Insertions and deletions as phylogenetic signal in alignment-free sequence comparison

Niklas Birth[1], Thomas Dencker[1] and Burkhard Morgenstern[1,2,3]

[1]Department of Bioinformatics, Institute of Microbiology and Genetics, Universität Göttingen, Goldschmidtstr. 1
[2]Göttingen Center of Molecular Biosciences (GZMB), Justus-von-Liebig-Weg 11
[3]Campus-Institute Data Science (CIDAS), Goldschmidtstr. 1, 37077 Göttingen, Germany

February 4, 2021

Abstract

Most methods for phylogenetic tree reconstruction are based on sequence alignments; they infer phylogenies based on aligned nucleotide or amino-acid residues. Gaps in alignments are usually not used as phylogenetic signal, even though they can, in principle, provide valuable information. In this paper, we explore an alignment-free approach to utilize insertions and deletions for phylogeny inference. We are using our previously developed approach *Multi-SpaM*, to generate local gap-free four-way alignments, so-called *quartet blocks*. For pairs of quartet blocks involving the same four sequences, we consider the distances between these blocks in the four sequences, to obtain hints about insertions or deletions that may have occurred since the four sequences evolved from their last common ancestor. This way, a pair of quartet blocks can support one of the three possible quartet topologies for the four involved sequences. We use this information as input for *Maximum-Parsimony* and for the software program *Quartet Max-Cut* to reconstruct phylogenetic trees that are only based on insertions and deletions.

1 Introduction

The foundation of most phylogenetic studies are multiple sequence alignments (MSAs), either of partial or complete genomes or of individual genes or proteins. If MSAs of multiple genes or proteins are used, there are two possibilities to construct a phylogenetic tree: (1) the alignments can be concatenated to form a so-called *superalignment* or *supermatrix*. Tree building methods such as *Maximum-Likelihood* [43, 13], *Bayesian Approaches* [37] or *Maximum-Parsimony* [8, 10, 45] can then be applied to these superalignments. (2) One can calculate a separate tree for each gene or protein family and then use a *supertree approach* [4] to amalgamate these different trees into one final tree, with methods such as *ASTRAL* [52] or *MRP* [34].

Multiple sequence alignments usually contain gaps representing insertions or deletions (*indels*) that are assumed to have happened since the aligned sequences evolved from their last common ancestor. Gaps, however, are usually not used for phylogeny reconstruction. Most of the above tree-reconstruction methods are based on substitution models for nucleotide or amino-acid residues. Here, alignment columns with gaps are either completely ignored, or gaps are treated as 'missing information', for example in the frequently used tool *PAUP** [45]. Some models have been proposed that can include gaps in the *Maximum-Likelihood* analysis such as *TKF91* [47] and TKF92 [48], see also [17, 1, 27]. Unfortunately, these models do not scale well to genomic data. Thus, indels are rarely used as a source of information for the phylogenetic analysis.

In those studies that actually make use of indels, this additional information is usually encoded in some simple manner. The most straightforward encoding is to treat the gap character as a fifth character for DNA comparison, or as a 21st character in protein comparison, respectively. This means that the lengths of gaps are not explicitly considered, so a gap of length $\ell > 1$ is considered to represent ℓ independent insertion or deletion events. Some more issues with this approach are discussed in [39]; these authors introduced the "simple encoding" of indel data as an alternative. For every indel in the multiple sequence alignment, an additional column is appended. This column contains a present/absent encoding for an indel event which is defined as a gap with given start and end positions. If a longer gap is fully contained in a shorter gap in another sequence, it is considered as *missing information*. Such a simple binary encoding is an effective way of using the length of the indels to gain additional information and can be used in some *maximum-parsimony* framework. A disadvantage of these approaches is their relatively long runtime. The above authors also proposed a more complex encoding of gaps [39] which they further refined in a subsequent paper [31].

The commonly used approaches to encode gaps for phylogeny reconstruction are compared in [32].

The "simple encoding" of gaps has been used in many studies; one recent study obtained additional information on the phylogeny of Neoaves which was hypothesized to have a "hard polytomy" [19]. Despite such successes, indel information is still largely ignored in phylogeny reconstruction. Oftentimes, it is unclear whether using indels is worth the large overhead and increased runtime. On the hand, it has also been shown that gaps can contain substantial phylogenetic information [7].

All of the above mentioned approaches to use indel information for phylogeny reconstruction require MSAs of the compared sequences. Nowadays, the amount of the available molecular data is rapidly increasing, due to the progress in next-generation sequencing technologies. If the size of the analyzed sequences increases, calculating multiple sequence alignments quickly becomes too time consuming. Thus, in order to provide faster and more convenient methods to phylogenetic reconstruction, many alignment-free approaches have been proposed in recent years. Most of these approaches calculate pairwise distances between sequences, based on sequence features such as k-mer frequencies [40, 33, 22] or the number [26] or length [23, 49, 29] of word matches. Distance methods such as *Neighbor-Joining* [38] or *BIONJ* [11] can then reconstruct phylogenetic trees from the calculated distances. For an overview, the reader is referred to recent reviews of alignment-free methods [50, 15, 3].

Some recently proposed alignment-free methods use inexact word matches between pairs of sequences [51, 16, 25], where mismatches are allowed to some degree. Such word matches can be considered as pairwise, gap-free "mini-alignments". So, strictly spoken, these methods are not "alignment-free". In the literature, they are still called "alignment-free", as they circumvent the need to calculate full sequence alignments of the compared sequences. The advantage of such "mini-alignments" is that inexact word matches can be found almost as efficiently as exact word matches, by adapting standard word-matching algorithms.

A number of these methods use so-called *spaced-words* [18, 22, 30]. A spaced-word is a word composed of nucleotide or amino-acid symbols that contains additional *wildcard* characters at certain positions, specified by a pre-defined binary pattern P representing 'match positions' and 'don't-care positions'. If the same 'spaced word' occurs in two different sequences, this is called a *Spaced-word Match* or *SpaM*, for short. One way of using spaced-word matches – or other types of inexact word matches – in alignment-free sequence comparison is to use them as a proxy for full alignments, to estimate the number of mismatches per position in the (unknown) full sequence align-

ment. This idea has been implemented in the software *Filtered Spaced Word Matches (FSWM)* [25]; it has also been applied to protein sequences [24], and to unassembled reads [21]. Other approaches have been proposed recently, that use the *number* of *SpaMs* to estimate phylogenetic distances between DNA sequences [30, 36], see [28] for a review of the various *SpaM*-based methods.

The *SpaM* approach has also been applied to *multiple* sequence comparison. *Multi-SpaM* [6] is a recent extension of the *FSMW* idea that finds spaced-word matches with respect to some binary pattern P, each space-word match involving *four* different input sequences. Such a spaced-word match is called a *quartet P-block*, or *quartet block*, for short. Each *quartet block*, thus, consists four occurrences of the same spaced-word, with respect to a specific pattern P. For each such block, the program then identifies the optimal quartet tree topology based on the nucleotides aligned to each other at the *don't-care* position of P, using the program *RAxML* [43]. Finally, the quartet trees calculated in this way are used to find a supertree of the full set of input sequences. To this end, *Multi-SpaM* uses the program *Quartet MaxCut* [42].

In the present paper, we use the *quartet blocks* to use insertions and deletions as phylogenetic signal. More specifically, for *pairs* of quartet blocks that involve the same four sequences, we consider the distances between the two blocks in these four sequences. Different distances indicate insertion or deletion events since the four sequences evolved from their last common ancestor. If, for example, a pair of blocks involves sequences $S_1, \ldots, S_4$ and the distances between these blocks are equal for S_1 and S_2 as well as for S_3 and S_4 but different between S_1, S_2 and S_3, S_4, this would support a quartet tree where S_1 and S_2 are neighbours, as well as S_3 and S_4.

To evaluate the phylogenetic signal that is contained in such pairs of quartet blocks, we first evaluated the quartet topologies inferred by our method directly, by comparing them to trusted reference trees. Next, we used two different methods to infer a phylogenetic tree for the full set of input sequences, based on the quartet trees obtained from our quartet block pairs. First, we construct a super tree from the quartet trees that we inferred as outlined above, using the software *MAx quartet tree* that we already used in *Multi-SpaM*. Second, we used distances between pairs of blocks to generate a data matrix which we used as input for *maximum parsimony*, to find a tree that minimizes the number of insertions and deletions that we have to assume, given the different distances between the quartet blocks. We evaluate these approaches on data sets that are commonly used as benchmark data in alignment-free sequence comparison. Our evaluation showed that the majority of the inferred quartet trees is correct and should therefore be

useful additional information for phylogeny reconstruction. Moreover, the quality of the trees that we inferred from our quartet block pairs alone is roughly comparable to the quality of trees obtained with existing alignment-free methods.

2 Design and Implementation

2.1 Spaced words, quartet blocks and distances between quartet blocks

We are using standard notation from stringology as defined, for example, in [14]. For a sequence S over some alphabet, $S(i)$ denotes the i-th symbol of S. In order to investigate the information that can be obtained from putative indels in an alignment-free context, we use the *P-blocks* generated by the program *Multi-SpaM* [6]. At the start of every run, a binary pattern $P \in \{0,1\}^\ell$ is specified for some integer ℓ. Here, a "1" in P denotes a *match position*, a "0" stands for a *don't-care position*. The number of *match positions* in P is called its *weight* and is denoted by w. By default, we are using parameter values $\ell = 110$ and $w = 10$, so by default the pattern P has 100 *don't-care* positions.

A *spaced word* W with respect to a pattern P is a word over the alphabet $\{A, C, G, T\} \cup \{*\}$ of the same length as P, and with $W(i) = *$ if and only if i is a *don't care position* of P, i.e. if $P(i) = 0$. If S is a sequence of length N over the nucleotide alphabet $\{A, C, G, T\}$, and W is a spaced word, we say that W *occurs* at some position $i \in \{1, \ldots, \ell\}$, if $S(i + j - 1) = W(j)$ for every match position j in P. For two sequences S and S' and positions i and i' in S and S', respectively, we say that there is a *spaced-word match (SpaM)* between S and S' at (i, i'), if the same spaced word W occurs at i in S and at i' in S'. A *SpaM* can be considered as a local pairwise alignment without gaps. Given a nucleotide substitution matrix, the *score* of a spaced-word match is defined as the sum of the substitution scores of the nucleotides aligned to each other at the *don't-care* positions of the underlying pattern P. In *FSWM* and *Multi-SpaM*, we are using a substitution matrix described in [5]. In *FSWM*, only *SpaMs* with positive scores are used. It has been shown that this *SpaM-filtering* step can effectively eliminate most random spaced-word matches [25].

The program *Multi-SpaM* is based on *quartet (P)-blocks*, where a quartet block is defined as four occurrences of some spaced word W in four different sequences. For a set of $N \geq 4$ input sequences, a quartet block can be thus considered as a local gap-free four-way alignment. To exclude spurious

random quartet blocks, *Multi-SpaM* removes quartet blocks with low scores. More precisely, a quartet block is required to contain one occurrence of the spaced-word W, such that the other three occurrences of W have positive scores with this first occurrence. In this paper, we are considering *pairs* of quartet blocks involving the same four sequences, and we are using the distances between the two blocks in these sequences as phylogenetic signal.

2.2 Phylogeny inference using distances between quartet blocks

Let us consider two *quartet blocks* B_1 and B_2 – not necessarily based on the same binary pattern – involving the same four sequences $S_1, \ldots, S_4$, and let D_i be the distance between B_1 and B_2 in sequence $S_i, i = 1, \ldots, 4$. More specifically, we define D_i as the length of the segment in S_i between the two spaced-word occurrences corresponding to B_1 and B_2, see Figures 1 and 2 for examples. Let us assume that the blocks B_1 and B_2 are representing true homologies, i.e. for each of them the respective segments go back to a common ancestor in evolution. If then we find that two of these distances are different from each other, this would imply that an insertion or deletion has happened between B_1 and B_2, since the two sequences have evolved from their last common ancestor; if the two distances are equal, no such insertion or deletion needs to be assumed.

There are three possible fully resolved (i.e. binary) quartet topologies for the four sequences $S_1, \ldots, S_4$ that we denote by $S_1S_2|S_3S_4$ etc. In the sense of the *parsimony* paradigm, we can consider the distance between two blocks as a *character* and D_i as the corresponding *character state* associated with sequence S_i. If two distances, say D_1 and D_2 are equal, and the other two distances, D_3 and D_4 are also equal to each other, but different from S_1 and S_2, respectively, this would support the tree topology $S_1S_2|S_3S_4$: with this topology, one would have to assume only one insertion or deletion to explain the character states, while for $S_1S_3|S_2S_4$ or $S_1S_4|S_2S_3$, two insertions or deletions would have to be assumed. In this situation, we say that the pair (B_1, B_2) *strongly* supports topology $S_1S_2|S_3S_4$.

Next, we consider the situation where two of the distances are equal, say $D_1 = D_2$, and D_3 and D_4 would be different from each other, and also different from D_1 and D_2. From a parsimony point-of-view, all three topologies would be equally good in this case, since each of them would require two insertions or deletions. it may still seem more plausible, however, to prefer the topology $S_1S_2|S_3S_4$ over the two alternative topologies. In fact, if we would use a simple probabilistic model where an insertion/deletion event has

a fixed probability p, with $0 < p < 0.5$, along each branch of the topology, then it is easy to see that the topology $S_1S_2|S_3S_4$ would have a higher likelihood than the two alternative topologies. In this situation, we say that the pair (B_1, B_2) *weakly* supports the topology $S_1S_2|S_3S_4$. Finally, we call a pair of quartet blocks *informative*, if it – strongly or weakly – supports one of the three quartet topologies for the involved four sequences.

Sequence													Distance D_i
S_1	A	G	G	C	A	A	C	G	G	T			2
S_2	A	G	G	C	A	T	C	G	G	T			2
S_3	A	G	G	C	A	A	C	T	C	G	G	T	4
S_4	A	G	G	C	A	A	C	T	C	G	G	T	4

Figure 1: Distances between two quartet blocks involving the same four sequences. In the sense of *maximum parsimony* with respect to insertions and deletions, these distances would *strongly* support the topology $S_1S_2|S_3S_4$.

Sequence													Distance D_i
S_1	A	G	G	C	A	A	C	G	G	T			2
S_2	A	G	G	C	A	T	C	G	G	T			2
S_3	A	G	G	C	A	A	C	T	C	G	G	T	4
S_4	A	G	G	C	A	A	T	C	G	G	T		3

Figure 2: Distances between two quartet blocks as in Figure 1. Here, the distances would *weakly* support the topology $S_1S_2|S_3S_4$.

For a set of input sequences $S_1, \ldots, S_N, N \geq 4$, we implemented two different ways of inferring phylogenetic trees from quartet-block pairs. With the first method, we calculate the *quartet topology* for each quartet-block pair that supports one of the three possible quartet topologies. We then a calculate *supertree* from these topologies. Here, we use the program *Quartet MaxCut* [41, 42] that we already used in our previous software program *Multi-SpaM* where we inferred quartet topologies from the nucleotides aligned at the *don't-care* positions of quartet blocks.

Our second method uses the distances between quartet blocks as input for *Maximum-Parsimony* [8, 10]. To this end, we generate a character matrix as follows: the rows of the matrix correspond, as usual, to the input sequences, and each informative quartet block pair corresponds to one column. The distances between the two quartet blocks are encoded by characters 0, 1 and 2, such that equal distances in an informative quartet-block pair are encoded

Sequences	Character states				
S_1	1	1	-	1	1
S_2	1	-	-	-	0
S_3	0	1	0	-	-
S_4	-	0	1	0	1
S_5	-	2	1	0	0
S_6	0	-	0	1	-

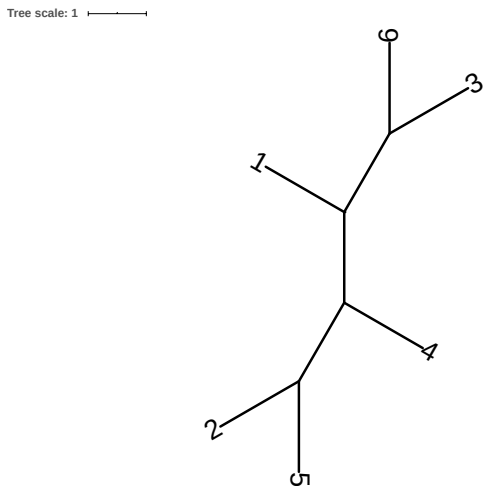

Figure 3: Character matrix for a set of 6 sequences, encoding distances between blocks for five quartet-block pairs, together with the tree topology calculated from this matrix. Each column corresponds to one informative block pair. Distances are represented by characters '0', '1' and '2'. The characters themselves are arbitrary, we only encode if the distance between the two blocks is equal or different in the four involved sequences. Dashes in a column represent 'missing information', for sequences that are not involved in the respective block pair. The matrix represents four quartet-block pairs that *strongly* support one quartet topology, namely column 1 supporting $S_1S_2|S_3S_6$, column 3 supporting $S_3S_6|S_4S_5$, column 4 supporting $S_1S_6|S_4S_5$ and column 5 supporting $S_1S_4|S_2S_5$. Column 2 corresponds to a quartet-block pair that *weakly* supports the topology $S_1S_3|S_4S_5$. The topology on the right-hand side was calculated from the matrix with the program *pars* form the *PHYLIP* package [9].

by the same character. For sequences not involved in a quartet-block pair, the corresponding entry in the matrix is empty and is considered as 'missing information'. In Figure 1, for example, the entries for S_1, S_2, S_3, S_4 would be 0, 0, 1, 1; in Figure 2, the corresponding entries would be 0, 0, 1, 2.

In order to find suitable quartet-block pairs for the two described approaches, we are using our software *Multi-SpaM*. This program samples up to 1 million quartet blocks. We use the quartet blocks generated by *Multi-SpaM* as *reference blocks*, and for each reference block B_1, we search for a second block in a window of L nucleotides to the right of B_1 for a second block B_2 involving involving the same four sequences. We use the first block that we find in this window, provided that the involved spaced-word matches

are *unique* within the window. If the pair $(B_1, B_2$ supports a topology of the involved four sequences – either strongly or weakly –, we use this block pair, otherwise the pair (B_1, B_2) is discarded.

3 Test results

In order to evaluate the above described approaches to phylogeny reconstruction, we used sets of genome sequences from *AF-Project* [53] that are frequently used as benchmark data for alignment-free methods. In addition, we used a set of 19 *Wolbachia* genomes [12]. For these sets of genomes, trusted phylogenetic trees are available that can be used as reference trees; these genomes have also been used as benchmark data to evaluate *Multi-SpaM* [6].

3.1 Quartet trees from quartet-block distances

First, we tested, how many *informative* quartet blocks we could find, i.e. how many of the identified quartet-block pairs would either *strongly* or *weakly* support one of the three possible quartet topologies for the corresponding four sequences. For each set of genome sequences, we generated 1,000,000 block pairs for each sequence set. The first ('reference') block in each pair was generated with *Multi-SpaM* [6] using default parameters, the second block was searched in a window of 500 *nt* to the right of the first block. For the second block, we used a pattern $P = 1111111$, i.e. blocks of exact word matches of length seven.

Table 1 shows the percentage of informative quartet block pairs, among the quartet block pairs that we used. To evaluate the correctness of the obtained quartet topologies, we compared them to the topologies of the respective quartet sub-trees of the reference trees. For this, we used the *ETE3* toolkit [20] to compute the *Robinson-Foulds (RF)* distance [35] between the two quartet topologies. If the *RF* distance is zero, the inferred quartet topology is in accordance with the reference tree.

We want to use the quartet trees that we obtain from informative quartet block pairs, to generate a tree of the full set of input sequences. Therefore, it is not sufficient for us to have a high percentage of correct quartet trees, but we also want to know how many of the sequence quartets are covered by these quartet trees. Generally, the results of super-tree methods depend on the *coverage* of the used quartet topologies [2, 44]. For a set of n input sequences, there are $\binom{n}{4}$ possible 'sequence quartets', i.e. sets of four sequences. Ideally, for every such set, we should have at least one quartet tree, in order to find the

	Strong support			Weak support			Strong & weak combined		
	# inf	% corr	% cov	# inf	% corr	% cov	# inf	% corr	% cov
E. coli 27	16,185	76.00	42.78	11,472	55.52	36.45	27,657	67.61	58.57
E. coli 29	1,658	78.93	34.67	10,588	58.86	27.83	27,169	71.20	48.59
Fish mito	3,049	67.48	15.76	6,975	61.48	27.44	10,024	63.36	34.54
Yersinia	4,871	45.91	100.00	4,516	41.48	98.86	9,387	43.77	100.00
Wolbachia	46,972	93.91	79.79	30,887	74.46	85.75	77,859	86.32	95.63

Table 1: Test results on different sets of genomes. As benchmark data, we used four sets of genome sequences from *AF-Project* [53] and one set of *Wolbachia* genomes [12]. For each data set, we generated 1,000,000 pairs of quartet blocks as described in the main text. The table shows the number of *informative* block pairs ('# inf'), i.e. the number of block pairs for which we obtained either strong or weak support for one of the three possible quartet topologies of the involved sequences. In addition we show the percentage of *correct* quartet topologies (with respect to the respective reference tree), out of all informative block pairs, as well as the 'coverage' by quartet blocks, i.e. the percentage of sequence quartets for which we found at least one informative block pair.

correct super tree. Table 1 reports the *quartet coverage*, i.e. the percentage of all sequence quartets, for which we obtained at least one quartet tree.

Note that *Multi-SpaM* uses randomly sampled quartet blocks, the program can thus return different results for the same set of input sequences. We therefore performed 10 program runs on each set of sequences and report the average *correctness* and *coverage* of these test runs.

3.2 Full phylogeny reconstruction

Finally, we applied our quartet-block pairs to reconstruct full tree topologies for the above sets of benchmark sequences. Here, we used two different approaches, namely *Quartet MaxCut* and *Maximum-Parsimony*, as described above. As is common practice in the field, we evaluated the quality of the reconstructed phylogenies by comparing the the respective reference trees from *AFproject* using the *normalized Robinson-Foulds (RF) distances* between the inferred and the reference topologies. For a data set with n taxa, the *normalized RF distances* are obtained from the *RF distances* by dividing them by $2 * n - 6$, i.e. by the maximum possible *distance* for trees with n leaves. The results of other alignment-free methods on these data are reported in [6, 53].

We applied the program *Quartet MaxCut* first to the quartet topologies derived from the set of *all* informative quartet-block pairs. As a comparison,

	Indel Information				
	Quartet MaxCut			Parsimony	*Multi-SpaM*
	Strong	Weak	Combined		
E. coli 29	12.8	19.7	15.8	13.4	12.6
E. coli 27	10.8	17.2	12.4	10.6	8.8
Fish mito	23.6	24.0	18.4	23.8	7.8
Yersinia	6.0	9.0	6.0	6.0	6.2
Plants	7.4	8.4	8.6	7	8
Wolbachia	6.0	7.4	6.8	6.0	6

Table 2: Average *Robinson-Foulds (RF)* distances between trees, reconstructed with various alignment-free methods, for five sets of genome sequences from *AFproject* [53] and one set of genomes from *Wolbachia*. For each data sets, the average over 10 program runs was taken, except for the *Plants* data set, where only one single program run was carried out.

we then inferred topologies using only those quartet-block pairs that *strongly* support one of the three possible topologies for the four involved sequences. The results of these test runs are shown in Table 2.

Next, we used the program *PAUP** [45] to calculate the most parsimonious tree, using the distances between quartet blocks as characters, as explained above. Here, we used the *TBR* [46] heuristic. In some cases, this resulted in multiple optimal, i.e. most parsimonious trees. In these cases, we calculated the average and error bars over all optimal trees. The results of these test runs are also shown in Table 2, together with the results from *Multi-SpaM*.

4 Discussion

Sequence-based phylogeny reconstruction usually relies on nucleotide or amino-acid residues aligned to each other in multiple alignments. Information about insertions and deletions (indels) is neglected in most studies, despite evidence that this information may be useful for phylogeny inference. There are several difficulties when indels are to be used as phylogenetic signal: it is difficult to derive probabilistic models for insertions and deletions, and there are computational issues if gaps of different lengths are spanning multiple columns in multiple alignments. Finally, gapped regions in sequence alignments are often considered less reliable than un-gapped regions, so the precise length

of insertions and deletions that have happened may not be easy to infer from multiple alignments.

In recent years, many fast alignment-free methods have been proposed to tackle the ever increasing amount of sequence data. Most of these methods are based on counting or comparing *words*, and gaps are usually not allowed within these words. It is therefore not straight-forward to adapt standard alignment-free methods to use indels as phylogenetic information.

In the present paper, we proposed to use *blocks* of aligned sequence segments to obtain information about possible insertions and deletions since the compared sequences have evolved from a common ancestor. *Within* these blocks, no gaps are allowed. To obtain hints about possible insertions or deletions, we consider *pairs* of such blocks involving the same sequences, and we look at the distances between these blocks in the involved sequences. If the distances between two blocks are different for two sequences, this indicates that there has been an insertion or deletion since they evolved from their last common ancestor. If the two distances are the same, no such event needs to be assumed. This information can be used to infer a tree topology for the sequences involved in a pair of blocks. To our knowledge, this is the first attempt to use insertions and deletions as phylogenetic signal in an alignment-free context.

In this study, we restricted ourselves, for simplicity, to blocks involving *four* input sequences each. We did not consider the *length* of hypothetical insertions and deletions, but only asked whether or not such an event might have happened between two sequences in the region bounded by two blocks. Since we can assume that indels are relatively rare events the *maximum parsimony* paradigm seems to be suitable in this situation. In the sense of *parsimony*, however, only those block pairs that *strongly* support one of three possible quartet topologies are informative. Indeed, if two distances between two blocks are equal, and the third and fourth distance are different – and different from each other –, then each of the three possible quartet topologies would require two insertion or deletion events.

Intuitively, however, one may want to use the information from such quartet blocks pairs that, in our terminology, *weakly* support one of the possible topologies. It is easy to see that, with a simple probabilistic model under which an insertion between two blocks occurs with a probability $p < 0.5$, independently of the length of the insertion and the distance between the blocks, a *weakly* supported topology would have a higher likelihood than the two alternative topologies – although from a parsimony point-of-view all topologies are equally good. So it might be interesting to apply such a simple probabilistic model to our approach, instead of maximum parsimony.

Also, while we restricted ourselves to quartet blocks in this study, it

might be worthwhile to use blocks involving more than four sequences, and to consider pairs of such blocks that share at least four sequences.

Using standard benchmark data, we could show that phylogenetic signal from putative insertions and deletions between quartet blocks is mostly in accordance with the reference phylogenies that we used as standard of truth. Interestingly, the quality of the tree topologies that we constructed from our "informative" pairs of quartet blocks – i.e. from indel information alone – is roughly comparable to the quality of topologies obtained with existing alignment-free methods. The phylogenetic signal from indels that we used, however, is complementary to the information that is used by those existing approaches. We expect therefore, that the accuracy of alignment-free phylogeny methods can be further improved by combining these two complementary sources of information.

References

[1] Alexander V. Alekseyenko, Christopher J. Lee, and Marc A. Suchard. Wagner and Dollo: a stochastic duet by composing two parsimonious solos. *Systematic Biology*, 57:772–784, 2008.

[2] Eliran Avni, Zahi Yona, Reuven Cohen, and Sagi Snir. The Performance of Two Supertree Schemes Compared Using Synthetic and Real Data Quartet Input. *J. Mol. Evol.*, 86:150–165, 2018.

[3] 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.

[4] Olaf R.P. Bininda-Emonds. The evolution of supertrees. *Trends in Ecology and Evolution*, 19:315 – 322, 2004.

[5] Francesca Chiaromonte, Von Bing Yap, and Webb Miller. Scoring pairwise genomic sequence alignments. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, and Teri E. Klein, editors, *Pacific Symposium on Biocomputing*, pages 115–126, Lihue, Hawaii, 2002.

[6] Thomas Dencker, Chris-André Leimeister, Michael Gerth, Christoph Bleidorn, Sagi Snir, and Burkhard Morgenstern. *Multi-SpaM*: a Maximum-Likelihood approach to phylogeny reconstruction using multiple spaced-word matches and quartet trees. *NAR Genomics and Bioinformatics*, 2:lqz013, 2020.

[7] Christoph Dessimoz and Manuel Gil. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biol.*, 11:R37, 2010.

[8] James S. Farris. Methods for computing Wagner trees. *Systematic Biology*, 19:83–92, 1970.

[9] Joseph Felsenstein. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, 5:164–166, 1989.

[10] Walter Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, 20:406–416, 1971.

[11] Olivier Gascuel. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14:685–695, 1997.

[12] Michael Gerth and Christoph Bleidorn. Comparative genomics provides a timeframe for Wolbachia evolution and exposes a recent biotin synthesis operon transfer. *Nature Microbiology*, 2:16241, 2017.

[13] Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Oliver Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, 59:307–321, 2010.

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

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

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

[17] Ian H. Holmes. Solving the master equation for Indels. *BMC Bioinformatics*, 18:255, 2017.

[18] 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.

[19] Peter Houde, Edward L. Braun, Nitish Narula, Uriel Minjares, and Siavash Mirarab. Phylogenetic signal of indels and the neoavian radiation. *Diversity*, 11, 2019.

[20] Jaime Huerta-Cepas, François Serra, and Peer Bork. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.*, 33:1635–1638, 2016.

[21] Anna Katharina Lau, Svenja Dörrer, Chris-André Leimeister, Christoph Bleidorn, and Burkhard Morgenstern. *Read-SpaM*: assembly-free and alignment-free comparison of bacterial genomes with low sequencing coverage. *BMC Bioinformatics*, 20:638, 2019.

[22] 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.

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

[24] Chris-Andre Leimeister, Jendrik Schellhorn, Svenja Dörrer, Michael Gerth, Christoph Bleidorn, and Burkhard Morgenstern. Prot-SpaM: Fast alignment-free phylogeny reconstruction based on whole-proteome sequences. *GigaScience*, 8:giy148, 2019.

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

[26] Ross A. Lippert, Haiyan Huang, and Michael S. Waterman. Distributional regimes for the number of k-word matches between two random sequences. *Proceedings of the National Academy of Sciences*, 99:13980–13989, 2002.

[27] István Miklós, Gerton A. Lunter, and Ian Holmes. A Long Indel Model For Evolutionary Sequence Alignment. *Molecular Biology and Evolution*, 21:529–540, 2004.

[28] Burkhard Morgenstern. Sequence comparison without alignment: The SpaM approaches. In Kazutaka Katoh, editor, *Multiple Sequence Alignment*, Methods in Molecular Biology, pages 121–134. Springer, 2020.

[29] 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.

[30] 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.

[31] Kai Müller. Incorporating information from length-mutational events into phylogenetic analysis. *Mol. Phylogenet. Evol.*, 38(3):667–676, Mar 2006.

[32] T. Heath Ogden and Michael S. Rosenberg. How should gaps be treated in parsimony? A comparison of approaches using simulation. *Mol. Phylogenet. Evol.*, 42:817–826, 2007.

[33] 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.

[34] Mark A. Ragan. Phylogenetic inference based on matrix representation of trees. *Mol. Phylogenet. Evol.*, 1:53–58, 1992.

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

[36] Sophie Röhling, Alexander Linne, Jendrik Schellhorn, Morteza Hosseini, Thomas Dencker, and Burkhard Morgenstern. The number of k-mer matches between two DNA sequences as a function of k and applications to estimate phylogenetic distances. *PLOS ONE*, 15:e0228070, 2020.

[37] Fredrik Ronquist and John P. Huelsenbeck. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572–1574, 2003.

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

[39] Mark P. Simmons and Helga Ochoterena. Gaps as characters in sequence-based phylogenetic analyses. *Syst. Biol.*, 49:369–381, 2000.

[40] 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.

[41] Sagi Snir and Satish Rao. Quartets MaxCut: A divide and conquer quartets algorithm. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 7:704–718, 2010.

[42] Sagi Snir and Satish Rao. Quartet MaxCut: A fast algorithm for amalgamating quartet trees. *Molecular Phylogenetics and Evolution*, 62:1 – 8, 2012.

[43] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30:1312–1313, 2014.

[44] M. Shel Swenson, Rahul Suri, C. Randal Linder, and Tandy Warnow. An experimental study of Quartets MaxCut and other supertree methods. *Algorithms Mol Biol*, 6:7, 2011.

[45] David Swofford. PAUP*. phylogenetic analysis using parsimony (*and other methods). version 4.0b10. *Sinauer Associates, Sunderland, Massachusetts*, 2003.

[46] David L. Swofford and Garry J. Olsen. Phylogeny reconstruction. In D.M. Hillis and C. Moritz, editors, *Molecular Systematics*, pages 407–511. Sinauer Associates, 1990.

[47] Jeffrey L. Thorne, Hirohisa Kishino, and Joseph Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.*, 33:114–124, 1991.

[48] Jeffrey L. Thorne, Hirohisa Kishino, and Joseph Felsenstein. Inching toward reality: An improved likelihood model of sequence evolution. *Journal of Molecular Evolution*, 34:3–16, 1992.

[49] 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.

[50] Susana Vinga and Jonas Almeida. Alignment-free sequence comparison - a review. *Bioinformatics*, 19:513–523, 2003.

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

[52] Chao. Zhang, Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(Suppl 6):153, 2018.

[53] Andrzej Zielezinski, Hani Z Girgis, Guillaume Bernard, Chris-André Leimeister, Kujin Tang, Thomas Dencker, Anna Katharina Lau, Sophie Röhling, 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. *Genome Biology*, 20:144, 2019.