An Alignment-free Method for Phylogeny Estimation using Maximum Likelihood

Tasfia Zahin*, **Md. Hasin Abrar***, **Mizanur Rahman, Tahrina Tasnim, Md. Shamsuzzoha Bayzid**, **Atif Rahman**†

Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology, Bangladesh

Abstract

Phylogenetic analysis i.e. construction of an accurate phylogenetic tree from genomic sequences of a set of species is one of the main challenges in bioinformatics. The popular approaches to this require aligning each pair of sequences to calculate pairwise distances or aligning all the sequences to construct a multiple sequence alignment. The computational complexity and difficulties in getting accurate alignments have led to development of alignment-free methods to estimate phylogenies. However, the alignment free approaches focus on computing distances between species and do not utilize statistical approaches for phylogeny estimation. Herein, we present a simple alignment free method for phylogeny construction based on contiguous sub-sequences of length k termed k-mers. The presence or absence of these k-mers are used to construct a phylogeny using a *maximum likelihood* approach. The results suggest our method is competitive with other alignment-free approaches, while outperforming them in some cases.

Keywords: phylogeny, alignment-free, k-mer, maximum likelihood

1 Introduction

A phylogenetic or evolutionary tree represents the evolutionary history of different biological organisms. The earliest phylogenetic tree was portrayed by Darwin in his book "The Origin of Species" [1]. Efficient and accurate construction of phylogenies from genomic data of various species is a fundamental problem in the fields of biology such as bioinformatics and systematics.

Phylogeny reconstruction methods can be broadly classified into two groups: *distance based* and *character based*. *Distance based* methods take a distance matrix containing pairwise distances among the species as input, which is calculated from the sequences in a pre-processing step. On the other hand, character based methods make use of the sequences typically in the form of a multiple sequence alignment. Popular distance based methods include *UPGMA* [2], *Neighbor-joining* [3] etc. They are fast and can handle many sequences but they perform well when the species involved have high similarity. *Maximum parsimony* [4], on the other hand, is a character based approach, where a character matrix is taken as input. The best tree under maximum parsimony criterion is the one that minimizes the number of changes in the nucleotide sequences over time. *Maximum Likelihood* [5], a probabilistic character based approach, uses a specific model of sequence evolution to find a best scoring tree that maximizes likelihood of observing the set of input sequences. This approach is quite realistic in nature and can be used for species that vary widely in terms of similarity.

Both *distance based* and *character based* approaches usually require prior alignment of input sequences. Alignment of sequences is a process that inserts gaps within the sequences in such a way that the identical nucleotides of different species align next to each other as much as possible. In distance based methods, sequences are aligned pairwise whereas in character based methods the sequences of all the species are aligned to construct a multiple sequence alignment. The quality of alignments greatly affects the result of the analysis. Finding the best alignment for multiple sequences is not trivial. As the length of two sequences increases, the number of possible alignments increases exponentially and it becomes difficult to find an optimal alignment. Usually the alignment is done progressively, and various heuristics are used. Constructing the distance matrix through alignment of each pair of sequences is also computationally expensive.

To overcome these difficulties, phylogenetic analyses have shifted towards approaches that are no longer confined to alignment needs. A number of *alignment free* methods have been introduced, that construct tree models of genetic relations among species without the need for alignments, saving a lot of time and memory in the process. Moreover, sometimes parts of the genomic sequences become shuffled or swapped which cause alignment based methods to perform poorly. Alignment free ones, however, are robust to such rearrangement events and also efficient for large sequence lengths.

A multitude of alignment-free methods have been developed recently that have been comprehensively reviewed in [6, 7]. In the early days, exact matches were used as a basis for measuring similarity between sequences. Later works have extended these to allow a few mismatches. Notably, *cophylog* [8] makes use of matching word counts in the sequences

*These authors contributed equally
†Corresponding author. Email: atif@cse.buet.ac.bd

with the allowance of a mismatch, followed by pairwise distance calculation and tree generation. *andi* [9] looks for mismatches surrounded by long exact matches. The mismatches in these words are then counted to find the number of substitutions between two sequences. *Multi-SpaM* [10] works on the *SWM or Space Word Match* [11] approach to identify quartet groups, *i.e.* a group of four space words with matching nucleotides at the match positions and probable mismatches at the don't care positions. Trees obtained from each group are then combined to form the final tree.

However, most of the alignment free methods developed so far are distance based methods and hence they do not allow model based phylogeny estimation that are known to be more robust than distance based approaches. An illustration of classification of phylogeny estimation approaches was presented in [6] which is shown in Figure 1. Höhl and Ragan [12] proposed a Bayesian approach for phylogeny inference using presence and absence of k-mers. In this paper, we follow a similar approach and present *K-Phylo*, an alignment-free method for phylogeny reconstruction that uses maximum likelihood for tree estimation. It is based on presence or absence of k-mers in the input sequences. We propose an approach for k-mer length selection and apply our method on standard datasets used to assess alignment free methods.

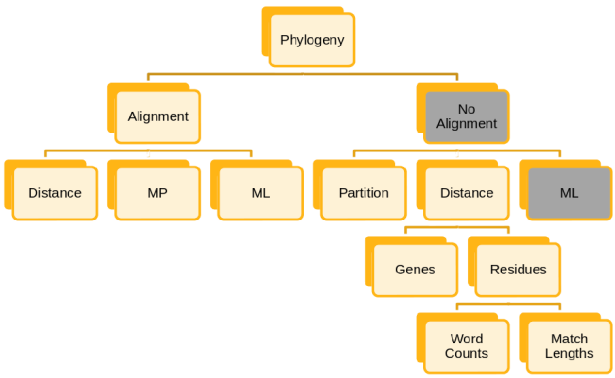

Figure 1: Classification of phylogeny estimation methods [6]. The blocks in grey are highlighted to mark their usage in the model we developed.

2 Methods

K-Phylo is a character based, alignment-free method that circumvents the complexity of multiple sequence alignment and combines the merits of maximum likelihood estimation in tree construction. Figure 2 shows the steps involved in the reconstruction process of this method. First, a k-mer counting tool is used to determine the set of k-mers present in each sequence for various k. Then these are combined to construct binary matrices denoting presence or absence of the k-mers in different species and the best k is chosen. Finally, the matrix corresponding to the best k is fed to a suitable phylogeny construction software to produce the output tree topology. Each step is described in more detail in the following sub-sections.

2.1 Generating K-mers

In this step, k-mers are extracted from the input DNA sequences using Jellyfish 2.2.4 tool [13] for a range of values of *k*. Both the actual k-mer and its reverse compliment equivalent are retrieved. For example, if a k-mer is *ACGTA* then its compliment is *TGCAT* and reverse compliment is *TACGT*. DNA strands exist in double helix form and sub-sequences from both strands are likely to appear in the input sequences. This is why both variations are included in the count. The output of this step is a list of k-mers.

2.2 Generating Binary Matrices

Depending on whether the extracted k-mer instances exist in the given sequences or not, an output similar to the binary matrix in Table 1 is produced. This matrix constitutes of only 0's and 1's. Its rows and columns represent the k-mers and input species respectively. An entry in the matrix contains 1 if the k-mer representing the row exists in the sequence of the specie representing the column, and 0 otherwise. One such matrix is produced for each value of k in a particular range which is further explained in 2.3. We have used hashing for this particular task. All the generated k-mers are read from a file and a unique index is generated for each of them. These are inserted in a hash table along with the species identification number. The k-mers are then read one by one from this hash table, while placing the appropriate value in the desired position of the matrix.

k-mer	S1	S2	S3	S4
ATTGCA	0	0	1	0
AATTCA	0	1	1	0
AGTGCT	1	0	1	0
CGTGCC	0	0	0	0
GGTGCC	0	0	0	1
GGTGCG	1	0	0	1

Table 1: Binary Matrix

2.3 Finding an Appropriate k-mer Length

A major challenge in k-mer based phylogeny estimation methods is the selection of appropriate k-mer length. As the k-mer length increases, the probability of finding this k-mer in the nucleotide sequences reduces and vice versa (Figure 3). A small k-value means the k-mer will be common and will result in large number of 1's in the binary matrix while a large k-value would produce long runs of 0's. An optimal k-mer length should be such that it captures the similarities of close species and the dissimilarities of distant ones, making the best usage of information offered by the dataset.

One of the ways of k selection is explained in [14]. The limitation of this selection process is that it is solely dependent on sequence length and does not take into account resemblances between sequences. Another mechanism in [15] explains a method of finding a range of feasible values of k as a

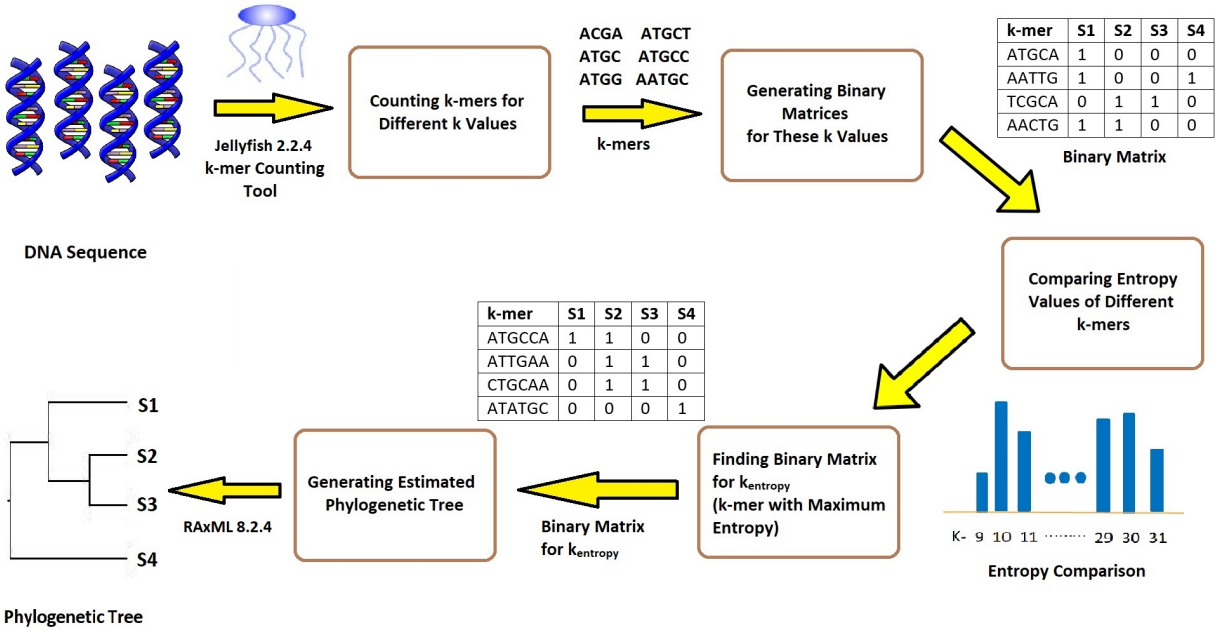

Figure 2: First, different k-mers are listed from the input sequences. Then separate binary matrices from all these k-mer counts are produced. From the binary matrices of different k-mers, $k_{entropy}$ is chosen based on the cumulative entropy values. Lastly, the binary matrix produced from this k-mer length is fed into RAxML for the estimation of the final phylogenetic tree.

function of sequence length, but does not refer to selection of any specific value of k. So this selection process is also not very suitable for our method. *K-Phylo* uses a simple way of choosing an appropriate k-mer length - *entropy*.

Figure 3: Probability of existence of k-mers depends on the k-mer length. In this figure, the k-mer *AT* of length 2 is found in all 3 taxa. However, the k-mer *ATAGCGC* of length 7 is found only in the source taxon (T1).

Entropy is the randomness or information loss of a system. It represents disorder (e.g. variation between two DNA sequences) and can be used to model the diversity of different genome sequences [16]. Since it is a good measure of randomness, it can be expected to capture the uncertainty present in a binary matrix. Maximum entropy occurs when k-mers of a certain length have exactly 50% presence among all the taxa. Long runs of 0's and 1's would reduce overall entropy while short runs would do the reverse. Information entropy is the average rate at which information is produced by a stochastic source of data. We use the following equation for entropy calculations.

$$H(X) = -\sum p(X) \log p(X)$$

Empirical evidence suggests that k values less than 9 make the k-mers too common while those greater than 31 make exact matches difficult to find in the sequences. *K-Phylo* is therefore run from k-mer length 9 to 31, and entropy values are calculated for each value of k. From the binary matrix corresponding to a certain value of k, we choose 5000 random k-mer rows and calculate the cumulative entropy using the mentioned equation. The cumulative entropy values from different k-mer lengths are then compared. The value of k resulting in the maximum entropy is chosen to be the most suitable one, indicating highest uniformity. Let us call this length $k_{entropy}$.

2.4 Generating Phylogenetic Tree

Once we have obtained the $k_{entropy}$ value, the next step is to construct a tree topology from the binary matrix produced for this length. To do this, this matrix is fed into RAxML software [17] as input. From the various maximum likelihood models available, we have used BINGAMMA since our input data contains merely 0's and 1's. The BINGAMMA model is defined for binary data and assumes a gamma prior on site mutation rates. This model takes in binary sequences and outputs a tree topology. Since a character substitution affects many k-mers in the sequences, we focus on topology at this stage and leave branch length estimation as future work. We also note that RAxML assumes the sites are independent which is violated here due to the same reason.

RF Distance	K-Phylo	RAxML(w alignment)
Average	0.9	0.8
MAX	2	2
MIN	0	0
Mode	0	0
Standard Deviation	0.99	0.98

Table 2: Comparison on 20 simulated datasets

2.5 Implementation

K-Phylo is implemented in C++ and Python. Besides, it uses the following tools for k-mer counting and phylogeny estimation:

- **K-mer Counting Tool - Jellyfish 2.2.4**
  A rigorous comparison of k-mer counting methods is presented by Zhang et al. [13] On the basis of this comparison we can find that Jellyfish is fast, and supports dynamic memory. Jellyfish is therefore preferable for large genome sequences. It extracts both the k-mer and its reverse complement (explained in Section 2.1) in the counting step.

- **Phylogenetic Tree Construction Software - RAxML 8.2.4**
  RAxML stands for **R**andomized **Ax**elerated **M**aximum **L**ikelihood [17]. It is a popular phylogenetic analysis software and can handle large datasets. It uses partial likelihood vectors over and over and thus by rearranging distance, it restores branch length and topology. For tree production, we have used its BINGAMMA model.

3 Datasets and Results

3.1 Datasets

We have used both simulated and biological datasets for testing *K-Phylo*. 20 different simulated nucleotide sequence data are generated using *SeqGen* [18], varying the number of species, source tree and branch lengths. It is provided with a source tree (in Newick format) as input and is used to generate nucleotide sequences according to the input tree. Some constraints are applied to the source phylogenetic trees used for generating the simulated sequences. It is ensured that the input species are neither too distant, nor too close. Branch lengths of extreme values, too small or too large, are avoided. The model we have used is general time reversible model with sequence length set to 50000. The source tree is treated as benchmark and the primary performance metric used for comparison of tree topology is Robinson Foulds (RF) distance between estimated tree and its benchmark. This metric gives a measure of distance between two trees by counting the number of dissimilar partitions. The RF distance between the estimated and benchmark tree is found using *KTreeDist* [19]. Performance of simulated data is compared with that of an alignment-based method, GTRGAMMA model from RAxML.

Five real datasets, relevant to our research, are used to compare our method to the existing different methods. These are full mitochondrial genome sequences of seven primates [6], full genome sequences of 8 Yersinia strains [20], full genome sequences of 27 *E.coli/Shigella* strains [20], assembled genomes of 25 fish species of the suborder *Labroidei* [21] and assembled 29 *E.coli/Shigella* strains [8]. In case of the real datasets other than the seven primates dataset, the sequences, benchmark trees and RF distances - all are obtained from the site AFproject [7] - a platform for comparison of alignment-free methods on different datasets. For seven primates, sequences and benchmark are obtained from [6].

3.2 Selection of k-mer length using entropy

A matter of interest is whether the chosen k value ($k_{entropy}$) with maximum entropy actually corresponds to the estimated tree with minimum RF distance or how far it is from the ideal k generated by *K-Phylo*. For example, from Figure 4, in the seven primates dataset, different k-mer lengths resulted in trees having different RF distances but *K-Phylo* successfully captured the k-mer length which generated a tree having RF distance of 0. Similarly, in the 27 *E.coli/Shigella* strains dataset, although several k-mer lengths result in trees having various RF distances like 8, 10, 12 or even 18, *K-Phylo* picked the $k_{entropy}$ value of 25 which produced a tree having the lowest RF distance (8) compared to the RFs generated from other k values using *K-Phylo*.

From here on, we consider the tree corresponding to $k_{entropy}$ and compare its RF distance with those achieved by similar methods. In this paper, we compare *K-Phylo* to *co-phylog* [8], *andi* [9], *Multi-SpaM* [10], *Average Common Substring* [22], *Composition Vector* [23], *kr* [24], *FFP* [25], *Grammar-Based Distance* [26] and *Skmer* [27] on the datasets mentioned. The ranking of these methods shown in Figure 5 are taken from AFproject [7].

3.3 Simulated Data

For each simulated dataset, we construct a tree by aligning the sequences using *Clustal Omega* [28] and then running RAxML with the GTRGAMMA model. We also construct a tree by running *K-Phylo* in an alignment free approach. The trees generated by *K-Phylo* from most of the simulated samples have RF distances of 0 from the source tree, *i.e.* they are exactly same. In many cases, the RF values between estimated tree and source tree are as good as the RF values between RAxML (with alignment) and the source tree. In some rare cases, RAxML (with alignment) exceeds *K-Phylo* in terms of performance.

The results are apparent from Table 2. We can see that the average RF distance between *K-Phylo* and source tree is very close to that of the approach using alignment. The maximum and minimum RF distances are similar for both approaches considering sample simulated sequences. The mode values of RF difference are the same, while the averages are almost equal.

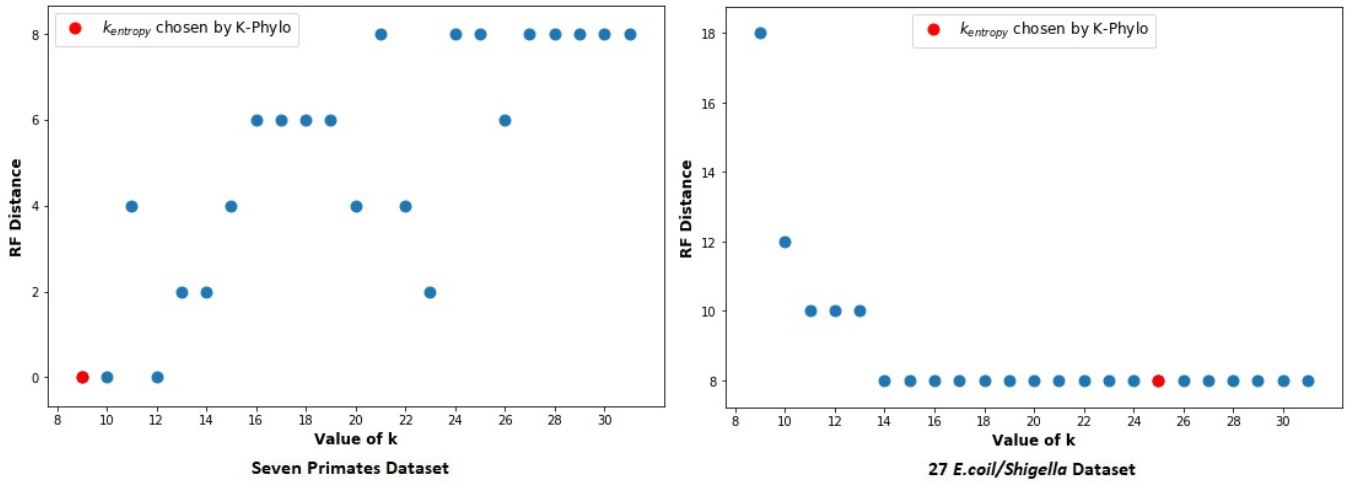

Figure 4: Choosing a suitable k-mer length in *K-Phylo* - while different k-mer lengths result in trees having different RF distances, *K-Phylo* picks $k_{entropy}$ that results in a tree having the best RF distance possible on seven primates and the 27 *E.coli/Shigella* strains dataset.

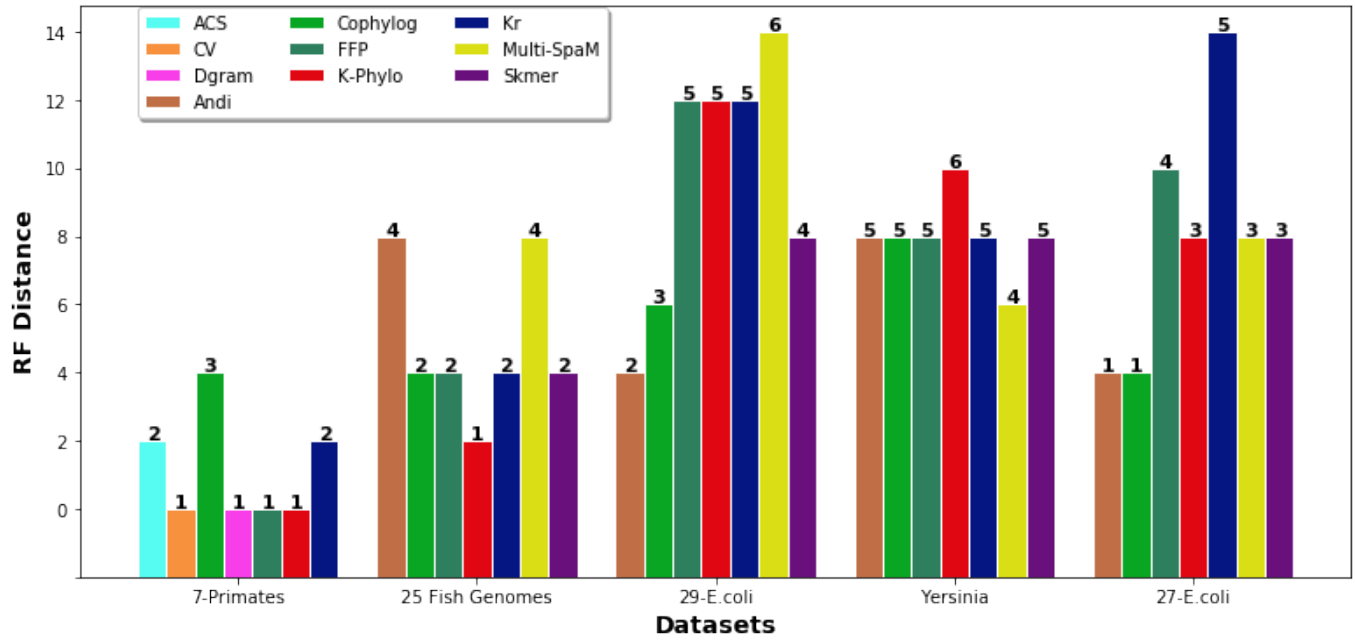

Figure 5: RF distance comparison among *K-Phylo* and several different methods on real datasets. Labels on top of the bars represent the corresponding method's respective rank from [6] (for the seven primates dataset) and AFproject [7] (for rest of the dataset). On the seven primate dataset, *K-Phylo* generates the exact benchmark tree leading to a RF distance of 0. This gives our method a ranking of 1. On the 25 fish species dataset, tree generated from *K-Phylo* has RF distance of 2, placing it in rank 1. On the 29 *E.coli* dataset, *K-Phylo* secures the 5th position, performing better than *Multi-SpaM*. On the *Yersinia* dataset, *K-Phylo* generates a RF distance of 10, securing the 6th position. Finally, on the 27 *E.coli* dataset, *K-Phylo* tree has RF distance of 8 resulting in rank 3.

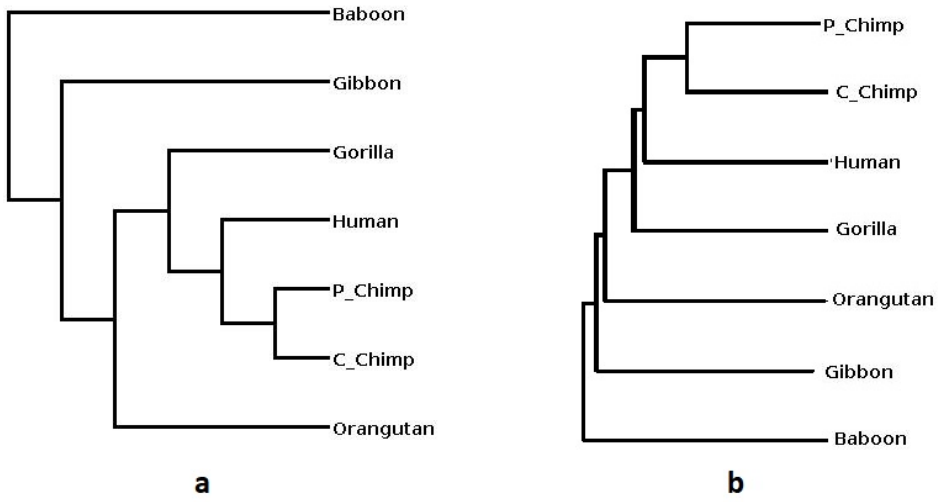

Figure 6: (a) Estimated phylogeny using *K-Phylo* and (b) the benchmark tree on seven primates dataset. [6] Here, *K-Phylo* achieves RF distance of 0 from the benchmark tree.

3.4 Primate Dataset

There are full genome sequences of seven primate species in this dataset. Here, *K-Phylo* produces an exact tree as the benchmark tree with RF distance of 0. The benchmark tree used is from [6]. Figure 5 shows that *K-Phylo* is better than or as good as all other methods. Although maximum entropy occurs at k equals 9, we found 0 RF distance for some other k-mer lengths as well. The tree estimated by our method along with the benchmark is demonstrated in Figure 6. The calculated entropy values and RF Distances corresponding to different values of k are listed in Table S2 of Supplementary Data.

3.5 25 *Fish Species*

In this dataset, there are assembled mitochondrial genomes of 25 fish species of the suborder *Labroidei*. On the best k-mer length (9), the resultant tree (Figure 7) from *K-Phylo* has RF distance of 2 which ranks it 1st among the 90 methods available in AFproject [7]. This can be visualized from Figure 5, which also highlights the fact that the estimated tree from *K-Phylo* has lower RF distance than that achieved by any other methods in the graph. Therefore, on this dataset, *K-Phylo* outperforms the methods it is compared with. The calculated entropy values and RF Distances corresponding to different values of k are listed in Table S2 of Supplementary Data.

3.6 29 *E.coli/Shigella* Strains

This dataset contains assembled 29 *E.coli/Shigella* strains. The tree from *K-Phylo* has RF distance of 12 on the best k-mer length (25), ranking it 5th among the 90 methods listed in AFproject [7]. Referring to Figure 5, it is evident that our method performed better than Multi-SpaM on this dataset, but did no better than some of the compared methods. The tree estimated by *K-Phylo* on this dataset and the benchmark tree are available in Figure S3 of Supplementary Data. The calculated entropy values and RF Distances corresponding to different values of k are listed in Table S3 of Supplementary Data.

3.7 *Yersinia* Strains

This dataset holds full genome sequences of 8 *Yersinia* strains. On the best k-mer length (32), *K-Phylo* generates an RF distance of 10 which gives it rank 6 among the 70 methods presented in AFproject [7]. On this dataset, each of the trees produced from the different k values (9 to 31) has RF distance of 10. From Figure 5, it can be seen that several other popular methods have RF distances very close to *K-Phylo*. It should be noted that results on this dataset are surprising as methods performing well on the rest of the datasets performed poorly here and vice versa as claimed in [15]. The tree estimated by *K-Phylo* on this dataset and the benchmark tree are available in Figure S1 of Supplementary Data. The calculated entropy values and RF Distances corresponding to different values of k are listed in Table S3 of Supplementary Data.

3.8 27 *E.coli/Shigella* Strains

This dataset contains 27 full genome sequences of *E.coli/Shigella* strains. *K-Phylo* achieves RF distance of 8 on its best k-mer length (25). This ranks our method 3 among the 70 methods available in AFproject [7]. Figure 5 reveals that our method performs better than *kr* and *FFP* on this dataset. The tree estimated by *K-Phylo* on this dataset and the benchmark tree are available in Figure S2 of Supplemen-

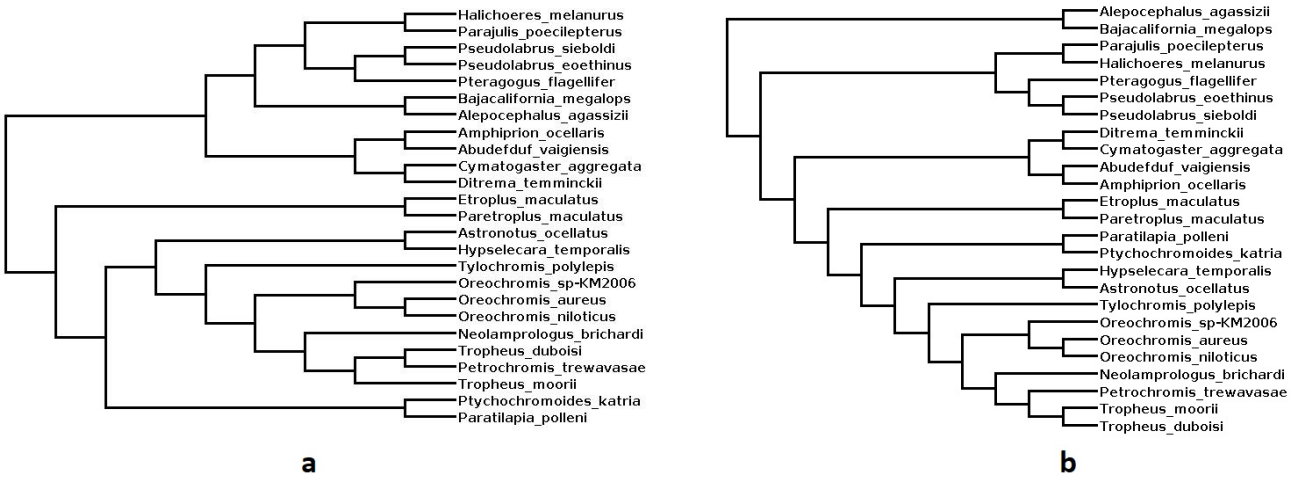

Figure 7: (a) Estimated phylogeny using *K-Phylo* and (b) the benchmark tree on 25 fish genomes dataset. *K-Phylo* has RF distance of 2 which ranks it 1st in comparison with the methods presented in AFproject[7].

tary Data. The calculated entropy values and RF Distances corresponding to different values of k are listed in Table S4 of Supplementary Data.

4 Conclusion

We presented an alignment free approach for phylogeny construction. It is based on presence or absence of k-mers in genomic sequences and estimates the tree using a maximum likelihood approach. While the method performs well on some datasets, it does not work well if the species involved are distant since in this case very few k-mers are conserved across the species. In future the performance of our method may be improved by considering k-mer counts instead of using presence or absence only. Moreover, the current version of this method uses an existing likelihood based phylogeny estimation tool and is concerned with tree topology only. Another future extension will be to develop a model to estimate branch lengths using this approach.

5 Supplementary Data

Additional information are available in the supplementary data.

References

[1] Charles Darwin. *On the origin of species, 1859*. Routledge, 2004.

[2] Robert R Sokal. A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, 38:1409–1438, 1958.

[3] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

[4] David W Mount. Maximum parsimony method for phylogenetic prediction. *Cold Spring Harbor Protocols*, 2008(4):pdb–top32, 2008.

[5] John P. Huelsenbeck. *Statistical Phylogenetics*. Wiley, 2011.

[6] Bernhard Haubold. Alignment-free phylogenetics and population genetics. *Briefings in Bioinformatics*, 15(3):407–418, 11 2013.

[7] Andrzej Zielezinski, Hani Z Girgis, Guillaume Bernard, Chris-Andre Leimeister, Kujin Tang, Thomas Dencker, Anna K Lau, Sophie Röhling, JaeJin Choi, Michael S Waterman, et al. Benchmarking of alignment-free sequence comparison methods. *BioRxiv*, page 611137, 2019.

[8] Huiguang Yi and Li Jin. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic acids research*, 41(7):e75–e75, 2013.

[9] Bernhard Haubold, Fabian Klötzl, and Peter Pfaffelhuber. andi: Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics*, 31(8):1169–1175, 2014.

[10] Thomas Dencker, Chris-Andre Leimeister, Michael Gerth, Christoph Bleidorn, Sagi Snir, and Burkhard Morgenstern. Multi-spam: a maximum-likelihood approach to phylogeny reconstruction based on multiple spaced-word matches. *arXiv preprint arXiv:1803.09222*, 2018.

[11] Chris-Andre Leimeister, Marcus Boden, Sebastian Horwege, Sebastian Lindner, and Burkhard Morgenstern. Fast alignment-free sequence comparison using spaced-word frequencies. *Bioinformatics*, 30(14):1991–1999, 2014.

[12] Michael Höhl and Mark A Ragan. Is multiple-sequence alignment required for accurate inference of phylogeny? *Systematic Biology*, 56(2):206–221, 2007.

[13] Qingpeng Zhang, Jason Pell, Rosangela Canino-Koning, Adina Chuang Howe, and C Titus Brown. These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure. *PloS one*, 9(7):e101271, 2014.

[14] Brian B Luczak, Benjamin T James, and Hani Z Girgis. A survey and evaluations of histogram-based statistics in alignment-free sequence comparison. *Briefings in Bioinformatics*, 20(4):1222–1237, 12 2017.

[15] Sophie Röhling, Thomas Dencker, and Burkhard Morgenstern. The number of k-mer matches between two dna sequences as a function of k. *BioRxiv*, page 527515, 2019.

[16] William B Sherwin. Entropy and information approaches to genetic diversity and its expression: genomic geography. *Entropy*, 12(7):1765–1798, 2010.

[17] Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[18] Andrew Rambaut and Nicholas C. Grass. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238, 06 1997.

[19] Víctor Soria-Carrasco, Gerard Talavera, Javier Igea, and Jose Castresana. The k tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. *Bioinformatics*, 23(21):2954–2956, 2007.

[20] Guillaume Bernard, Cheong Xin Chan, and Mark A Ragan. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. *Scientific reports*, 6:28970, 2016.

[21] 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, 2013.

[22] Igor Ulitsky, David Burstein, Tamir Tuller, and Benny Chor. The average common substring approach to phylogenomic reconstruction. *Journal of Computational Biology*, 13(2):336–350, 2006.

[23] Ji Qi, Bin Wang, and Bai-Iin Hao. Whole proteome prokaryote phylogeny without sequence alignment: a k-string composition approach. *Journal of molecular evolution*, 58(1):1–11, 2004.

[24] Bernhard Haubold, Peter Pfaffelhuber, Mirjana Domazet-Los o, and Thomas Wiehe. Estimating mutation distances from unaligned genomes. *Journal of Computational Biology*, 16(10):1487–1500, 2009.

[25] Gregory E Sims and Sung-Hou Kim. Whole-genome phylogeny of escherichia coli/shigella group by feature frequency profiles (ffps). *Proceedings of the National Academy of Sciences*, 108(20):8329–8334, 2011.

[26] David J Russell, Samuel F Way, Andrew K Benson, and Khalid Sayood. A grammar-based distance metric enables fast and accurate clustering of large sets of 16s sequences. *BMC bioinformatics*, 11(1):601, 2010.

[27] 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(1):34, 2019.

[28] Fábio Madeira, Young Mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian R N Tivey, Simon C Potter, Robert D Finn, and Rodrigo Lopez. The embl-ebi search and sequence analysis tools apis in 2019. *Nucleic acids research*, 47(W1):W636—W641, July 2019.