BMC Bioinformatics

RESEARCH

Open Access

An alignment-free method for phylogeny estimation using maximum likelihood

Tasfia Zahin[1†], Md. Hasin Abrar[1†], Mizanur Rahman Jewel[1], Tahrina Tasnim[1], Md. Shamsuzzoha Bayzid[1] and Atif Rahman[1*]

[†]Tasfia Zahin and Md. Hasin Abrar contributed equally to this work.

*Correspondence: atif@cse.buet.ac.bd

[1] Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka 1205, Bangladesh

Abstract

Background: While alignment has traditionally been the primary approach for establishing homology prior to phylogenetic inference, alignment-free methods offer a simplified alternative, particularly beneficial when handling genome-wide data involving long sequences and complex events such as rearrangements. Moreover, alignment-free methods become crucial for data types like genome skims, where assembly is impractical. However, despite these benefits, alignment-free techniques have not gained widespread acceptance since they lack the accuracy of alignment-based techniques, primarily due to their reliance on simplified models of pairwise distance calculation.

Results: Here, we present a likelihood based alignment-free technique for phylogenetic tree construction. We encode the presence or absence of *k*-mers in genome sequences in a binary matrix, and estimate phylogenetic trees using a maximum likelihood approach. A likelihood based alignment-free method for phylogeny estimation is implemented for the first time in a software named PEAFOWL, which is available at: https://github.com/hasin-abrar/Peafowl-repo. We analyze the performance of our method on seven real datasets and compare the results with the state of the art alignment-free methods.

Conclusions: Results suggest that our method is competitive with existing alignment-free tools. This indicates that maximum likelihood based alignment-free methods may in the future be refined to outperform alignment-free methods relying on distance calculation as has been the case in the alignment-based setting.

Keywords: Phylogenetics, Alignment-free, *k*-mer, Likelihood

Background

A phylogenetic tree depicts the evolutionary history of a given set of species. Efficient and accurate construction of phylogenies from genome data is one of the most important problems in biology and is a major research focus in bioinformatics and systematics. Phylogeny construction methods can be broadly classified into two groups: distance based and character based. Distance based methods compute the distances from the genomic sequences of each pair of species to construct a distance matrix. Tree

construction algorithms are then applied to this matrix to estimate the tree topology. Popular distance based methods include *UPGMA* [1], *neighbor-joining* [2], etc. They are fast and can handle many sequences but their performance is dependent on the accuracy of the distance matrix. Character based methods, on the other hand, make use of the sequences typically in the form of a multiple sequence alignment (MSA). *Maximum parsimony* [3] is a character based approach where a character matrix is taken as input and the best tree under the maximum parsimony criterion is the one that minimizes the number of changes in the nucleotide sequences over time. *Maximum likelihood* [4], a probabilistic character based approach, uses specific models of sequence evolution to find a tree that maximizes the likelihood of observing the set of input sequences. This approach is quite realistic in nature and suitable for species that vary widely in terms of similarity unlike the maximum parsimony approach.

Previous studies indicate that, in general, maximum likelihood approaches are superior in terms of performance over distance based methods. Maximum likelihood based methods were observed to estimate correct trees better than the neighbor joining method when the underlying assumptions behind the methods are not satisfied [5]. In addition, maximum likelihood based methods are also more robust than distance based methods using least square criterion [6].

However, in the alignment-based paradigm, both distance based and character based approaches require prior alignment of input sequences. The quality of alignment greatly affects the resulting phylogeny. Sequence alignment is memory and time consuming, and hence is difficult to scale to large sequences and whole genomes. Moreover, finding an optimal multiple sequence alignment is known to be computationally intractable as the number of possible alignments increases exponentially with increasing sequence lengths [7]. Furthermore, alignment-based methods assume a preserved linear order of homology, and therefore the presence of rearrangement events, such as translocation, inversion, etc. within whole genome sequences complicates sequence alignment—making it even more challenging to construct accurate phylogenetic trees from whole genomes [7].

To overcome the aforementioned difficulties, phylogenetic analyses that are not confined to alignment needs are gaining increasing attention, saving substantial time and memory in the phylogeny estimation process. The methods are collectively known as *alignment-free* methods. They are robust to rearrangement events and suitable for phylogeny estimation from large sequences and even whole genomes. However, despite their practical advantages, alignment-free techniques have not demonstrated the same level of accuracy as alignment-based methods. It is important to acknowledge that we do not anticipate alignment-free methods to match the accuracy of alignment-based methods, particularly when dealing with small, rearrangement-free sequences such as single genes. This is because alignment-free methods still require effective strategies for handling homology, a challenge that is no less complex than alignment itself.

A multitude of recently developed alignment-free methods have been comprehensively reviewed in [8, 9]. Among these, *co-phylog* [10] searches for short alignments of fixed length in the sequences allowing a mismatch in the middle. Evolutionary distances are calculated from these sub-sequences, followed by tree generation. *andi* [11] looks for mismatches surrounded by long exact matches. Counts of mismatches are used to estimate the number of substitutions between two sequences. *Mash* [12] is based on

the MinHash technique to find representative sketches of sequences from which Jaccard indices are estimated as a distance measure. *Multi-SpaM* [13] uses the *Space Word Match (SWM)* [14] 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.

However, despite their potential, alignment-free methods have not yet been found to be as accurate as alignment-based methods. Majority of the alignment-free methods developed so far are distance based and hence do not allow model based phylogeny estimation that are known to be more robust than the former. Höhl and Ragan [15] proposed a Bayesian approach for phylogeny inference based on the existence of *k*-mers (contiguous subsequences of length *k*) in the sequences.

In this paper, motivated by the observation that methods using maximum likelihood outperform distance based methods in the alignment-based setting, we present an alignment-free method for phylogenetic tree construction that utilizes maximum likelihood estimation. We first construct a matrix encoding the presence or absence of *k*-mers within the sequences, and then use an existing model for binary traits to construct a phylogeny that maximizes likelihood. The method is implemented in a tool called PEAFOWL (**P**hylogeny **E**stimation through **A**lignment **F**ree **O**ptimization **W**ith **L**ikelihood). We analyze the performance of our method by applying it on seven real datasets, including datasets from the AFproject [9] which is widely used for assessing alignment-free tools.

Methods

An overview of phylogenetic tree estimation using PEAFOWL is shown in Fig. 1. The method consists of four major steps. First, the set of *k*-mers present in each input sequence is generated. Second, a binary matrix is constructed, which encapsulates the presence or absence of the *k*-mers within the sequences. Third, a suitable value of *k*

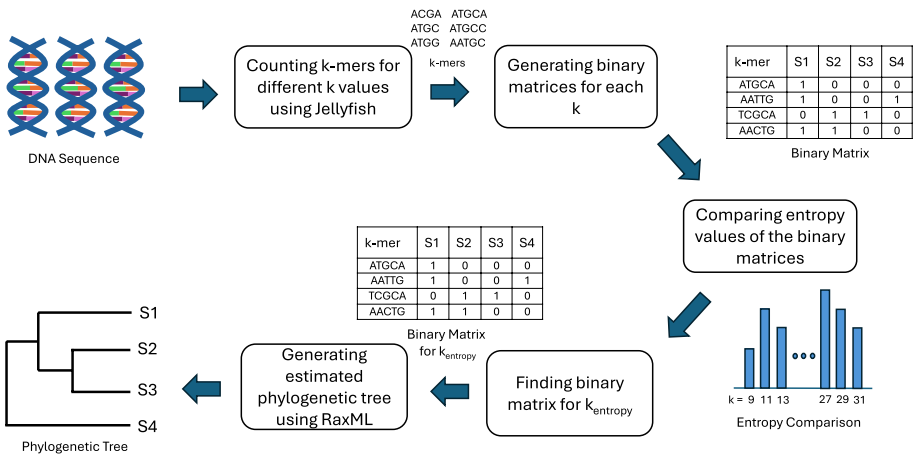


Fig. 1 Overview of phylogenetic tree estimation using PEAFOWL. At the beginning, *k*-mers of various sizes are listed from the input sequences using the *k*-mer counting tool Jellyfish. Then separate binary matrices are produced using these *k*-mers. From the binary matrices of different *k*-mer sizes, an appropriate *k*-mer length ($k_{entropy}$) is chosen based on cumulative entropy values. Lastly, the binary matrix corresponding to $k_{entropy}$ is provided as input to RAxML for the estimation of the phylogenetic tree

is chosen based on entropy values. Finally, a phylogenetic tree is constructed using maximum likelihood estimation. A sketch of the steps is presented in Algorithm 1 and described in more detail in the following sections.

Algorithm 1 Phylogeny estimation using PEAFOWL

Input: Set of genome sequences, $S = \{S_1, S_2, \ldots, S_m\}$
Output: Phylogenetic tree, T
$k_{min} \leftarrow 9$;
$k_{max} \leftarrow 31$;
Let C be a dictionary to store entropy values;
Let B be a dictionary to store binary matrices;
for $k \leftarrow k_{min}$ **to** k_{max} **by** 2 **do**
  Let $X_k \leftarrow [x_1, x_2, \ldots x_n]$ be the set of k-mers of size k from S;
  Let B_k be a matrix of size $n \times m$;
  for $i \leftarrow 1$ **to** n **do**
    for $j \leftarrow 1$ **to** m **do**
      if $x_i \in S_j$ **then**
        $B_k(i, j) \leftarrow 1$;
      else
        $B_k(i, j) \leftarrow 0$;
      end
    end
  end
  Let $R_k \leftarrow [r_1, r_2, \ldots, r_q]$ be q uniformly randomly selected rows from $[1, n]$;
  $C_k \leftarrow -\sum_{i=1}^{q} \sum_{x \in \{0,1\}} p_{r_i}(x) \log p_{r_i}(x)$ where
  $p_{r_i}(x) = \left(\sum_{j=1}^{m} \mathbb{1}_{B_k(r_i, j) = x} \right) / m$;
end
$k_{entropy} \leftarrow \arg\max_k C_k$;
$T \leftarrow RAxML(B_{k_{entropy}})$;
return T

Generating *k*-mers

The first step in PEAFOWL is to generate the lists of *k*-mers present in the input sequences. *k*-mers are generated from the input DNA sequences using Jellyfish [16] for odd values of *k* ranging from 9 to 31 (more details in Subsection Finding an appropriate k-mer length). As DNA is double-stranded and the sequences in the two strands are complements of each other, the input sequences can be from one strand or from both. In the latter case, it is more appropriate to consider a *k*-mer and its reverse complement as the same during counting, usually referred to as canonical counting. In the former case, a *k*-mer and its reverse complement can be treated independently, commonly known as non-canonical counting. Our method is designed to work in both possible modes, allowing the user to choose how reverse complements should be treated during the *k*-mer counting step. However, all the results shown in this paper except one (see Subsection Horizontal gene transfer (HGT)) are obtained using the canonical counting mode as the assembled sequences may correspond to either strand of DNA.

Generating binary matrices

The next step is to construct a binary matrix denoting whether the generated *k*-mers are present in the given sequences or not. This matrix consists of only 0's and 1's. Its rows and columns represent the *k*-mers and the input species, respectively. An entry in the matrix contains 1 if the *k*-mer representing the row (or its reverse complement) exists in the sequence of the species representing the column and 0 otherwise. One such matrix is produced for each value of *k*. We use hashing for this particular task. *k*-mers are read from a file, and a unique index is generated for each of them. The *k*-mers are inserted into a hash table along with the identification numbers of species they come from. The hash table indices are accessed one after another while placing an appropriate value in the desired position of the matrix.

Finding an appropriate *k*-mer length

A number of approaches have been proposed by researchers to choose a proper *k*-mer length for alignment-free analysis. [17] applies a logarithmic function on input sequence lengths to calculate a suitable value of *k*. The limitation of this selection process is that it does not take into account how closely related the species are. *Slope-SpaM* [18] analyzes match probability to calculate lower and upper bounds on *k*-mer length. However, it does not serve the need for a specific value of *k* that our model requires. More specifically, our target is to find a value of *k* that would capture the most informative binary matrix for tree generation.

The genetic diversity among the different genome sequences can be modeled by the concept of entropy [19]. This concept has been previously used by several other sequence analysis approaches [20, 21]. We utilize this in our method for *k*-mer length selection (Fig. 2). *k*-mers that can be found in almost all the species introduce many 1 s in the matrix, while rare ones introduce many 0 s. *k*-mers that do not fall in either of these extremities provide comparatively more information. A binary matrix rich in these types of *k*-mers will capture the relationship between species better than the others. Since entropy can capture the randomness in a system, we use this metric to compare the information content of the binary matrices, and choose an appropriate *k*-mer length. Cumulative entropy for a binary matrix is calculated using the following equation.

$$C^{entropy} = -\sum_{i=1}^{q} \sum_{x \in \{0,1\}} p_{r_i}(x) \log p_{r_i}(x)$$

where

$$p_{r_i}(x) = \frac{\sum_{j=1}^{m} \mathbb{1}_{X_{r_i,j}=x}}{m}$$

S1:AATGCCATAGCGCC
S2:ATGACCATCCGCCG
S3:AGTCATCAGCCGGC

Fig. 2 Choosing *k*-mer lengths. Existence of *k*-mers depends on the 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)

Here, q represents the number of k-mers used for entropy calculation, and m represents the total number of species. $X_{i,j}$ represents the state of an entry in the matrix corresponding to the i^{th} row and the j^{th} species, and can take values of either 0 or 1. Again, $\mathbb{1}_{X_{i,j}=x} = 1$ if $X_{i,j} = x$, and 0 otherwise. The equation adds the entropy values of q randomly selected rows $[r_1, r_2, \ldots, r_q]$ to get the cumulative entropy ($C^{entropy}$) for a binary matrix. Here, q is empirically chosen to be 5000.

Empirical evidence suggests that k values less than 9 cause k-mers to be excessively abundant, while those greater than 31 often lead to the presence of k-mers in only one or a few sequences [18]. For even values of k, a k-mer and its reverse complement may become the same causing inconsistency in the non-canonical counting mode [22]. To address these issues and to reduce the computational complexity, binary matrices are created for odd k-mer lengths ranging from 9 to 31. $C^{entropy}$ values from different k-mer lengths are compared and the value of k resulting in the maximum entropy is selected to be the most suitable one. We refer to this length as $k_{entropy}$.

Generating phylogenetic trees

Once we have obtained $k_{entropy}$, the final step is to construct a tree from the binary matrix corresponding to this length. This is done by providing the concerned matrix as input into a widely used tool for maximum likelihood phylogeny estimation named RAxML [23]. We use an existing model of substitution for binary traits BINGAMMA for our method. It is defined for binary data and assumes a gamma prior on the site mutation rates. The model takes in binary sequences and outputs an estimated tree topology, assuming sites to be independent. However, in reality, one character substitution in a sequence affects a number of neighboring k-mers at that site. This is why we focus on tree topology for now and leave branch length estimation as future work.

Implementation

Peafowl is implemented using C++ and shell scripts. In addition, it uses Jellyfish 2.2.4 for k-mer counting. A rigorous comparison of k-mer counting methods is presented by Zhang et al. [24]. We choose Jellyfish [16] on the basis of this comparison. The tool is fast, supports dynamic memory and is preferable for large genome sequences.

Peafowl also uses RAxML 8.2.4 for phylogeny estimation. RAxML stands for **R**andomized **Ax**elerated **M**aximum **L**ikelihood [23]. It is a popular phylogenetic analysis software that can handle large datasets and is useful for maximum likelihood based phylogeny inference.

Results

Datasets and benchmarking

We assess the performance of our method using seven real datasets. First, we analyze a 7 Primates dataset [8] and a Drosophila dataset from [25]. The 7 Primates dataset contains full mitochondrial genome sequences of 7 primates, and the Drosophila dataset consists of real genome skims of 14 Drosophila species subsampled to 100 Mb. We selected these datasets as the reference trees for these species are well established.

Next, we analyzed datasets from the AFproject [9] that have been widely used for benchmarking alignment-free methods. We selected the five real datasets under the

Genome-based Phylogeny and Horizontal Gene Transfer categories that had assembled genomes. They include assembled sequences of 29 *E.coli/Shigella* strains [10], assembled mitochondrial genomes of 25 fish species of the suborder Labroidei [26], full genome sequences of 14 plant species [27], full genome sequences of 27 *E.coli/Shigella* strains [28], and full genome sequences of 8 Yersinia strains [28].

Genome sequences, benchmark trees, and results of the last five datasets were obtained from the AFproject [9]. For the primates and Drosophila datasets, sequences and the benchmark trees are obtained from [8] and [25], respectively. The primary performance metric used throughout this paper is the *Robinson Foulds (RF)* [29] distance. It gives a measure of the distance between two trees by counting the number of dissimilar partitions. The distance is divided by the maximum possible RF value to obtain the normalized RF distance (nRF). The smaller this score, the more congruent the estimated and the reference trees.

Our method is run on these datasets (except one) with the-*r* parameter (canonical *k*-mer counting) i.e. reverse complements are considered the same *k*-mer. The tree corresponding to $k_{entropy}$ is treated as the final tree. The nRF distance between this tree and the reference tree is compared to those achieved by state-of-the-art methods from the AFproject [9]. The benchmarked methods include *FFP* [30, 31], *co-phylog* [10], *Mash* [12], *Skmer* [25] *FSWM/Read-SpaM* [32, 33] *andi* [11], *phylonium* [34], *Multi-SpaM* [13], and *CAFE-cvtree* [35]. It has been observed that no single method benchmarked by AFproject [9] achieves the best scores across all datasets. The aforementioned methods include the top performers.

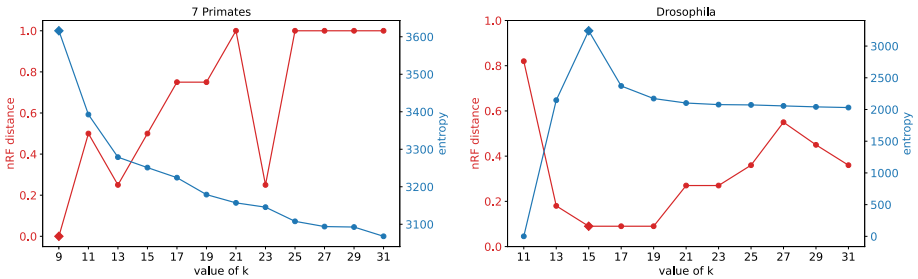
Some of the benchmarked tools generate a distance matrix as output and not a phylogenetic tree. Therefore, for the 7 primates and Drosophila datasets, we apply neighbor-joining and UPGMA implementation of MEGA-X [36] on the distance matrix to get the estimated trees, and find the RF distances using PHYLIP [37]. nRF values of benchmarked tools reported here are from trees produced by neighbor-joining. Results from UPGMA are available in the supplementary materials. We limit the scope of our work to phylogenetic trees upto a maximum of 30 species.

Selection of *k*-mer lengths

We first explore how the estimated trees vary with different *k*-mer lengths. The variation of nRF and entropy with change in *k*-mer size for the 7 Primates and Drosophila datasets are illustrated in Fig. 3. Similar plots for the remaining datasets are shown in Supplementary Figures S1–S5. We observe that, for the 7 Primates dataset, the minimum nRF distance of 0 and the maximum entropy is obtained when *k* equals 9. In all cases, we find that the lowest nRF distances occur at the *k*-mer lengths with the highest entropy i.e. $k_{entropy}$. We observe that in Fig. 3a there is a drop in nRF at k=23. This might be because the dataset contains only seven species, so a change in a single branch leads to a substantial decrease in the nRF value. In the subsequent sections, we only report nRF distances corresponding to the tree obtained using $k_{entropy}$.

7 Primates and drosophila datasets

The nRF distances for Peafowl and other methods for the 7 Primates and Drosophila datasets are demonstrated in Fig. 4 and Supplementary Table S1. Peafowl along with a



(a) (b)

Fig. 3 nRF and entropy vs. *k*-mer. Variation of normalized Robinson Foulds distance and entropy with change in *k*-mer length for **a** the 7-Primates dataset and **b** the Drosophila dataset. Diamond shaped markers represent values corresponding to $k_{entropy}$

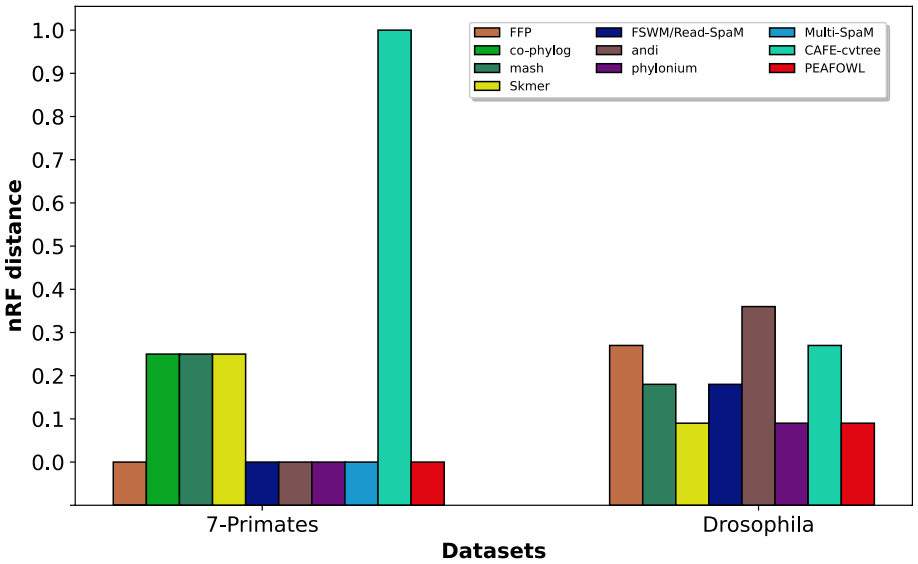


Fig. 4 Comparison of nRF distances. nRF distance comparison among PEAFOWL and state-of-the-art methods on the 7 Primates and Drosophila datasets

few other methods (e.g., andi, Multi-SpaM, FFP) correctly reconstructed the reference tree. It is worth noting that highly accurate methods like Mash, Skmer, and co-phylog placed Gibbon as the sister to hominine (gorillas, chimpanzees, and humans) and thus failed to reconstruct the well-established (orangutan, (gorilla, (chimpanzee, human))) relationship (see Fig. 5a).

For the *Drosophila* dataset, the trees with the lowest nRF distances were obtained by PEAFOWL, Skmer and phylonium (Fig. 4). Both PEAFOWL and Skmer produced the same tree which differs from the reference tree in one branch (see Fig. 5b). Skmer and PEAFOWL reconstructed the sister relationship of *Drosophila mauritiana* and *Drosophila simulans* which contradicts the reference tree supporting the (*Drosophila mauritiana*, (*Drosophila simulans*, *Drosophila sechellia*)) relationship.

An additional advantage of PEAFOWL is it allows generation of support values. Whereas most alignment-free tools produce only one tree, PEAFOWL can leverage

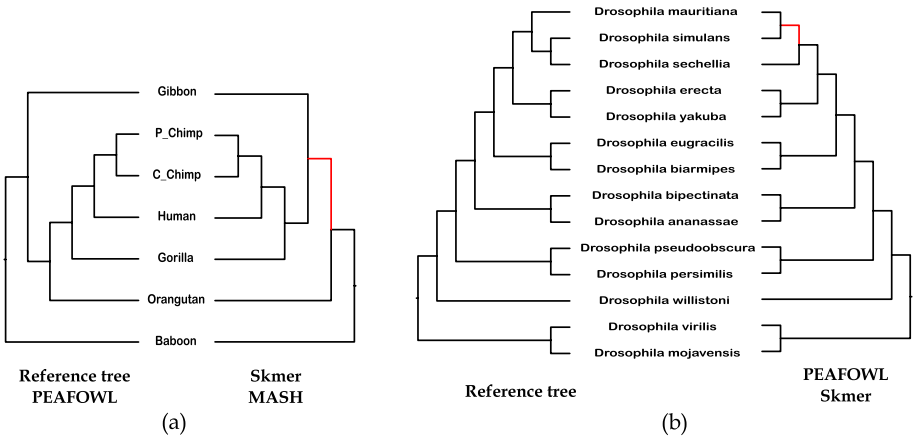


Fig. 5 Analysis of the Primate and Drosophila phylogenies. The internal branches in the estimated trees that are not found in the reference trees are shown in red. **a** The trees estimated by PEAFOWL (which is identical to the reference tree [8]), Skmer, and Mash. **b** The trees estimated by PEAFOWL and Skmer in comparison to the reference tree

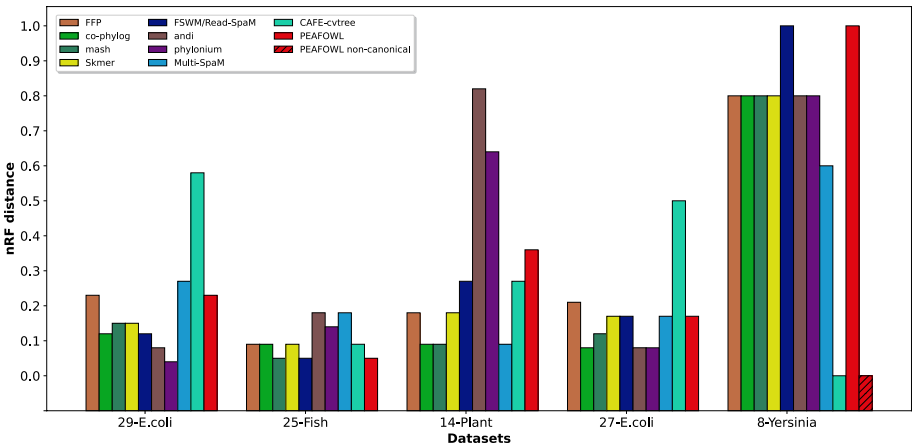


Fig. 6 Comparison of nRF distances on the AFproject datasets. nRF distance comparison among PEAFOWL and several different methods on real datasets from AFproject. Exact values can be found in supplementary materials

RAxML to produce bootstrap values, enhancing confidence in the results. Supplementary Figures S12 and S13 show the consensus trees for the primate and *Drosophila* datasets respectively along with the bootstrap support values generated using the rapid bootstrapping option in RAxML.

Genome-based phylogeny

The Genome-based Phylogeny group of the AFproject [9] includes assembled 29 *E.coli/Shigella* strains [10], assembled mitochondrial genomes of 25 fish species of the suborder Labroidei [26], and full genome sequences of 14 plant species [27]. A comparison of the nRF distances achieved by various methods is shown in Fig. 6. PEAFOWL attains nRF values of 0.23, 0.05, and 0.36 on these datasets, respectively.

On the 25 Fish dataset, our method is one of the best performing tools, achieving the lowest nRF distance of 0.05 along with Mash and FSWM. Estimated and reference trees for fish genome are shown in Supplementary Figure S6.

However, on the 29 *E.coli/Shigella* and the 14 plant datasets, PEAFOWL is outperformed by other methods. The best performing method on the 29 *E.coli/Shigella* dataset is phylonium whereas co-phylog, Mash and Multi-SpaM generate the most accurate trees on the 14 plant dataset.

Estimated and reference trees for the 29 *E.coli* and 14 plant datasets are available in Supplementary Figure S7 and S8. It is worth noting that the reference tree for the 29 *E.coli* dataset was constructed using an alignment-based approach from the assembled genomes [10] and has not been thoroughly validated subsequently. For the plant dataset, the $k_{entropy}$ value in PEAFOWL was calculated based on running the method over a k-mer range of 9 to 17 instead of 9 to 31 to avoid resource exhaustion. Results are not included in the entropy variation plot for k equals 9 and 17 due to the presence of all 1's in the binary matrix, resulting in zero entropy and computational limitation, respectively.

Horizontal gene transfer (HGT)

This category of data from the AFproject [9] includes full genome sequences of 27 *E.coli/Shigella* strains [28] and 8 Yersinia strains [28]. These two datasets are known to have undergone extensive genome rearrangements [9]. They exhibit horizontal gene transfer properties that may cause distant species to show sibling-like properties (such as similar k-mers).

The performance of various alignment-free tools on the Yersinia dataset is shown in Fig. 6. An observation was made previously [9] that whole-genome analysis tools tend to construct trees relatively discordant to the reference tree on Yersinia sequences than traditional approaches. This seems true for PEAFOWL as well, with an nRF of 1 on this dataset. The best performing method in this dataset is CAFE-cvtree. However, most tools perform poorly in this case, with only two having an nRF value below 0.8. It has been conjectured that the complex nature of the genus and substantial rearrangement events may promote this discrepancy [9].

We further explored this issue and noted that the eight Yersinia genomes are very similar in sequence but share genome rearrangements, and the reference tree was constructed using genomic inversion events inferred from a whole-genome alignment [28]. Since in the canonical counting mode, k-mers and their reverse complements are counted together, inversion events are undetected except at the two ends of the inversions. So, we also run our method without the-r parameter i.e. perform non-canonical counting, implying that k-mers and their reverse complements are treated as separate entities by the counting tool. Remarkably, with the non-canonical counting mode, PEAFOWL reconstructed a tree that is identical to the reference tree (Fig. 7, Supplementary Figures S9, S10).

Supplementary Tables S2 and S3 report the entropy and nRF values (corresponding to the highest entropy) achieved by PEAFOWL in the canonical and non-canonical settings, respectively. We observe that the entropy values in the non-canonical mode are

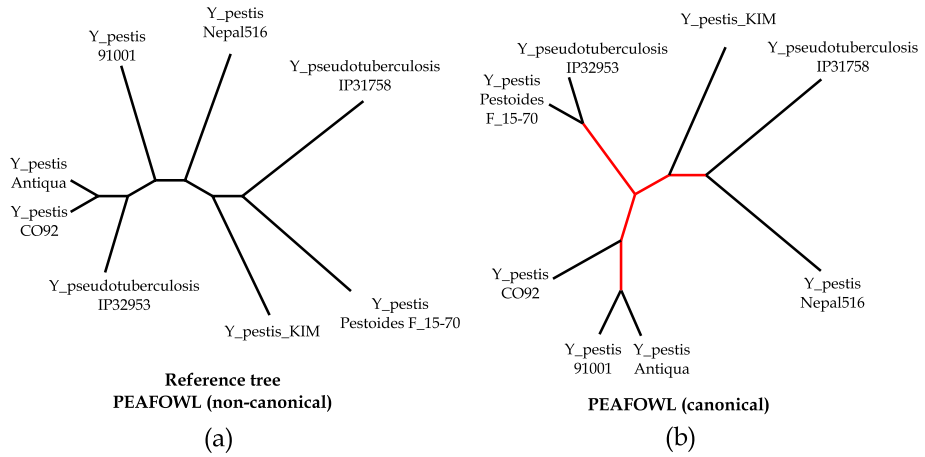


Fig. 7 Analysis of the Yersinia phylogenies. **a** The tree estimated by PEAFOWL with non-canonical counting mode, which is identical to the reference tree. **b** PEAFOWL-estimated tree with canonical mode of counting. The branches in the estimated tree that differ from the reference tree are shown in red

Table 1 Runtime and peak memory usage of PEAFOWL

Dataset	Size	Time	Peak memory usage (GB)
7 Primates	114 Kb	36 sec	5
14 *Drosophila*	2.1 Gb	25 hour 16 min 12 sec	23
25 Fish	411 Kb	1 min 12 sec	5
14 Plant	4.47 Gb	8 hour 18 min 0 sec	40
29 *E.coli*	137 Mb	1 hour 20 min 24 sec	6
27 *E.coli*	128 Mb	1 hour 38 min 24 sec	6
8 *Yersinia*	35.7 Mb	28 min 12 sec	6

substantially higher than those in the canonical mode for the Yersinia dataset. On the 27 *E.coli/Shigella* dataset, our method achieves nRF distance of 0.17, but the best performers include co-phylog, andi and phylonium with an nRF of 0.08 (Fig. 6). Estimated and reference trees for *E.coli/Shigella* are shown in Supplementary Figure S11.

Runtime and memory usage

All the datasets are run in an AMD Ryzen 9 7950X 16-Core Processor machine with 64 GB RAM. Table 1 summarizes the unzipped size of the datasets used, corresponding runtime, and peak memory usage by PEAFOWL. Values for the plant dataset are corresponding to a k-mer range of 9 to 17. A breakdown of runtimes of various steps of PEAFOWL as well as the dimensions of the k-mer presence/absence matrices are provided in Supplementary Table S4.

Conclusions

In this paper, we presented PEAFOWL, an alignment-free method for phylogeny estimation using maximum likelihood. It circumvents the complexity of multiple sequence alignment and combines the merits of maximum likelihood estimation in tree

construction. We evaluated the performance of Peafowl on seven real datasets and compared the results with the state of the art alignment-free methods. We observe that Peafowl generates trees with the lowest nRF distances in three of the datasets, while phylonium acheives the lowest nRF in four datasets (Figs. 4 and 6). Moreover, the tree estimated by Peafowl on one other dataset (Yersinia dataset) matches the reference tree when it is run in a mode suited to capture inversions. Our experimental results suggest that the performance of various methods may substantially vary across different datasets. Therefore, selecting suitable methods becomes particularly challenging when the data are heterogeneous, which is often so for genome-scale phylogenetic data. Consequently, alignment-free tree estimation, far from being a "solved problem", merits further attention and improvements.

Peafowl has several limitations and can be extended and improved in a number of ways. First, it does not work well if the species are distant since very few *k*-mers are conserved across the species in this case. Its performance also suffers if sequences contain considerable missing regions. In the future, these issues may be addressed. Second, the current version works on assembled sequences or genomes. A future direction will be to extend it to support phylogeny estimation from unassembled sequencing reads. Third, trees are presently estimated using an existing model for binary traits based on presence or absence of *k*-mers, and actual counts are ignored. As such, the estimated branch lengths are often inaccurate. In the future, models suitable for *k*-mer counts may be developed, which may then be utilized to accurately infer branch lengths. Finally, our tool is substantially slower than distance-based alignment-free methods which limited our experiments to datasets containing up to 30 taxa. We find that the step that combines the Jellyfish generated *k*-mer lists of different species into the *k*-mer presence/absence matrices for various values of *k* to select the *k*-mer length corresponding to the maximum entropy takes most of the time. A future direction will be to resolve this issue to accommodate larger taxonomic groups. Moreover, a sketch-based approach such as Mash [12] can be explored and only a subset of *k*-mers may be considered.

Abbreviations

UPGMA	Unweighted pair group method with arithmetic mean
MSA	Multiple sequence alignment
SWM	Space word match
HGT	Horizontal gene transfer
nRF	Normalized Robinson–Foulds

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06080-w.

Supplementary file 1. An alignment-free method for phylogeny estimation using maximum likelihood" contains all supplementary tables and figures.

Author contributions

AR and MSB conceived and supervised the project. TZ, MHA, MR, TT, MSB and AR designed the methodology. TZ, MHA, MR, TT developed the software. TZ and MHA performed the analysis. TZ, MHA, MSB and AR wrote and edited the manuscript.

Funding

This research received no specific grant from any funding agency.

Availability of data and materials

The 7 primates dataset is available at http://guanine.evolbio.mpg.de/aliFreeReview/. The Drosophila dataset can be downloaded from https://github.com/danrdanny/Drosophila15GenomesProject/tree/master/assembledGenomes

which was processed according to [25]. The AF project datasets are available at https://afproject.org/app/. The code to generate the results is available at https://github.com/hasin-abrar/Peafowl-repo.

Declarations

Ethical approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
We declare that the authors have no Conflict of interest as defined by BMC, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Received: 31 October 2024 Accepted: 10 February 2025
Published online: 07 March 2025

References

1. Sokal RR. A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull. 1958;38:1409–38.
2. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406–25.
3. Mount DW. Maximum parsimony method for phylogenetic prediction. Cold Spring Harbor Protocols. 2008;2008(4):32.
4. Huelsenbeck JP. Statistical phylogenetics. Hoboken: Wiley; 2011.
5. Huelsenbeck JP. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Mol Biol Evolut. 1995;12(5):843–9.
6. Yang Z. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst Biol. 1994;43(3):329–42.
7. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol. 2017;18:1–17.
8. Haubold B. Alignment-free phylogenetics and population genetics. Briefings Bioinform. 2013;15(3):407–18.
9. Zielezinski A, Girgis HZ, Bernard G, Leimeister C-A, Tang K, Dencker T, Lau RS, Choi JJ, Waterman MS, et al. Benchmarking of alignment-free sequence comparison methods. Genome Biol. 2019;20(1):1–18.
10. Yi H, Jin L. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. Nucleic Acids Res. 2013;41(7):e75–e75.
11. Haubold B, Klötzl F, Pfaffelhuber P. Andi: fast and accurate estimation of evolutionary distances between closely related genomes. Bioinformatics. 2014;31(8):1169–75.
12. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy Adam M. Mash: fast genome and metagenome distance estimation using Minhash. Genome Biol. 2016;17(1):132.
13. Dencker T, Leimeister CA, Gerth M, Bleidorn C, Snir S, Morgenstern B.'multi-spam': a maximum-likelihood approach to phylogeny reconstruction using multiple spaced-word matches and quartet trees. NAR Genom Bioinform 202; 2(1): 13.
14. Leimeister C-A, Boden M, Horwege S, Lindner S, Morgenstern B. Fast alignment-free sequence comparison using spaced-word frequencies. Bioinformatics. 2014;30(14):1991–9.
15. Höhl M, Ragan MA. Is multiple-sequence alignment required for accurate inference of phylogeny? Syst Biol. 2007;56(2):206–21.
16. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–70.
17. Luczak BB, James BT, Girgis HZ. A survey and evaluations of histogram-based statistics in alignment-free sequence comparison. Briefings Bioinform. 2017;20(4):1222–37.
18. Röhling S, Linne A, Schellhorn J, Hosseini M, Dencker T, Morgenstern B. The number of k-mer matches between two DNA sequences as a function of k and applications to estimate phylogenetic distances. PLoS ONE. 2020;15(2): e0228070.
19. Sherwin WB. Entropy and information approaches to genetic diversity and its expression: genomic geography. Entropy. 2010;12(7):1765–98.
20. Das S, Deb T, Dey N, Ashour AS, Bhattacharya DK, Tibarewala DN. Optimal choice of k-mer in composition vector method for genome sequence comparison. Genomics. 2018;110(5):263–73.
21. Wu YQ, Yu Z-G, Tang R-B, Han G-S, Anh VV. An information-entropy position-weighted k-mer relative measure for whole genome phylogeny reconstruction. Front Genet. 2021;12: 766496.
22. Zentgraf J, Rahmann S. Fast gapped k-mer counting with subdivided multi-way bucketed cuckoo hash tables. In 22nd International Workshop on Algorithms in Bioinformatics (WABI 2022). Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.
23. Stamatakis A. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
24. Zhang Q, Pell J, Canino-Koning R, Howe AC, Brown CT. These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure. PloS one. 2014;9(7): e101271.

25. Sarmashghi S, Bohmann K, Gilbert MTP, Bafna V, Mirarab S. Skmer: assembly-free and alignment-free sample identification using genome skims. Genome Biol. 2019;20(1):34.
26. Fischer C, Koblmüller S, Gülly C, Schlötterer C, Sturmbauer C, Thallinger GG. 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. 2013;8(6): e67048.
27. Hatje K, Kollmar M. A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. Front Plant Sci. 2012;3:192.
28. Bernard G, Chan CX, Ragan MA. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. Sci Rep. 2016;6:28970.
29. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci. 1981;53(1–2):131–47.
30. Sims GE, Jun SR, Wu GA, Kim SH. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc Nat Acad Sci. 2009;106(8):2677–82.
31. Choi JJ, Kim S-H. A genome tree of life for the fungi kingdom. Proc Natl Acad Sci. 2017;114(35):9391–6.
32. Leimeister C-A, Sohrabi-Jahromi S, Morgenstern B. Fast and accurate phylogeny reconstruction using filtered spaced-word matches. Bioinformatics. 2017;33(7):971–9.
33. Lau A-K, Dörrer S, Leimeister C-A, Bleidorn C, Morgenstern B. Read-spam: assembly-free and alignment-free comparison of bacterial genomes with low sequencing coverage. BMC Bioinformatics. 2019;20(20):638.
34. Klötzl F, Haubold B. Phylonium: fast estimation of evolutionary distances from large samples of similar genomes. Bioinformatics. 2020;36(7):2040–6.
35. Lu YY, Tang K, Ren J, Fuhrman JA, Waterman MS, Sun F. CAFE: aCcelerated Alignment-FrEe sequence analysis. Nucleic Acids Res. 2017;45(W1):W554–9.
36. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. Mega x: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35(6):1547–9.
37. Felsenstein Joseph. PHYLIP (phylogeny inference package), version 3.5 c. Joseph Felsenstein., 1993.

Publisher's Note