page of the project provides a variety benchmark sequence sets, and the results of alignment-free methods on these sequences can be uploaded

to the server. Their quality is calculated and can be compared to a large number of other alignment-free methods, among them the state-of-the-art methods.

To evaluate *Slope-SpaM*, we downloaded four sets of full-genome sequences that are available in the categories *genome-based phylogeny* and *horizontal gene transfer*, namely (1) a set of 29 *E.coli/Shigella* genomes [53], (2) a set of 25 mitochondrial genomes from different fish species of the suborder *Labroidei* [16], (3) another set of 27 *E.coli/Shigella* [48] and (4) a set of 8 genomes of different strains of *Yersinia* [12]. Some of these data sets have been used previously by developers of alignment-free tools to benchmark their methods.

We ran *Slope-SpaM* on the four sets of genomes and uploaded the obtained distance matrices to the *AFproject* web server for evaluation. For these categories of benchmark data, the *AFproject* server calculates phylogenetic trees from the produced distance matrices using *Neighbor Joining* [45]. It then compares these trees to trusted reference trees of the respective data sets under the normalized *Robinson-Foulds (nRF)* metric. Here, the standard *Robinson-Foulds (RF)* distance [44] that measures the dissimilarity between two tree topologies is divided by the maximal possible *RF* distance of two trees with the same number of leaves; the *nRF* distances can therefore take values between 0 and 1. On the *AFproject* server, the benchmark results are then ranked in order of increasing *nRF* distance, *i.e.* in order of decreasing quality.

The results of our test runs were as follows: For the first set of *E.coli* genomes (29 genomes), the *nRF* distance of the obtained tree to the reference tree was 0.54, corresponding to rank 11 out of 18. For the fish mitochondrial genomes, the tree produced with the *Slope-SpaM* distances had a nRF distance of 0.32 to the reference three, corresponding to rank 5 out of 10. For the second *E.coli/Shigella* data (27 genomes), the *nRF* distance between the *Slope-SpaM* tree and the reference tree was again 0.54, which was this time position 11 out of 16. For the set of 8 *Yersinia* genomes, the topology of the *Slope-SpaM* tree was identical with the reference tree topology, so the *nRF* distance was 0, corresponding to rank 1 out of 6.

6 Discussion

A number of alignment-free methods have been proposed in the literature to estimate the nucleotide match probability p for a pair of DNA sequences from the number N_k or D_2 of k-mer matches for a fixed value of k. This can be done by setting N_k in relation to the total number of k-mers in the compared sequences [40] or to the length of the sequences [36]. A certain draw-back of these approaches is that, in order to accurately estimate the match probability p, not only N_k and the sequence lengths (or total number of k-mers, respectively), but also the extent of the homology between the compared sequences must be known.

Indeed, if two sequences share only short, local homologies with each other, an observed number N_k of k-mer matches could indicate a high match probability p within those homologous regions – while the same value of N_k would correspond to a lower p if the homology would extend over the entire length of the sequences. Since it is, in general, not known which proportion of the input sequences is homologous to each other, the above methods assume, for simplicity, that the compared sequences are homologous over their entire length. Obviously, this assumption affects the accuracy of these approaches since, in reality, genome sequences often share only local homologies.

In the present paper, we introduced *Slope-SpaM*, another approach to estimate the match probability p between two DNA sequences – and thereby their *Jukes-Cantor* distance – from the number N_k of word matches; we generalized this approach to *spaced-word matches*, based on patterns with k match positions. The main difference between *Slope-SpaM* and previous methods is that, instead of using only one single word length or pattern weight k, our program calculates N_k, together with an expression $F(k)$ that depends on N_k, for a whole range of values of k. The nucleotide match probability p in an alignment of the input sequences is then estimated from the slope of F. From the definition of $F(k)$, it is easy to see that the slope of F and our estimate of p depend on the values of N_k, but not on the extent of the homology between the compared sequences.

Test results on simulated sequences show that our distance estimates are accurate for distances up to around 0.5 substitutions per sequence position. Our experimental results also confirm that *Slope-SpaM* still produces accurate distance estimates if sequences contain only local regions of homology. In Fig. 2 and Fig. 4, sequence pairs

with a *Jukes-Cantor* distance of 0.2 were used. While the sequences in Fig. 2 are similar to each other over their entire length, in Fig. 4 a local region of sequence 'homology' is embedded in non-related random sequences twice the length of the homology region. Nevertheless, the slope of the function F is almost the same in both figure, leading to similar estimates of the match probability p. In fact, with the values shown in Fig. 4 (top), we would have obtained the same correct estimate for p that we obtained with the data from Fig. 2, if our program had chosen a minimum value $k_1 = 11$ for the word length k, instead of $k_1 = 10$. Alternatively, a sufficiently large maximal value k_2 would have led to a more accurate estimate of p.

On the real-world genomes, the performance of *Slope-SpaM* was in the medium range, compared to the other methods for which results are available on the *AFproject* server. On one real-world data set that is available on this server – a set of 14 plant genomes –, *Slope-SpaM* did dot produce distance values for all pairs of sequences, we therefore had to omit this set from the program evaluation. we expect that a better way of selecting the program parameters, in particular the range of word lengths (or pattern weights) k may further improve the accuracy of *Slope-SpaM*. Since the approach implemented in *Slope-SpaM* is novel and very different to existing alignment-free approaches, substantial improvements are possible by systematically analyzing the influence of the applied sequence parameters. Also, we used the *Jukes-Cantor* model, the simplest possible model for nucleotide substitutions. Using more sophisticated substitution models may also contribute to improve future versions of our program.

6.1 Acknowledgements

We thank Fengzhu Sun for his useful comments on an earlier version of this manuscript, Andrzej Zielezinski and Wojchiech Karlowski for making the *AFproject* server available, Jendrik Schellhorn for the test runs for the semi-artificial sequences with local homologies, Christoph Bleidorn and Micha Gerth for discussions about the *Wolbachia* genomes and Chris-André Leimeister and Peter Meinicke for general discussions on the project.

References

[1] http://150.254.120.147:7000/.

[2] Saulo Alves Aflitos, Edouard Severing, Gabino Sanchez-Perez, Sander Peters, Hans de Jong, and Dick de Ridder. Cnidaria: fast, reference-free clustering of raw and assembled genome and transcriptome NGS data. *BMC Bioinformatics*, 16:352, 2015.

[3] Nathan A. Ahlgren, Jie Ren, Yang Young Lu, Jed A. Fuhrman, and Fengzhu Sun. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Research*, 45:39–53, 2017.

[4] Metin Balaban, Shahab Sarmashghi, and Siavash Mirarab. APPLES: Fast distance-based phylogenetic placement. *bioRxiv*, 10.1101/475566, 2019.

[5] Guillaume Bernard, Cheong Xin Chan, Yao-Ban Chan, Xin-Yi Chua, Yingnan Cong, James M. Hogan, Stefan R. Maetschke, and Mark A. Ragan. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Briefings in Bioinformatics*, 22:426–435, 2019.

[6] Karel Břinda, Alanna Callendrello, Lauren Cowley, Themoula Charalampous, Robyn S Lee, Derek R MacFadden, Gregory Kucherov, Justin O'Grady, Michael Baym, and William P Hanage. Lineage calling can identify antibiotic resistant clones within minutes. *bioRxiv*, 10.1101/403204, 2018.

[7] Andrei Z. Broder. Identifying and filtering near-duplicate documents. In *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, COM '00, pages 1–10, Berlin, Heidelberg, 2000. Springer-Verlag.

[8] Raquel Bromberg, Nick V. Grishin, and Zbyszek Otwinowski. Phylogeny reconstruction with alignment-free method that corrects for horizontal gene transfer. *PLOS Comput Biol*, 12:e1004985, 2016.

[9] Daniel G. Brown. *Bioinformatics Algorithms: Techniques and Applications*, chapter A survey of seeding for sequence alignment, pages 126–152. Wiley-Interscience, New York, Fev. 2008.

[10] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12:59–60, 2015.

[11] Karel Břinda, Maciej Sykulski, and Gregory Kucherov. Spaced seeds improve k-mer-based metagenomic classification. *Bioinformatics*, 31:3584–3592, 2015.

[12] Aaron E. Darling, István Miklós, and Mark A. Ragan. Dynamics of genome rearrangement in bacterial populations. *PLOS Genetics*, 4(7), 2008.

[13] Lavinia Egidi and Giovanni Manzini. Design and analysis of periodic multiple seeds. *Theoretical Computer Science*, 522:62 – 76, 2014.

[14] Huan Fan, Anthony R. Ives, Yann Surget-Groba, and Charles H. Cannon. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics*, 16:522, 2015.

[15] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA, USA, 2004.

[16] Christoph Fischer, Stephan Koblmüller, Christian Gülly, Christian Schlötterer, Christian Sturmbauer, and Gerhard G. Thallinger. Complete mitochondrial DNA sequences of the threadfin cichlid (Petrochromis trewavasae) and the blunthead cichlid (Tropheus moorii) and patterns of mitochondrial genome evolution in cichlid fishes. *PLoS One*, 8(6):e67048–e67048, 2013.

[17] Michael Gerth, Marie-Theres Gansauge, Anne Weigert, and Christoph Bleidorn. Phylogenomic analyses uncover origin and spread of the Wolbachia pandemic. *Nature Communications*, 5:5117, 2014.

[18] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, Cambridge, UK, 1997.

[19] Lars Hahn, Chris-André Leimeister, Rachid Ounit, Stefano Lonardi, and Burkhard Morgenstern. *rasbhari*: optimizing spaced seeds for database searching, read mapping and alignment-free sequence comparison. *PLOS Computational Biology*, 12(10):e1005107, 2016.

[20] Bernhard Haubold. Alignment-free phylogenetics and population genetics. *Briefings in Bioinformatics*, 15:407–418, 2014.

[21] Bernhard Haubold, Fabian Klötzl, and Peter Pfaffelhuber. andi: Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics*, 31:1169–1175, 2015.

[22] Bernhard Haubold, Peter Pfaffelhuber, Mirjana Domazet-Loso, and Thomas Wiehe. Estimating mutation distances from unaligned genomes. *Journal of Computational Biology*, 16:1487–1500, 2009.

[23] Sebastian Horwege, Sebastian Lindner, Marcus Boden, Klaus Hatje, Martin Kollmar, Chris-André Leimeister, and Burkhard Morgenstern. *Spaced words* and *kmacs*: fast alignment-free sequence comparison based on inexact word matches. *Nucleic Acids Research*, 42:W7–W11, 2014.

[24] Lucian Ilie, Silvana Ilie, and Anahita M. Bigvand. SpEED: fast computation of sensitive spaced seeds. *Bioinformatics*, 27:2433–2434, 2011.

[25] Thomas H. Jukes and Charles R. Cantor. *Evolution of Protein Molecules*. Academy Press, New York, 1969.

[26] Gregory Kucherov. Evolution of biosequence search algorithms: a brief survey. *Bioinformatics*, btz272, 2019.

[27] Anna Katharina Lau, Chris-André Leimeister, and Burkhard Morgenstern. *Read-SpaM*: assembly-free and alignment-free comparison of bacterial genomes. *bioRxiv*, doi:10.1101/550632, 2019.

[28] Chris-André Leimeister, Marcus Boden, Sebastian Horwege, Sebastian Lindner, and Burkhard Morgenstern. Fast alignment-free sequence comparison using spaced-word frequencies. *Bioinformatics*, 30:1991–1999, 2014.

[29] Chris-André Leimeister and Burkhard Morgenstern. *kmacs*: the k-mismatch average common substring approach to alignment-free sequence comparison. *Bioinformatics*, 30:2000–2008, 2014.

[30] Chris-Andre Leimeister, Jendrik Schellhorn, Svenja Schöbel, Michael Gerth, Christoph Bleidorn, and Burkhard Morgenstern. Prot-SpaM: Fast alignment-free phylogeny reconstruction based on whole-proteome sequences. *GigaScience*, giy148, 2018.

[31] Chris-André Leimeister, Salma Sohrabi-Jahromi, and Burkhard Morgenstern. Fast and accurate phylogeny reconstruction using filtered spaced-word matches. *Bioinformatics*, 33:971–979, 2017.

[32] Ming Li, Bin Ma, Derek Kisman, and John Tromp. PatternHunter II: Highly sensitive and fast homology search. *Genome Informatics*, 14:164–175, 2003.

[33] Benjamin Linard, Krister Swenson, and Fabio Pardi. Rapid alignment-free phylogenetic identification of metagenomic sequences. *bioRxiv*, 2018.

[34] Yang Young Lu, Kujin Tang, Jie Ren, Jed A. Fuhrman, Michael S. Waterman, and Fengzhu Sun. CAFE: aCcelerated Alignment-FrEe sequence analysis. *Nucleic Acids Research*, 45:W554–W559, 2017.

[35] Bin Ma, John Tromp, and Ming Li. PatternHunter: faster and more sensitive homology search. *Bioinformatics*, 18:440–445, 2002.

[36] Burkhard Morgenstern, Svenja Schöbel, and Chris-André Leimeister. Phylogeny reconstruction based on the length distribution of k-mismatch common substrings. *Algorithms for Molecular Biology*, 12:27, 2017.

[37] Burkhard Morgenstern, Bingyao Zhu, Sebastian Horwege, and Chris-André Leimeister. Estimating evolutionary distances between genomic sequences from spaced-word matches. *Algorithms for Molecular Biology*, 10:5, 2015.

[38] Kevin D. Murray, Christfried Webers, Cheng Soon Ong, Justin Borevitz, and Norman Warthmann. kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity. *PLOS Computational Biology*, 13:e1005727, 2017.

[39] Laurent Noé. Best hits of 11110110111: model-free selection and parameter-free sensitivity calculation of spaced seeds. *Algorithms for Molecular Biology*, 12:1, 2017.

[40] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome Biology*, 17:132, 2016.

[41] Ji Qi, Hong Luo, and Bailin Hao. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Research*, 32(suppl 2):W45–W47, 2004.

[42] Gesine Reinert, David Chew, Fengzhu Sun, and Michael S. Waterman. Alignment-free sequence comparison (I): Statistics and power. *Journal of Computational Biology*, 16:1615–1634, 2009.

[43] Jie Ren, Xin Bai, Yang Young Lu, Kujin Tang, Ying Wang, Gesine Reinert, and Fengzhu Sun. Alignment-free sequence analysis and applications. *Annual Review of Biomedical Data Science*, 1:93–114, 2018.

[44] David F Robinson and Les Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

[45] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

[46] Shahab Sarmashghi, Kristine Bohmann, M. Thomas P. Gilbert, Vineet Bafna, and Siavash Mirarab. Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biology*, 20:34, 2019.

[47] Gregory E. Sims, Se-Ran Jun, Guohong A. Wu, and Sung-Hou Kim. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences*, 106:2677–2682, 2009.

[48] Elizabeth Skippington and Mark A. Ragan. Within-species lateral genetic transfer and the evolution of transcription al regulation in Escherichia coli and Shigella. *BMC Genomics*, 12:532, 2011.

[49] Kai Song, Jie Ren, Gesine Reinert, Minghua Deng, Michael S. Waterman, and Fengzhu Sun. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings in Bioinformatics*, 15:343–353, 2014.

[50] Kai Song, Jie Ren, Zhiyuan Zhai, Xuemei Liu, Minghua Deng, and Fengzhu Sun. Alignment-free sequence comparison based on next-generation sequencing reads. *Journal of Computational Biology*, 20:64–79, 2013.

[51] Igor Ulitsky, David Burstein, Tamir Tuller, and Benny Chor. The average common substring approach to phylogenomic reconstruction. *Journal of Computational Biology*, 13:336–350, 2006.

[52] Lin Wan, Gesine Reinert, Fengzhu Sun, and Michael S Waterman. Alignment-free sequence comparison (II): theoretical power of comparison statistics. *Journal of Computational Biology*, 17:1467–1490, 2010.

[53] Huiguang Yi and Li Jin. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Research*, 41:e75, 2013.

[54] Qian Zhang, Se-Ran Jun, Michael Leuze, David Ussery, and Intawat Nookaew. Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of k-mer. *Scientific Reports*, 7:40712, 2017.

[55] Andrzej Zielezinski, Hani Z Girgis, Guillaume Bernard, Chris-Andre Leimeister, Kujin Tang, Thomas Dencker, Anna K Lau, Sophie Roehling, JaeJin Choi, Michael S Waterman, Matteo Comin, Sung-Hou Kim, Susana Vinga, Jonas S Almeida, Cheong Xin Chan, Benjamin James, Fengzhu Sun, Burkhard Morgenstern, and Wojciech M Karlowski. Benchmarking of alignment-free sequence comparison methods. *bioRxiv*, 10.1101/611137, 2019.

[56] Andrzej Zielezinski, Susana Vinga, Jonas Almeida, and Wojciech M. Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, 18:186, 2017.