GRaphical footprint based Alignment-Free method (GRAFree) for reconstructing evolutionary Traits in Large-Scale Genomic Features

Aritra Mahapatra[1], Jayanta Mukherjee[2]

Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur, India, 721302

1 aritra.mhp@iitkgp.ac.in
2 jay@cse.iitkgp.ac.in

abstract

In our study, we attempt to extract novel features from mitochondrial genomic sequences reflecting their evolutionary traits by our proposed method GRAFree (GRaphical footprint based Alignment-Free method). These features are used to build a phylogenetic tree given a set of species from insect, fish, bird, and mammal. A novel distance measure in the feature space is proposed for the purpose of reflecting the proximity of these species in the evolutionary processes. The distance function is found to be a metric. We have proposed a three step technique to select a feature vector from the feature space. We have carried out variations of these selected feature vectors for generating multiple hypothesis of these trees and finally we used a consensus based tree merging algorithm to obtain the phylogeny. Experimentations were carried out with 157 species covering four different classes such as, Insecta, Actinopterygii, Aves, and Mammalia. We also introduce a measure of quality of the inferred tree especially when the reference tree is not present. The performance of the output tree can be measured at each clade by considering the presence of each species at the corresponding clade. GRAFree can be applied on any graphical representation of genome to reconstruct the phylogenetic tree. We apply our proposed distance function on the selected feature vectors for three naive methods of graphical representation of genome. The inferred tree reflects some accepted evolutionary traits with a high bootstrap support. This concludes that our proposed distance function can be applied to capture the evolutionary relationships of a large number of both close and distance species using graphical methods.

Introduction

In studying phylogeny of different species using molecular data, mostly mutations, insertion, and deletion of residues in various homologous segments of DNA sequences are observed by computational biologists[18], [27]. This approach is sensitive to the selection of segments (e.g. genes, coding segments, etc.) of the sequence. Moreover, the homologous segments are very small portions ($< 2\%$) of the whole genome [51]. The roles of majority

of the genome sequences ($\approx$ 98%) are unknown. Hence those parts are considered as *"junk"* [15], [49].

The mitochondrial genomes (mtDNA) are relatively simpler than the whole genome. It consists of a limited number of genes, tRNAs, etc. Moreover, the "junk" segments are negligible with respect to the length of the mtDNA (generally $\approx$ 1%). Most importantly, mtDNA are haploid, inherited maternally in most animals [12], and recombination is very rare event in it [16]. So the changes of mtDNA sequence occur mainly due to mutations.

There are various challenges in using mtDNA sequences in computation and analysis. During the evolution process the genes of mtDNA very often change their order within the mtDNA and also get fragmented [33], [22], [5], [2]. This violates the collinearity of homologous regions very often [77]. The length of the mtDNA as well as the length of genes are also different for different species which makes it difficult to align the homologous regions. Apart from these facts, the complexity, versatility, and the huge length of the data make it difficult to develop any simple method in comparative genomics [46]. Conventional methods compute the distance between sequences through computationally intensive process of multiple sequence alignment [29], which remains a bottleneck in using whole genomic sequences for constructing phylogeny [25]. As a result, there exist a few works which attempt to discover evolutionary features in the larger apparent non-homologous regions of the genomic sequences using alignment-free methods.

The existing alignment-free methods can be broadly categorized into four types:

1. **k-mer/word frequency based methods:** The comparison between two sequences are derived by the variation of the frequency of optimized k-mer. Feature frequency profile (FFP) [62], [63], composition vector (CV) [69], return time distribution (RTD) [30], frequency chaos game representation (FCGR) [28] [23]) used this method.

2. **Substring based methods:** The pairwise distances are measured by the average length of maximum common substrings of two sequences, e.g., average common substring (ACS) [67], average common substring with k-mismatches (ACS-k) [35], mutation distance [24].

3. **Information theory based methods:** The alignment-free sequence comparison method become effective by inheritting different concepts from information theory like entropy, mutual information, etc. Base base correlation [43], Information correlation and partial information correlation (IC-PIC) [20], and Lempel-Ziv compress [50] proposed various information theory based methods to compare two sequences.

4. **Graphical representation based methods:** Here the DNA/amino acid sequences are represented in multidimensional space. The pairwise distances are obtained by comparing the graphs. Iterated map (IM) [1] adopted this technique to compute the distance between two sequences.

Most of the methods have various limitations. They are not suitable to deal with a large number of taxa, and the size of the input sequences is also limited [6, 45, 7]. It is found that an online tool named Alfree [77] can accept the total length of all sequences up to two lakhs. Similarly another

online tool CVTree3 [55], [78] works on coding sections only. The offline version of CVTree [55], [78], D_2^* [59], [68] is a very expensive process with respect to both memory and time. More over, the genomic data often works better or increase support for smaller datasets. For the larger dataset of very diverge species the phylogenetic tree construction methods have often failed [54].

Due to these difficulties, conventional methods of phylogenetic reconstruction are restricted to working with whole genome sequences as well as large dataset. For the last three decades, several methods have been introduced to represent the DNA sequence mathematically (both numerically and graphically) [48]. It has been hypothesized that each species carries unique patterns over their DNA sequence which makes a species different from others [34]. Exploration of those distributions for unique characterization is the key motivation behind the mathematical representation of a genome. There exist various representations of the large genome sequences through line graph by mapping the nucleotides to various numeric representations. Considering the genome sequences as the signal (called genomic signal), these methods analyze respective sequences using different signal processing techniques. Several techniques have been proposed to represent DNA sequences graphically in 2D space. One of the way to represent is by considering the structural groups of DNA sequences, such as purine (A, G) and pyrimidine (C, T) [47], amino (A, C) and keto (T, G) [37], and strong H-bond (C, G) and weak H-bond (A, T) [21]. The graphical representation has inherent a serious limitation of overlapping paths which causes loss of information [56]. In some techniques, the sequence is represented as an entity in higher dimensions such as, in 3-D [57, 39, 9, 26, 42], 4-D [11], 5-D [40], and 6-D [41]. The increase of dimension reduces the probability of occurrence of degeneracy, but it causes difficulty in visualization. In few schemes, like Worm Curve [58], DV-Curve [75], cell representation [71], etc., the DNA sequences are represented in a non-overlapping fashion.

GRAFree can be applied on any graphical method. GRAFree also lifts the loss of information due to overlapping paths by considering the coordinates of each nucleotide. In this study, we consider three sets of structural groups of nucleotides (purine, pyrimidine), (amino, keto), and (weak H-bond, strong H-bond) separately for representing DNA by a sequence of points in a 2-D integral coordinate space. This point set is called **Graphical Foot Print (GFP)** of a DNA sequence. We propose a technique for extracting features from GFPs and use them for constructing phylogenetic trees. As there are three different types of numerical representation of nucleotides, there are three different hypotheses for the phylogeny. Each of them is found to be statistically significant compared to a tree randomly generated. We also generate a consensus tree from these three hypotheses by applying a tree merging algorithm called COSPEDTree [3, 4].

Experimentations were carried out with a large dataset of total 157 species from four different classes, namely, Insecta (insect), Actinopterygii (ray-finned fish), Aves (bird), and Mammalia (Mammal).

The contributions made in this work are highlighted below:

- Revisiting the concept of Graphical Foot Print (GFP), 2D representation of DNA sequences, and introducing the concept of drift in GFP, which is found to be translation invariant for a sequence.

- Representation of a fragment of drift by a novel 5 dimensional descriptor. All the 5-D descriptors together represent a genotype char-

acteristic for a species.

- Proposed a new distance function to measure the dissimilarity among species, and use the distance matrix for generating phyolgenetic tree by distance based methods such as UPGMA [64]. The distance function is proved to be a metric.

- Proposed a technique to select the value of the parameters involved in computing the distance matrix.

1 Materials and Methods

Feature space

Definition 1. ***Graphical Foot Print (GFP).***
Let a sequence, $\mathcal{S} \in \Sigma^+$, $\Sigma = \{A, T, G, C\}$. For each combination of Purine (R)/Pyrimidine (Y), Amino (M)/Keto (K), and Strong H-bond (S)/Weak H-bond (W),the GFP of $\mathcal{S}$, $\phi(\mathcal{S})$, is the locus of 2-D points in an integral coordinate space, such that (x_i, y_i) is the coordinate of the alphabet s_i, $\forall\ s_i \in \mathcal{S}$, for $i = 1, 2, \ldots, n$, and $x_0 = y_0 = 0$.
Case-1: for Purine/Pyrimidine

$$\begin{aligned} x_i &= x_{i-1} + 1; \quad \textit{if } s_i = G \\ &= x_{i-1} - 1; \quad \textit{if } s_i = A \\ &= 0; \quad \textit{otherwise} \\ y_i &= y_{i-1} + 1; \quad \textit{if } s_i = C \\ &= y_{i-1} - 1; \quad \textit{if } s_i = T \\ &= 0; \quad \textit{otherwise} \end{aligned} \tag{1}$$

Case-2: for Strong H-bond/Weak H-bond

$$\begin{aligned} x_i &= x_{i-1} + 1; \quad \textit{if } s_i = C \\ &= x_{i-1} - 1; \quad \textit{if } s_i = G \\ &= 0; \quad \textit{otherwise} \\ y_i &= y_{i-1} + 1; \quad \textit{if } s_i = T \\ &= y_{i-1} - 1; \quad \textit{if } s_i = A \\ &= 0; \quad \textit{otherwise} \end{aligned} \tag{2}$$

Case-3: for Amino/Keto

$$\begin{aligned} x_i &= x_{i-1} + 1; \quad \textit{if } s_i = A \\ &= x_{i-1} - 1; \quad \textit{if } s_i = C \\ &= 0; \quad \textit{otherwise} \\ y_i &= y_{i-1} + 1; \quad \textit{if } s_i = T \\ &= y_{i-1} - 1; \quad \textit{if } s_i = G \\ &= 0; \quad \textit{otherwise} \end{aligned} \tag{3}$$

We denote GFPs of Case-1, Case-2 and Case-3, as GFP-RY (Φ_{RY}), GFP-SW (Φ_{SW}) and GFP-MK (Φ_{MK}), respectively.

Definition 2. ***Drift of GFP.***

Let $\mathcal{S}$ be a DNA sequence and s_i be the alphabet ($s_i \in \{A, T, G, C\}$) at the i^{th} position of $\mathcal{S}$. Let $\phi_i(\mathcal{S})$ denote the corresponding (x_i,y_i) the coordinate of s_i in $\Phi(\mathcal{S})$.

Then for length L, drift at the i^{th} position is defined as,

$\delta_i^{(L)} = \phi_{i+L}(\mathcal{S}) - \phi_i(\mathcal{S})$, *where* $(i + L) \leq |\mathcal{S}|$

Considering the drifts for every i^{th} location of the whole sequence, the sequence of drifts is denoted by

$\Delta^{(L)} = [\delta_0^{(L)}, \delta_1^{(L)}, \delta_2^{(L)}, \delta_3^{(L)}, \ldots, \delta_m^{(L)}]$, *where* $(m + L) = |\mathcal{S}|$

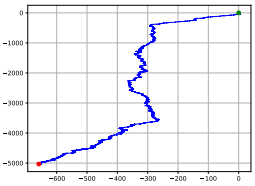
For GFP-RY (refer to Definition 1), an element (x_i, y_i) in $\Delta_{RY}^{(L)}$ provides excess numbers of G from A and C from T in segment of length L starting from the i^{th} location, respectively. Similarly, in GFP-SW, they are the excess numbers of C from G and T from A (represents as $\Delta_{SW}^{(L)}$), and in GFP-MK, they correspond to the excess numbers of A from C and T from G (represents as $\Delta_{MK}^{(L)}$), respectively.

We also call the elements of $\Delta^{(L)}$ as points, as they can be plotted on a 2-D coordinate system. We call this plot as the scatter plot of the drift sequence. Similarly, we get a scatter plot of a GFP. Compared to $\Phi_i(\mathcal{S})$, $\Delta^{(L)}$ is translation invariant as its set of points does not depend on the starting point of the sequence. It has been observed that in many cases the scatter plots of Δ have similar structure for closely spaced species mentioned in literature. In Fig. 1 we demonstrate the scatter plots of GFPs and drift sequences of two species from each class namely, *Drepanotermes* sp. and *Macrognathotermes errator* from insects, *Bathygadus antrodes* and *Bregmaceros nectabanus* from fishes, *Jacana jacana* and *Raphus cucullatus* from birds, *Canis familiaris* and *Panthera tigris tigris* from mammals. It can be observed that the species from same class (insect, fish, bird, or mammal) have the similar pattern in their drift sequences which intuitively indicates that the intraclass species are closer that the interclass species. It can also be observed that differences between two GFPs get reflected in their respective drifts. It is noted that the GFPs of *Bathygadus antrodes*, *Bregmaceros nectabanus*, and *Canis familiaris* (refer to Fig. 1 (c, d, g), respectively) have the similar patters, where as their drifts, shown in Fig. 1 (k, l, o), respectively, are quite different.

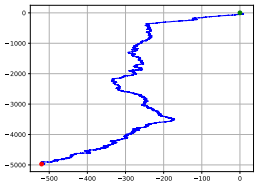
We represent spatial distribution of these points of Δ by an elliptical model using a five dimensional feature descriptor: $(\mu, \Lambda, \lambda, \theta)$, where $\mu = (\mu_x, \mu_y)$ is the center of the coordinates, Λ and λ are major and minor eigen values of the covariance matrix, and θ is the angle formed by the eigen vector corresponding to Λ with respect to the x-axis. We make $\mathcal{F}$ number of non overlapping equal length fragments from Δ and represent each fragment using the five dimensional feature descriptor.

Distance function and its properties

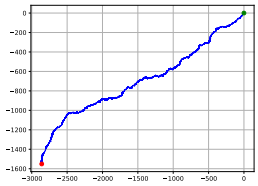
For two sequences $\mathcal{P}$ and $\mathcal{Q}$ with the feature descriptors of i^{th} fragments $(\mu_{\mathcal{P}i}, \Lambda_{\mathcal{P}i}, \lambda_{\mathcal{P}i}, \theta_{\mathcal{P}i})$ and $(\mu_{\mathcal{Q}i}, \Lambda_{\mathcal{Q}i}, \lambda_{\mathcal{Q}i}, \theta_{\mathcal{Q}i})$, where $i \leq \mathcal{F}$, $\mu_{\mathcal{P}i} = (\mu_{x\mathcal{P}i}, \mu_{y\mathcal{P}i})$ and $\mu_{\mathcal{Q}i} = (\mu_{x\mathcal{Q}i}, \mu_{y\mathcal{Q}i})$, we propose the following distance function between them,



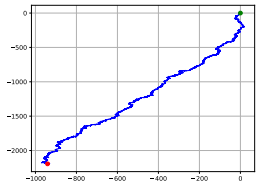
(a) ϕ of *Drepanotermes* sp.



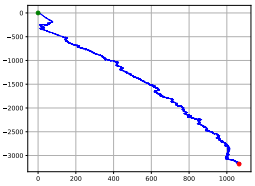
(b) ϕ of *Ma. errator*



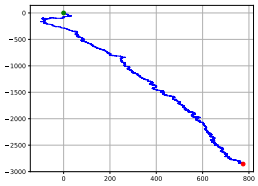
(c) ϕ of *Ba. antrodes*



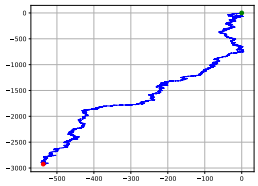
(d) ϕ of *Br. nectabanus*



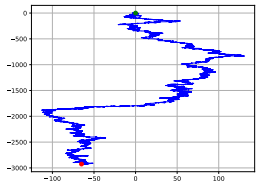
(e) ϕ of *J. jacana*



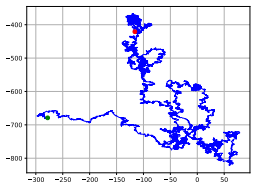
(f) ϕ of *R. cucullatus*



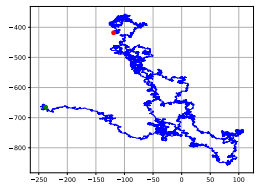
(g) ϕ of *C. familiaris*



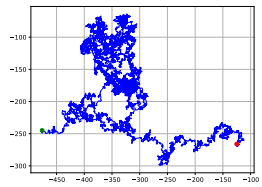
(h) ϕ of *P. t. tigris*



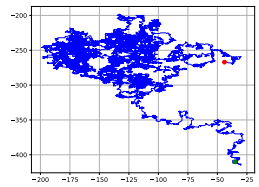
(i) $\Delta^{(2000)}$ of *Drepanotermes* sp.



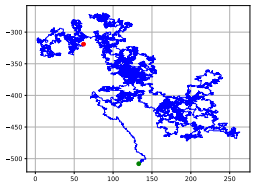
(j) $\Delta^{(2000)}$ of *Ma. errator*



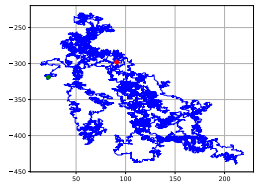
(k) $\Delta^{(2000)}$ of *Ba. antrodes*



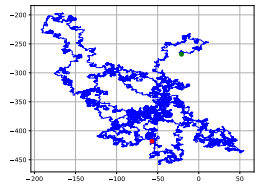
(l) $\Delta^{(2000)}$ of *Br. nectabanus*



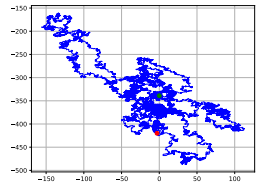
(m) $\Delta^{(2000)}$ of *J. jacana*



(n) $\Delta^{(2000)}$ of *R. cucullatus*



(o) $\Delta^{(2000)}$ of *C. familiaris*



(p) $\Delta^{(2000)}$ of *P. t. tigris*

Figure 1: Φ_{RY} and Δ_{RY} of few species from our dataset. Fig. 1 (a, b) are the Φ_{RY} generated from the insects namely, *Drepanotermes* sp. and *Macrognathotermes errator*, respectively. Fig. 1 (c, d) are the Φ_{RY} generated from the fishes namely, *Bathygadus antrodes* and *Bregmaceros nectabanus*, respectively. Fig. 1 (e, f) are the Φ_{RY} generated from the birds namely, *Jacana jacana* and *Raphus cucullatus*, respectively. Fig. 1 (g, h) are the Φ_{RY} generated from the mammals namely, *Canis familiaris* and *Panthera tigris tigris*, respectively. Fig. 1 (i-p) are the Δ_{RY} for $L = 2000$ of the corresponding species. The green and red dots are the start and end points of the graph, respectively.

$$D(\mathcal{P},\mathcal{Q}) = \frac{1}{\mathcal{F}}\sum_{i=1}^{\mathcal{F}}[\alpha\sqrt{{\mu_{\mathcal{P}_i}}^T\mu_{\mathcal{P}i} + {\mu_{\mathcal{Q}_i}}^T\mu_{\mathcal{Q}i} - 2{\mu_{\mathcal{P}_i}}^T\mu_{\mathcal{Q}i}cos(\theta_{\mathcal{P}i} - \theta_{\mathcal{Q}i})} + (1-\alpha)\sqrt{(\Lambda_{\mathcal{P}i} - \Lambda_{\mathcal{Q}i})^2 + (\lambda_{\mathcal{P}i} - \lambda_{\mathcal{Q}i})^2}] \quad where, \alpha = [0,1] \quad (4)$$

Lemma 1. *The distance D between two sequences is a metric. For any three sequences, $\mathcal{P}$,$\mathcal{Q}$ and $\mathcal{R}$, we have*

1. *Non-negativity.* $D(\mathcal{P},\mathcal{Q}) \geq 0$
2. *Identity.* $D(\mathcal{P},\mathcal{Q}) = 0$ *if and only if* $\mathcal{P} = \mathcal{Q}$
3. *Symmetry.* $D(\mathcal{P},\mathcal{Q}) = D(\mathcal{Q},\mathcal{P})$
4. *Triangle inequality.* $D(\mathcal{P},\mathcal{Q}) + D(\mathcal{Q},\mathcal{R}) \geq D(\mathcal{P},\mathcal{R})$

Proof. The properties 1, 2 and 3 can be proved from the definition itself. Here we prove the property 4.

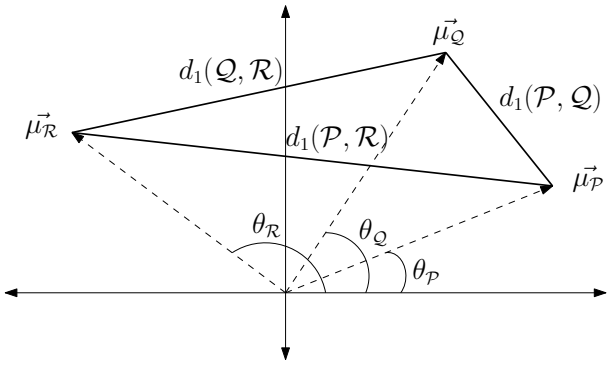


Figure 2: Computation of $D(\mathcal{P},\mathcal{Q})$

The distance between a single fragment of $\mathcal{P}$ and $\mathcal{Q}$ is,

$$\widehat{D}(\mathcal{P},\mathcal{Q}) = \alpha\sqrt{{\mu_{\mathcal{P}_i}}^T\mu_{\mathcal{P}i} + {\mu_{\mathcal{Q}_i}}^T\mu_{\mathcal{Q}i} - 2{\mu_{\mathcal{P}_i}}^T\mu_{\mathcal{Q}i}cos(\theta_{\mathcal{P}i} - \theta_{\mathcal{Q}i})} + (1-\alpha)\sqrt{(\Lambda_{\mathcal{P}i} - \Lambda_{\mathcal{Q}i})^2 + (\lambda_{\mathcal{P}i} - \lambda_{\mathcal{Q}i})^2} \quad where, \alpha = [0,1] \quad (5)$$

The distance between a single fragment of $\mathcal{Q}$ and $\mathcal{R}$ is,

$$\widehat{D}(\mathcal{Q},\mathcal{R}) = \alpha\sqrt{{\mu_{\mathcal{Q}_i}}^T\mu_{\mathcal{Q}i} + {\mu_{\mathcal{R}_i}}^T\mu_{\mathcal{R}i} - 2{\mu_{\mathcal{Q}_i}}^T\mu_{\mathcal{R}i}cos(\theta_{\mathcal{Q}i} - \theta_{\mathcal{R}i})} + (1-\alpha)\sqrt{(\Lambda_{\mathcal{Q}i} - \Lambda_{\mathcal{R}i})^2 + (\lambda_{\mathcal{Q}i} - \lambda_{\mathcal{R}i})^2} \quad where, \alpha = [0,1] \quad (6)$$

The distance between a single fragment of $\mathcal{P}$ and $\mathcal{R}$ is,

$$\widehat{D}(\mathcal{P},\mathcal{R}) = \alpha\sqrt{{\mu_{\mathcal{P}_i}}^T\mu_{\mathcal{P}i} + {\mu_{\mathcal{R}_i}}^T\mu_{\mathcal{R}i} - 2{\mu_{\mathcal{P}_i}}^T\mu_{\mathcal{R}i}cos(\theta_{\mathcal{P}i} - \theta_{\mathcal{R}i})} + (1-\alpha)\sqrt{(\Lambda_{\mathcal{P}i} - \Lambda_{\mathcal{R}i})^2 + (\lambda_{\mathcal{P}i} - \lambda_{\mathcal{R}i})^2} \quad where, \alpha = [0,1] \quad (7)$$

Let,

$$d_1(\mathcal{P},\mathcal{Q}) = \sqrt{\mu_{\mathcal{P}}^T\mu_{\mathcal{P}} + \mu_{\mathcal{Q}}^T\mu_{\mathcal{Q}} - 2\mu_{\mathcal{P}}^T\mu_{\mathcal{Q}}cos(\theta_{\mathcal{P}} - \theta_{\mathcal{Q}})}$$

$$d_1(\mathcal{Q},\mathcal{R}) = \sqrt{\mu_{\mathcal{Q}}^T\mu_{\mathcal{Q}} + \mu_{\mathcal{R}}^T\mu_{\mathcal{R}} - 2\mu_{\mathcal{Q}}^T\mu_{\mathcal{R}}cos(\theta_{\mathcal{Q}} - \theta_{\mathcal{R}})}$$

$$d_1(\mathcal{P},\mathcal{R}) = \sqrt{\mu_{\mathcal{P}}^T\mu_{\mathcal{P}} + \mu_{\mathcal{R}}^T\mu_{\mathcal{R}} - 2\mu_{\mathcal{P}}^T\mu_{\mathcal{R}}cos(\theta_{\mathcal{P}} - \theta_{\mathcal{R}})}$$

First, we prove,

$$d_1(P,Q) + d_1(Q,R) \geq d_1(P,R) \tag{8}$$

Fig. 2 represents the idea behind computing the distance between two sequences. So, from the figure we observe that $d_1(\mathcal{P},\mathcal{Q})$, $d_1(\mathcal{Q},\mathcal{R})$ and $d_1(\mathcal{P},\mathcal{R})$ form a triangle. Using triangulation inequality Eq. (8) can be proved.

For single fragment Let,

$$d_2(P,Q) = \sqrt{(\Lambda_{\mathcal{P}} - \Lambda_{\mathcal{Q}})^2 + (\lambda_{\mathcal{P}} - \lambda_{\mathcal{Q}})^2}$$

$$d_2(Q,R) = \sqrt{(\Lambda_{\mathcal{Q}} - \Lambda_{\mathcal{R}})^2 + (\lambda_{\mathcal{Q}} - \lambda_{\mathcal{R}})^2}$$

$$d_2(P,R) = \sqrt{(\Lambda_{\mathcal{P}} - \Lambda_{\mathcal{R}})^2 + (\lambda_{\mathcal{P}} - \lambda_{\mathcal{R}})^2}$$

Similarly, it can be proved that,

$$d_2(P,Q) + d_2(Q,R) \geq d_2(P,R) \tag{9}$$

Hence, by combining Eq. (8) and Eq. (9) it can be said that, $\widehat{D}(P,Q) + \widehat{D}(Q,R) \geq \widehat{D}(P,R)$.

Hence, $\widehat{D}$ is a metric. As D is the linear combination of $\widehat{D}$. So D is also be a metric. □

Taxon sampling and acquiring mitochondrial genome

We have selected various mitochondrial genome sequences sequenced by various researchers such as insects are selected from [8], [65], ray-finned fishes are selected from [61], [76], Aves data are selected from [19], [32], [53], and Mammalian data are selected from [31], [73], [74] [52]. We ignore those accession numbers which store some selected genes of mtDNA. Hence, we have studied over 157 species of four different classes - Insecta (insect), Actinopterygii (ray-finned fish), Aves (bird), and Mammalia (Mammal). The selected data have been downloaded from the NCBI database[1]. The average percentage of unrecognized nucleotide of all 157 mtDNA is 0.06% which inferred that the data we selected for this study are quite good in quality. Details of all the species are listed in Table 1.

Table 1: List of species

Species name	Accession number	Sequence length	A%	T%	G%	C%	Unrecognized%	AT%	GC%	AT skew	GC skew
Acinonyx jubatus	NC_005212.1	17047	33.10	27.53	13.58	25.79	0.00	60.63	39.37	0.09	-0.31
Ailuropoda melandeuca	EF196663.1	16747	31.83	29.36	14.92	23.87	0.03	61.19	38.78	0.04	-0.23
Allantus luctifer	KJ713152.1	15418	42.02	39.11	7.54	11.33	0.00	81.13	18.87	0.04	-0.20
Alopecoenas salamonis	KX902250.1	17141	30.93	24.25	13.32	31.50	0.00	55.18	44.82	0.12	-0.41
Anas platyrhynchos	EU009397.1	16604	29.20	22.21	15.78	32.81	0.00	51.41	48.59	0.14	-0.35
Apis mellifera syriaca	KP163643.1	15428	42.88	41.30	5.85	9.97	0.01	84.17	15.82	0.02	-0.26
Arctocephalus forsteri	NC_004023.1	15413	33.23	25.91	14.11	26.75	0.00	59.14	40.86	0.12	-0.31
Arctogadus glacialis	AM919429.1	16644	28.14	29.91	16.59	25.34	0.02	58.04	41.93	-0.03	-0.21
Ardea novaehollandiae	NC_008551.1	17511	31.77	23.41	13.44	31.37	0.00	55.19	44.81	0.15	-0.40
Boreogadus saida	NC_010121.1	16745	28.10	29.63	16.73	25.54	0.00	57.73	42.27	-0.03	-0.21
Nasutitermes triodiae	JX144940.1	15849	42.26	23.52	12.10	22.12	0.00	65.78	34.22	0.28	-0.29
Neofelis nebulosa	NC_008450.1	16844	31.72	27.13	14.79	26.37	0.00	58.85	41.15	0.08	-0.28

Continued on next page

[1]Website of NCBI database: `http://www.ncbi.nlm.nih.gov`

Table 1 – *Continued from previous page*

Species name	Accession number	Sequence length	A%	T%	G%	C%	Unrecognized%	AT%	GC%	AT skew	GC skew
Panthera tigris sumatrae	JF357970.1	17001	31.77	26.94	14.68	26.62	0.00	58.71	41.29	0.08	-0.29
Lota lota	AP004412.1	16527	28.37	27.59	16.32	27.72	0.00	55.96	44.04	0.01	-0.26
Halichoerus grypus	X72004.1	16797	32.96	25.30	14.28	27.46	0.00	58.27	41.73	0.13	-0.32
Didunculus strigirostris	KX902245.1	17071	30.58	24.24	13.71	31.47	0.01	54.82	45.18	0.12	-0.39
Megacrania alpheus adan	AB477471.1	17124	46.22	30.65	9.27	13.86	0.00	76.87	23.13	0.20	-0.20
Platalea leucorodia	KT901459.1	16846	31.12	24.26	13.84	30.78	0.00	55.38	44.62	0.12	-0.38
Cephus sareptanus	KM377624.1	15212	42.58	36.62	7.34	13.46	0.00	79.20	20.80	0.08	-0.29
Ixobrychus cinnamomeus	KJ190959.1	18640	32.25	25.25	13.23	29.27	0.00	57.51	42.49	0.12	-0.38
Melanogrammus aeglefinus	NC_007396.1	16585	28.50	30.49	16.33	24.67	0.00	58.99	41.01	-0.03	-0.20
Bathygadus antrodes	AP008988.1	17596	27.59	34.94	18.78	18.69	0.00	62.53	37.47	-0.12	0.00
Ursus malayanus	EF196664.1	16783	31.17	27.87	15.36	25.59	0.00	59.05	40.95	0.06	-0.25
Geotrygon violacea	NC_015207.1	16864	30.46	24.58	13.86	31.08	0.02	55.04	44.94	0.11	-0.38
Raphus cucullatus	KX902236.1	17092	30.28	25.81	13.56	30.34	0.01	56.08	43.90	0.08	-0.38
Columba janthina	KM926619.1	17469	30.38	24.08	13.54	32.00	0.00	54.46	45.54	0.12	-0.41
Nibea albiflora	HQ890947.1	16499	26.40	25.88	16.91	30.81	0.00	52.28	47.72	0.01	-0.29
Mephitis mephiti	HM106332.1	16538	34.04	29.01	13.25	23.70	0.00	63.05	36.95	0.08	-0.28
Trichiosoma anthracinum	KT921411.1	15392	43.35	37.42	7.75	11.48	0.01	80.76	19.23	0.07	-0.19
Collichthys niveatus	JN678726.1	16450	27.76	25.85	15.91	30.48	0.00	53.60	46.40	0.04	-0.31
Pennahia argentata Japan	KC545800.1	16486	27.51	26.44	15.93	30.12	0.00	53.95	46.05	0.02	-0.31
Lynx rufus	NC_014456.1	17056	32.39	26.59	14.26	26.75	0.00	58.99	41.01	0.10	-0.30
Panthera tigris amoyensis	NC_014770.1	17001	31.86	26.94	14.62	26.57	0.00	58.80	41.20	0.08	-0.29
Prionailurus bengalensis euptilura	NC_016189.1	16990	33.03	27.41	13.54	26.02	0.00	60.44	39.56	0.09	-0.32
Ciconia ciconia	AB026818.1	17347	30.54	23.13	14.35	31.98	0.00	53.66	46.34	0.14	-0.38
Streptopelia chinensis	KP273832.1	16966	30.09	23.92	13.93	32.06	0.00	54.01	45.99	0.11	-0.39
Ichthyaetus relictus	KC760146.1	16586	30.62	24.38	14.07	30.93	0.00	55.00	45.00	0.11	-0.37
Ectopistes migratorius	KX902243.1	16943	30.08	24.44	13.98	31.45	0.04	54.52	45.43	0.10	-0.38
Pollachius pollachius	NC_015097.1	16539	27.68	29.23	17.15	25.94	0.00	56.91	43.09	-0.03	-0.20
Felis catus	U20753.1	17009	32.59	27.08	14.15	26.19	0.00	59.67	40.33	0.09	-0.30
Sclerophasma paresisensis	DQ241798.1	15500	41.59	33.47	10.57	14.36	0.00	75.06	24.94	0.11	-0.15
Chalcophaps indica	HM746789.1	15363	30.69	23.78	13.50	32.03	0.00	54.47	45.53	0.13	-0.41
Eurynorhynchus pygmeus	KP742478.1	16707	31.29	24.85	13.84	30.02	0.00	56.14	43.86	0.11	-0.37
Sternula albifrons	KT350612.1	16357	31.11	26.12	13.74	29.03	0.00	57.22	42.78	0.09	-0.36
Orussus occidentalis	FJ478174.1	15947	38.75	37.46	8.23	15.55	0.00	76.21	23.79	0.02	-0.31
Mastotermes darwiniensis	JX144929.1	15487	39.63	28.39	11.99	19.99	0.00	68.02	31.98	0.17	-0.25
Tenthredo tienmushana	KR703581.1	14942	42.48	37.67	7.71	12.15	0.00	80.14	19.86	0.06	-0.22
Ursus americanus	AF303109.1	16841	31.14	28.26	15.52	25.08	0.00	59.40	40.60	0.05	-0.24
Synthliboramphus antiquus	AP009042.1	16730	31.10	24.75	13.56	30.60	0.00	55.85	44.15	0.11	-0.39
Platalea minor	EF455490.1	16918	31.13	24.24	13.97	30.66	0.00	55.37	44.63	0.12	-0.37
Theragra finnmarchica	AM489718.1	16571	28.10	29.51	16.70	25.69	0.00	57.61	42.39	-0.02	-0.21
Blattella germanica	EU854321.1	15025	39.19	35.36	10.44	15.01	0.00	74.56	25.44	0.05	-0.18
Jacana jacana	KJ631049.1	16975	31.88	24.35	13.15	30.62	0.00	56.23	43.77	0.13	-0.40
Gallus gallus	NC_001323.1	16775	30.25	23.79	13.51	32.45	0.00	54.04	45.96	0.12	-0.41
Patagioenas fasciata	KX902239.1	16970	30.20	24.32	13.85	31.61	0.02	54.52	45.46	0.11	-0.39
Goura cristata	KX902242.1	17082	29.69	24.03	14.33	31.93	0.02	53.72	46.26	0.11	-0.38
Reticulitermes santonensis	EF206315.1	16567	43.06	23.03	11.99	21.92	0.01	66.09	33.90	0.30	-0.29
Collichthys lucida	JN857362.1	16451	28.00	25.63	15.71	30.65	0.00	53.63	46.37	0.04	-0.32
Schedorhinotermes breinli	JX144935.1	15864	43.89	22.09	11.66	22.37	0.00	65.97	34.03	0.33	-0.31
Ursus ursinus	EF196662.1	16817	30.83	27.47	15.76	25.95	0.00	58.29	41.71	0.06	-0.24
Pennahia argentata China	HQ890946.1	16485	27.46	26.33	16.04	30.18	0.00	53.78	46.22	0.02	-0.31
Ciconia boyciana	AB026193.1	17622	30.85	22.87	14.29	31.99	0.00	53.72	46.28	0.15	-0.38
Tamolanica tamolana	DQ241797.1	16055	39.84	35.43	9.45	15.28	0.00	75.27	24.73	0.06	-0.24
Vanellus vanellus	KM577158.1	16795	31.44	24.03	13.76	30.77	0.00	55.47	44.53	0.13	-0.38
Panthera tigris altaica	HM185182.2	16995	31.86	26.94	14.57	26.62	0.01	58.79	41.19	0.08	-0.29
Ursus thibetanus	EF196661.1	16795	31.16	27.88	15.44	25.51	0.01	59.05	40.95	0.06	-0.25
Gadus ogac	NC_012323.1	15564	27.70	29.14	16.96	26.01	0.19	56.84	42.96	-0.03	-0.21
Micromesistius poutassou	FR751401.1	16573	27.42	27.10	17.23	28.24	0.00	54.53	45.47	0.01	-0.24
Recurvirostra avosetta	KP757766.1	16897	31.72	23.59	13.56	31.13	0.00	55.31	44.69	0.15	-0.39
Scolopax rusticola	KM434134.1	16984	31.79	25.02	13.34	29.85	0.00	56.81	43.19	0.12	-0.38
Panthera leo persica	JQ904290.1	16818	31.93	27.17	14.48	26.42	0.00	59.10	40.90	0.08	-0.29
Columba jouyi	KX902247.1	17179	30.48	24.05	13.71	31.75	0.02	54.53	45.46	0.12	-0.40
Merluccius merluccius	NC_015120.1	17078	26.61	25.07	16.86	31.46	0.00	51.68	48.32	0.03	-0.30
Chroicocephalus ridibundus	KM577662.1	16807	30.81	23.97	14.16	31.06	0.00	54.78	45.22	0.12	-0.37
Microhodotermes viator	JX144931.1	15704	44.79	22.44	11.67	21.10	0.00	67.23	32.77	0.33	-0.29
Cephus pygmeus	KM377623.1	16145	42.93	36.89	7.23	12.95	0.00	79.82	20.18	0.08	-0.28
Pollachius virens	FR751399.1	16556	27.68	29.39	17.09	25.85	0.00	57.06	42.94	-0.03	-0.20
Vanhornia eucnemidarum	DQ302100.1	16574	43.49	36.62	6.67	13.18	0.04	80.11	19.85	0.09	-0.33
Eupolyphaga sinensis	FJ830540.1	15553	40.42	31.61	10.49	17.47	0.00	72.04	27.96	0.12	-0.25
Caloenas nicobarica	KX902248.1	17090	30.09	24.97	14.27	30.67	0.00	55.06	44.94	0.09	-0.36
Haematopus ater	AY074886.2	16791	31.59	23.62	13.67	31.12	0.00	55.21	44.79	0.14	-0.39
Neotermes insularis	JX144933.1	15799	42.63	25.20	11.71	20.46	0.00	67.83	32.17	0.26	-0.27
Zenaida macroura	KX902235.1	17132	29.84	23.57	14.17	32.41	0.00	53.41	46.59	0.12	-0.39

Continued on next page

Table 1 – *Continued from previous page*

Species name	Accession number	Sequence length	A%	T%	G%	C%	Unrecognized%	AT%	GC%	AT skew	GC skew
Hemiphaga novaeseelandiae	NC_013244.1	17264	30.97	23.92	13.21	31.89	0.01	54.89	45.11	0.13	-0.41
Monocellicampa pruni	JX566509.1	15169	40.87	36.34	8.31	14.48	0.00	77.22	22.78	0.06	-0.27
Panthera uncia	EF551004.1	16773	31.94	27.09	14.48	26.49	0.00	59.03	40.97	0.08	-0.29
Branta canadensis	NC_007011.1	16760	30.18	22.60	15.14	32.07	0.00	52.79	47.21	0.14	-0.36
Caloenas maculata	KX902249.1	17036	29.21	25.31	14.80	30.67	0.01	54.53	45.47	0.07	-0.35
Leptotila verreauxi	NC_015190.1	17176	30.08	24.03	13.88	32.02	0.00	54.10	45.90	0.11	-0.40
Zenaida auriculata	HM640211.1	16781	29.71	23.58	14.21	32.48	0.02	53.29	46.69	0.12	-0.39
Trachyrincus murrayi	AP008990.1	16677	29.14	29.94	15.82	25.11	0.00	59.08	40.92	-0.01	-0.23
Arenaria interpres	AY074885.2	16725	30.64	24.71	13.94	30.72	0.00	55.34	44.66	0.11	-0.38
Asiemphytus rufocephalus	KR703582.1	14864	43.02	38.38	7.55	11.05	0.00	81.40	18.60	0.06	-0.19
Chroicocephalus brunnicephalus	JX155863.1	16769	30.70	24.03	14.16	31.11	0.00	54.73	45.27	0.12	-0.37
Ursus arctos	AF303110.1	17020	30.89	27.80	15.72	25.59	0.00	58.69	41.31	0.05	-0.24
Phodilus badius	KF961183.1	17086	30.41	21.36	14.28	33.66	0.29	51.77	47.94	0.17	-0.40
Coelorinchus kishinouyei	AP002929.1	15942	29.57	28.11	15.52	26.80	0.00	57.68	42.32	0.03	-0.27
Otidiphaps nobilis	KX902241.1	16570	29.66	24.68	13.65	30.46	1.56	54.34	44.10	0.09	-0.38
Columba livia	KJ722068.1	17235	30.22	24.05	13.97	31.76	0.00	54.27	45.73	0.11	-0.39
Larimichthys polyactis	FJ618559.1	16470	27.55	25.00	16.18	31.27	0.00	52.55	47.45	0.05	-0.32
Pezophaps solitaria	KX902238.1	16644	29.83	23.82	13.60	31.00	1.75	53.65	44.60	0.11	-0.39
Gadus morhua kildinensis	AM489716.1	16654	28.06	29.56	16.79	25.59	0.00	57.62	42.38	-0.03	-0.21
Panthera tigris tigris	JF357968.1	16976	31.84	26.97	14.60	26.58	0.01	58.81	41.19	0.08	-0.29
Panthera pardus	NC_010641.1	16964	31.81	27.07	14.54	26.57	0.00	58.88	41.12	0.08	-0.29
Panthera tigris corbetti	JF357971.1	16602	31.85	26.88	14.64	26.62	0.01	58.73	41.26	0.08	-0.29
Larus dominicanus	AY293619.1	16701	30.54	24.45	14.13	30.88	0.00	54.98	45.02	0.11	-0.37
Jacana spinosa	KJ631048.1	17079	31.54	24.86	13.09	30.42	0.09	56.40	43.51	0.12	-0.40
Gallicolumba luzonica	HM746790.1	15192	31.18	24.01	13.67	31.13	0.01	55.19	44.81	0.13	-0.39
Threskiornis aethiopicus	GQ358927.1	16960	31.08	23.97	13.96	30.98	0.00	55.06	44.94	0.13	-0.38
Columba rupestris	KX902246.1	17201	30.19	23.86	13.77	31.93	0.24	54.05	45.71	0.12	-0.40
Miichthys miiuy	HM447240.1	16493	27.49	24.43	15.89	32.19	0.00	51.92	48.08	0.06	-0.34
Nycticorax nycticorax	NC_015807.1	17829	32.37	23.53	14.13	29.96	0.00	55.90	44.10	0.16	-0.36
Egretta eulophotes	KJ190949.1	20058	31.43	23.72	13.55	31.30	0.00	55.15	44.85	0.14	-0.40
Heterotermes sp	JX144936.1	16370	42.79	22.10	12.27	22.83	0.00	64.89	35.11	0.32	-0.30
Vanellus cinereus	KM404175.1	17074	31.63	23.53	13.77	31.08	0.00	55.15	44.85	0.15	-0.39
Chroicocephalus saundersi	JQ071443.1	16725	30.41	23.98	14.39	31.21	0.01	54.39	45.60	0.12	-0.37
Eumetopias jubatus	AJ428578.2	16638	33.58	25.63	13.69	27.11	0.00	59.21	40.79	0.13	-0.33
Theragra chalcogramma pantophysin	AB182305.1	16571	28.10	29.50	16.69	25.71	0.00	57.61	42.39	-0.02	-0.21
Cetonurus globiceps	KF751382.1	17137	28.07	28.13	15.57	28.23	0.00	56.21	43.79	0.00	-0.29
Perga condei	AY787816.1	13416	42.75	35.15	8.25	13.82	0.02	77.91	22.07	0.10	-0.25
Egretta garzetta	NC_023981.1	17361	31.50	23.23	13.47	31.80	0.00	54.73	45.27	0.15	-0.40
Larimichthys crocea	EU339149.1	16466	27.55	25.46	16.30	30.69	0.00	53.01	46.99	0.04	-0.31
Ursus maritimus	AF303111.1	17017	30.87	27.77	15.82	25.54	0.00	58.64	41.36	0.05	-0.24
Bahaba taipingensis	JX232404.1	16500	27.55	25.15	15.90	31.41	0.00	52.70	47.30	0.05	-0.33
Locusta migratoria	X80245.1	15722	44.54	30.79	10.09	14.58	0.00	75.33	24.67	0.18	-0.18
Drepanotermes sp	JX144938.1	16542	42.45	24.76	12.02	20.77	0.00	67.20	32.80	0.26	-0.27
Cephus cinctus	FJ478173.1	19339	42.39	39.57	6.44	11.59	0.01	81.95	18.04	0.03	-0.29
Macrotermes subhyalinus	JX144937.1	16351	43.91	21.66	11.55	22.89	0.00	65.56	34.44	0.34	-0.33
Stercorarius maccormicki	KM401546.1	16669	30.94	24.39	13.82	30.85	0.00	55.34	44.66	0.12	-0.38
Canis lupus familiaris	NC_002008.4	16727	31.63	28.72	14.14	25.51	0.00	60.35	39.65	0.05	-0.29
Nibea coibor	KM373207.1	16509	26.91	25.40	16.32	31.37	0.00	52.30	47.70	0.03	-0.32
Manis tetradactyla	NC_004027.1	16571	33.31	29.75	13.73	23.21	0.00	63.06	36.94	0.06	-0.26
Bregmaceros nectabanus	AP004411.1	16030	28.62	31.15	14.93	25.28	0.01	59.78	40.22	-0.04	-0.26
Porotermes adamsoni	JX144930.1	16039	42.76	24.05	11.90	21.29	0.00	66.82	33.18	0.28	-0.28
Gallinago stenura	KY056596.1	16899	32.21	26.14	12.88	28.77	0.00	58.35	41.65	0.10	-0.38
Parapristipoma trilineatum	NC_009857.1	16546	27.99	26.89	16.57	28.54	0.00	54.88	45.12	0.02	-0.27
Zootermopsis angusticollis	JX144932.1	15483	46.08	23.32	10.70	19.91	0.00	69.40	30.60	0.33	-0.30
Dendrophysa russelii	JQ728562.1	16626	27.28	26.12	16.20	30.40	0.00	53.40	46.60	0.02	-0.30
Ascaloptynx appendiculatus	FJ171324.1	15877	40.34	35.23	9.70	14.73	0.00	75.57	24.43	0.07	-0.21
Macrognathotermes errator	JX144939.1	16330	42.33	24.49	11.88	21.30	0.00	66.82	33.18	0.27	-0.28
Turtur tympanistria	HM746793.1	15557	30.44	23.87	13.70	31.99	0.00	54.32	45.68	0.12	-0.40
Nipponia nippon	AB104902.1	16732	30.44	23.48	14.27	31.81	0.00	53.92	46.08	0.13	-0.38
Canis familiaris	U96639.2	16727	31.63	28.72	14.14	25.51	0.00	60.35	39.65	0.05	-0.29
Vespa bicolor	KJ735511.1	16937	40.74	40.98	5.47	12.81	0.00	81.72	18.28	0.00	-0.40
Pterocles gutturalis	KX902237.1	15637	29.31	25.19	13.19	27.18	5.14	54.50	40.37	0.08	-0.35
Squalogadus modificatus	AP008989.1	16550	29.35	29.12	15.35	26.19	0.00	58.47	41.53	0.00	-0.26
Diadegma semiclausum	EU871947.1	18728	44.08	43.33	5.05	7.54	0.00	87.41	12.59	0.01	-0.20
Phoca vitulina	X63726.1	16826	32.98	25.30	14.28	27.43	0.00	58.28	41.72	0.13	-0.32
Ventrifossa garmani	AP008991.1	17230	28.17	27.80	15.85	28.18	0.00	55.97	44.03	0.01	-0.28
Geopelia striata	HM746791.1	15859	30.59	23.63	13.89	31.88	0.01	54.22	45.77	0.13	-0.39
Coptotermes lacteus	JX144934.1	16326	42.94	21.46	12.07	23.53	0.00	64.40	35.60	0.33	-0.32
Cryptocercus relictus	JX144941.1	15373	45.34	28.15	10.13	16.32	0.05	73.49	26.46	0.23	-0.23
Periplaneta fuliginosa	AB126004.1	14996	42.13	33.02	10.35	14.50	0.00	75.15	24.85	0.12	-0.17
Puma concolor	NC_016470.1	17153	32.90	27.31	13.81	25.98	0.00	60.21	39.79	0.09	-0.31
Sardinops melanostictus	NC_002616.1	16881	25.32	25.98	20.34	28.35	0.00	51.31	48.69	-0.01	-0.16
Tremarctos ornatus	EF196665.1	16766	31.29	27.34	15.42	25.93	0.02	58.62	41.35	0.07	-0.25
Larus crassirostris	KM507782.1	16746	30.62	24.34	14.13	30.91	0.00	54.96	45.04	0.11	-0.37

215

Data acquisition

We have developed a python based web scraper which can download the whole genome sequences, gene, and amino acid sequences (from the NCBI database server) by specifying either the accession numbers of the target species, or the name of a particular gene. In the former case, the whole genome sequences, different gene and amino acid sequences of the query species are downloaded. In the second case, all of the homologs of the query gene are extracted from the server. The web scraper extracts information from the NCBI data repository using *Entrez Global Query Cross-Database Search System*[2]. Entrez is a primary text search and retrieval system of NCBI database. The search system provides nine e-utilities, out of which "ESearch", "ELink", and "EFetch" have been used in our tool. From the downloaded items we consider only the mitochondrial genome sequences used in the subsequent analysis. Accession numbers of the individual species used in the current study are listed in Table 1.

Phylogenetic inference

We apply the proposed distance measure (refer to Eq. (4)) over 157 selected species from four classes to compute pairwise distances between them with different values of length, L (from 50 to 5000), $\mathcal{F}$ (from 1 to 200), and α [0,1]. We compute all the feature sets separately for GFP-RY, GFP-SW, and GFP-MK (refer to Eq. (1), (2), (3), respectively). By applying Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [64] over these distance matrices, we get the phylogenetic tree for each such case. The inferred phylogenetic trees for GFP-RY, GFP-SW, and GFP-MK are represented as $\mathcal{T}_{RY}$, $\mathcal{T}_{SW}$, and $\mathcal{T}_{MK}$, respectively.

Finally, given L, $\mathcal{F}$, and α, we get three phylogenetic trees for GFP-RY, GFP-SW, and GFP-MK, $\mathcal{T}_{RY}$, $\mathcal{T}_{SW}$, and $\mathcal{T}_{MK}$, respectively. These trees are combined following a consensus tree merging algorithm (COSPEDTree-II [4]) and get $\mathcal{T}$.

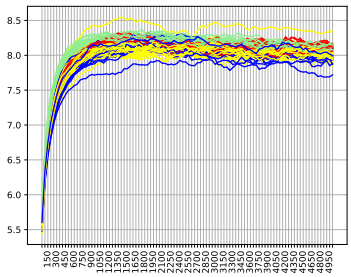
Selecting parameter values

We apply following three step technique to select the L, $\mathcal{F}$, and α.

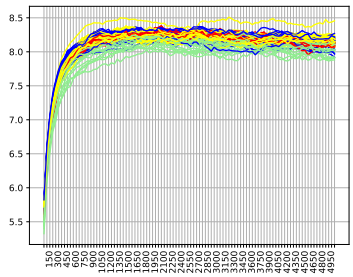
1. Selecting the value of L using Shannon entropy of the sequence of each species.
2. Considering the intraclass variances and interclass distances of the features of each species to select the value of $\mathcal{F}$.
3. By considering the same for the pairwise distances we select the value of α.

It is empirically noticed that the selection criteria we proposed derive the trees $\mathcal{T}_{RY}$, $\mathcal{T}_{SW}$, and $\mathcal{T}_{MK}$ infer better clades with the four different classes of species.

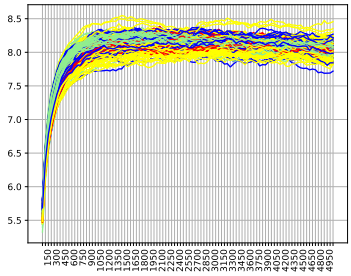
[2]Website of "NCBI Help Manual": `http://www.ncbi.nlm.nih.gov/books/NBK3831/`



(a) Shannon entropy for Δ_{RY}



(b) Shannon entropy for Δ_{SW}



(c) Shannon entropy for Δ_{MK}

Figure 3: Shannon entropy for all the value of L from 50 to 5000. Different color graphs represent the Shannon entropy of four different types of species, Mammalia (Red), Aves (Yellow), Actinopterygii (Blue), and Insecta (Green).

Selection of the value of L

Shannon entropy [60] is used to measure the randomness in genomic data [66]. For different values of L (from 50 to 5000 with the difference of 50), we compute the Shannon entropy of the drift sequence (Δ) of individual species. The high value of entropy infers that for the value of L, the $\Delta^{(L)}$ contains high number of unique point coordinates. Fig. 3 shows that initially the entropy of all the species increase by increasing L. At almost $L = 800$, the entropy of all the species become stabilized at a high level. By increasing the value of L after that does not change the entropy at any significant level. From Fig. 3 it can be noted that for $L \geq 800$, the $\Delta^{(L)}$ contains significantly large number of unique point coordinates than that of $L < 800$. It is also be pointed out that increasing the value of L reduces the number of point coordinates in the corresponding Δ. Hence, we choose the value of L as 800.

Selection of the value of $\mathcal{F}$

Here we consider the feature vector for each species, say χ_s, where s is a species. By applying the distance metric (refer to Eq. (4)), we compute the distances, say, $D(\chi_{si}, \mu_i)$, where χ_{si} is the feature vector of the species s selected from class i, and μ_i is the mean of the feature vector of all species from class i. The variance of the computed distances of class i, say, σ_i^2, represents the separation of the feature vectors of intraclass species.

So, for C number of classes the mean of intraclass variances,

$$\mu_{intraclass} = \frac{1}{C}\sum_{i=1}^{C}\sigma_i^2 \tag{10}$$

To derive the interclass distances, we consider the μ, where,

$$\mu = \frac{1}{C}\sum_{s=1}^{C}\mu_i \tag{11}$$

By applying the distance metric (refer to Eq. (4)) we compute $D(\mu_i, \mu)$ which represents the separation of the feature vectors of interclass species. So, for C number of classes the mean of interclass distance,

$$\mu_{interclass} = \frac{1}{C}\sum_{i=1}^{C}D(\mu_i, \mu) \tag{12}$$

Using this two elements we derive the discriminant score of the selected species as,

$$DS = \frac{\mu_{interclass}}{\mu_{intraclass}} \tag{13}$$

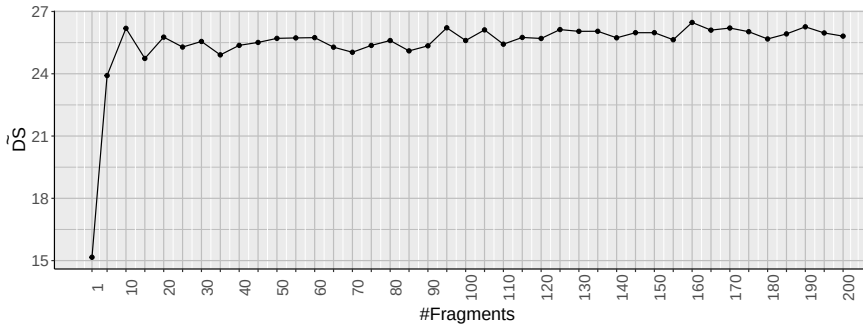
Maximizing the DS is equivalent to getting a good separation between the feature vectors of interclass species. We apply this method for different values of $\mathcal{F}$ (from 1 to 200) and the optimized values of L, here it is $L \geq 800$. It is found from Fig. 4 (a, c, and e) that for all the values of L, the value of DS increases with an increasing value of $\mathcal{F}$. It can be noticed that the overall value of DS is maximum for $L = 800$. So to select the value of $\mathcal{F}$, we consider $\widetilde{DS}$ for $L = 800$, where,

$$\widetilde{DS} = \frac{\mu_{interclass}}{\log \mu_{intraclass}} \tag{14}$$

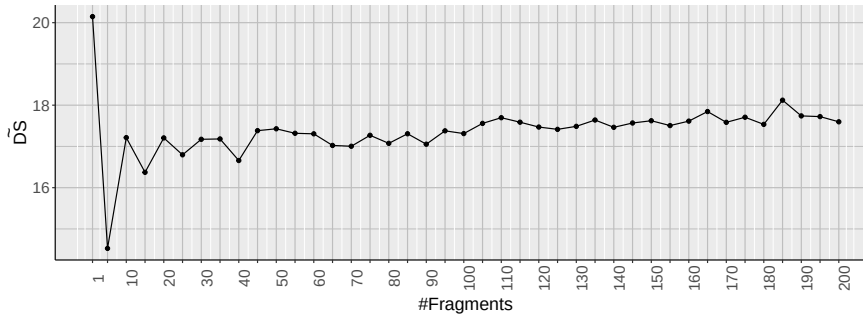
As the effect of $\mu_{intraclass}$ is scaled down in $\widetilde{DS}$, so $\widetilde{DS}$ represents the effect of $\mu_{interclass}$ on DS for the corresponding value of $\mathcal{F}$. It is also shown in Fig. 4 (b, d, and f) that, for $L = 800$, after a period the change of $\widetilde{DS}$ becomes less than 5%. We consider that as the stable state of $\widetilde{DS}$. This implies that after a certain value of $\mathcal{F}$, the interclass distance does not increase significantly with an increasing value of $\mathcal{F}$. Within this *stable state*, there are some segments where the changes of $\widetilde{DS}$ are less than 2%. We consider those as *stationary regions*. We consider the DS (obtained from Eq. (13)) of those stationary regions. We choose that value of $\mathcal{F}$ for which the maximum value of DS lies within these stationary regions. Empirically it is tested that for that value of $\mathcal{F}$, GRAFree infers tree with better clades for both GFP-RY, GFP-SW, and GFP-MK. Hence, it is considered that for the particular value of $\mathcal{F}$ the feature vector of the species represents the Δ better for comparative genomic study. Hence we select the value of $\mathcal{F}$ as 160, 165, and 165 for Δ_{RY}, Δ_{SW}, and Δ_{MK}, respectively.

Selection of the value of α

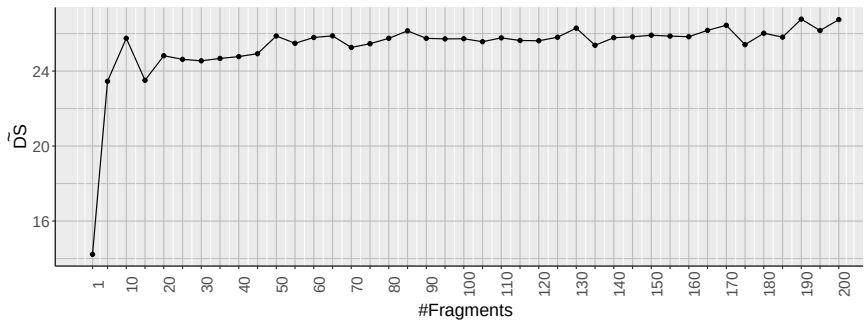
For a given L and $\mathcal{F}$, to choose the value of α within the range of [0,1], we consider the distance matrix and apply the same concept over the scaler values of the distance matrix to compute DS (refer to Eq. (13)). We derive the mean and variance of all pairwise distances, say, μ_i and σ_i^2 respectively,



(a) For $L = 800$ the $\widetilde{DS}$ for Δ_{RY}

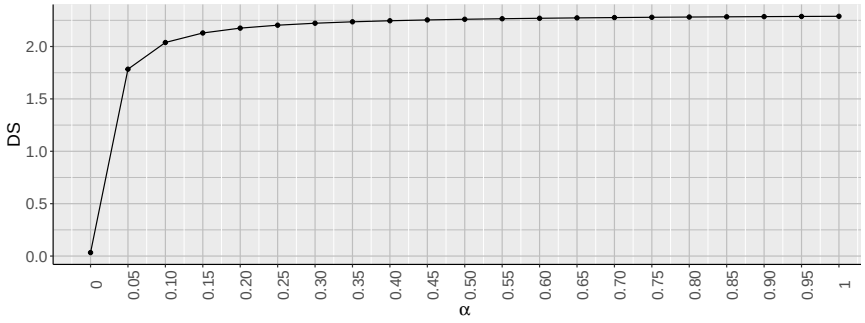


(b) For $L = 800$ the $\widetilde{DS}$ for Δ_{SW}

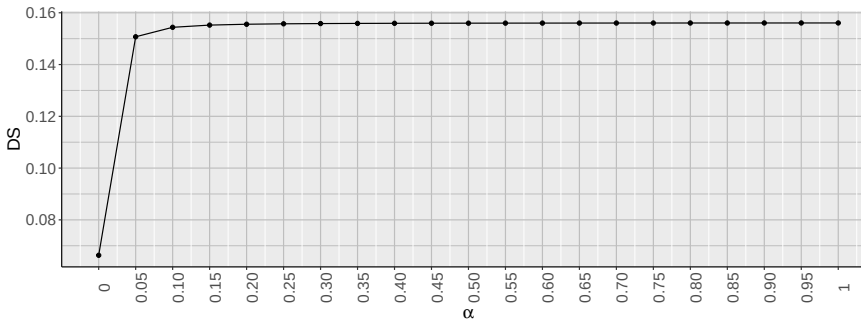


(c) For $L = 800$ the $\widetilde{DS}$ for Δ_{MK}

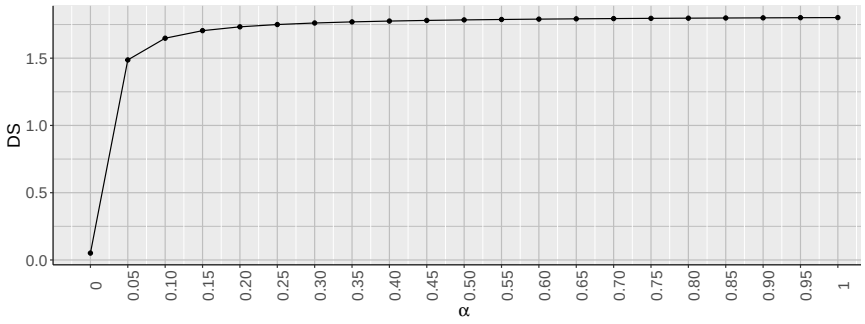
Figure 4: The *widetildeDS* for $L = 800$ for Δ_{RY}, Δ_{SW}, and Δ_{MK}, respectively.



(a) DS for GFP-RY



(b) DS for GFP-SW



(c) DS for GFP-MK

Figure 5: The DS for all the values of α from 0 to 1, given $L = 800$ and $\mathcal{F}$ =160, 165, and 165 for GFP-RY, GFP-SW, and GFP-MK, respectively.

between the species of class i. Similarly, compute the $\mu_{intraclass}$ as the mean of intraclass variances using the Eq. (10).

The mean of interclass distance is derived as following,

$$\mu_{interclass} = \frac{1}{C}\sum_{i=1}^{C}(\mu_i - \mu)^2 \tag{15}$$

It is also be observed that the value of DS becomes stabilized after a period. Now we choose that value of α for which the maximum value of DS is obtained from Eq. (13) after stabilized. From Fig. 5, it can be observed that for all the selected values of $\mathcal{F}$ (=160, 165, and 165 for Δ_{RY}, Δ_{SW}, and Δ_{MK}, respectively) and L (=800), the maximum value of DS is obtained for α =1.

Hence, we consider $L = 800$, $\mathcal{F} = 160$, and $\alpha = 1$ as the value of the parameters to derive the phylogenetic tree for GFP-RY. Similarly, for GFP-SW and GFP-MK, the value of L, $\mathcal{F}$, and α are chosen as 800, 165, 1 and 800, 165, 1, respectively. It is also noted empirically that the tree inferred using these parameters accumulates most of the intraclass species within same clade.

Performance measure

In this study, we consider four different classes with more than 50 families of species. For measuring the accuracy of the derived tree we consider four classes, seven orders, and four families which are monophyletic and have more than ten representative species in our dataset. Our primary objective of this proposed method of performance measuring is to cluster the monophyletic species according to their respective class, order, or family. This is a quantitative measure of the deformation of a given monophyletic clade of phylogenetic tree. This measurement is useful especially when we do not have the reference tree to compare. The *transfer level* ($TransLv$) proposed here is defined as the minimum number of levels require to move a species to another target clade. The objective behind the transfer of a species to another clade is either to place the species to its appropriate clade or to remove the species from an inaccurate clade. The *total transfer level* (TTL) of a clade is the sum of the transfer levels to make a clade correct. The proposed measure of accuracy of a clade is based on the total transfer level of the clade. Using the TTL we compute the proposed measure, *Deformity Index*, which is a quantitative measure of the deformation the clades of the tree. Hence, for the ideal case where each species is placed within the proper clade, the value of Deformity Index is zero. The computation of the Deformity Index of a clade is described in Algorithm 1.

Results and discussion

Bootstrapping

The conventional method of bootstrapping [17], [14] considers the aligned sequences to resample and replicate. As we are developing an alignment-free method of phylogeny construction, the conventional bootstrapping method may not be applicable for this case. The main motivation of bootstrapping is to generate the population from a single genome. It is observed that the average intraspecific genetic variation is within 1% [38], [70]. So here we propose a bootstrapping technique which considers the genetic variance of a sample space within 1%. For that we apply mutations at each location with a probability of 1% and consider an unbiased selection of the nucleotides at each location. We generate 100 replicas using this bootstrapping method and construct trees from each set of sequences using GRAFree method by setting values of L, $\mathcal{F}$, and α as discussed in the previous section. Felsenstein's bootstrapping method [17] assesses the robustness of phylogenetic trees using the presence and absence of clades. For the large scale genomics Felsenstein's bootstrap is not efficient to sum up the replicas. For the hundreds of species this method is inclined to produce low bootstrap support [36]. So here we apply a modification of Felsenstein's bootstrapping, where the presence of a clade is quantified using the transfer distance proposed in [36]. The transfer distance [10] or

Algorithm 1 Algorithm for measuring the Deformity Index

Input: Tree topology

Input: $MonoPhyl$, is a list of all species from a monophyletic clade considered to compute the Deformity Index.

Output: Deformity Index of the clade of the tree

//Compute the deformity index to place species to its appropriate clade.

Find the all unique clades having maximum number of species from $MonoPhyl$, say C

if $|c| = 1$, $\forall c \in C$, that means every $c \in C$ contains single species s, where $s \in MonoPhyl$ **then**

  $DeformityIndexAdd = \infty$

else

  for each c, $\forall c \in C$ **do**

    Derive $TransLv_s$ for each species s, where $s \notin c$

    $TTL_c = \sum_s TransLv_s$

  end for

end if

$DeformityIndexAdd = \frac{min(TTL_c)}{|MonoPhyl|}$,
where, $|MonoPhyl|$ = number of species in $MonoPhyl$

//Compute the deformity index to remove the species from an inaccurate clade

Find all species, S, placed under the clade of $MonoPhyl$, where $S \notin MonoPhyl$

Derive $TransLv_s$ for each species s, where $s \in S$

$TTL = \sum_s TransLv_s$

$DeformityIndexRemove = \frac{min(TTL)}{|MonoPhyl|}$,
where, $|MonoPhyl|$ = number of species in $MonoPhyl$

$DeformityIndex =$
$min(DeformityIndexAdd_i, DeformityIndexRemove_i)$

R-distance [13] is the minimum number of changes required to transform one partition to other. We computed the occurrence of each clade using the tool BOOSTER [36].

Observations from derived phylogenetic trees

Here, we present phylogenetic trees generated by our proposed method, GRAFree, using the whole mitochondrial genome sequences of the selected species. We consider the value of L, $\mathcal{F}$, and α derived from the selection technique proposed in the previous section. It is observed in Table 2 that the average Deformity Index of GFP-RY (please refer to Eq. (1)) is lower than that of GFP-SW and GFP-MK (please refer to Eq. (2) and (3), respectively). These results infer that the skew (AG skew and CT skew) represented by the Eq. (1) bears the signature of genomic contents related to the evolution more precisely than that of the other skews. Hence, the skew of the genomic signature may also require careful investigation on the matter of evolutionary relationships. The $\mathcal{T}_{RY}$, $\mathcal{T}_{SW}$, and $\mathcal{T}_{MK}$ are shown in Fig. 6. Fig. 7 presents the final tree after merging all three cases using the COSPED-II algorithm.

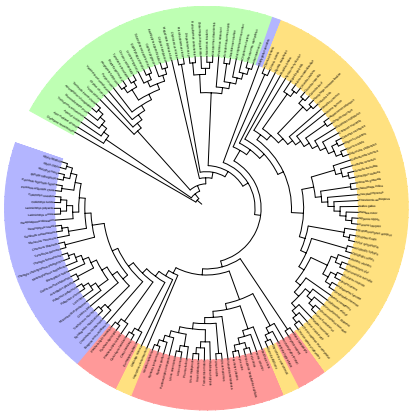
To measure the performance of $\mathcal{T}$, we chose 15 monophyletic clades of four classes, seven orders, and four families. It is observed that $\mathcal{T}$ has formed the monophyletic clades for the three major classes, mammals, fishes, and birds with minor deviations, whether insects are inferred as paraphylic. This tree also infers insects as the oldest class among these four classes followed by birds, mammals, and fishes. Mammals and fishes are inferred as the sister clades in $\mathcal{T}$. The deformity index of different trees are shown in Table 2.

Observations from reference methods

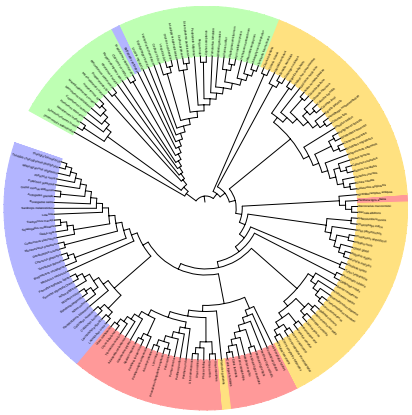
We have examined five different existing distance measures of alignment free method for phylogenetic reconstruction, i.e. FFP [62], [63], D_2^* [59], [68], Chebyshev, Canberra, and Co-phylog [72] using the tool ACcelerated Alignment-FrEe sequence analysis (CAFE) [44] on our dataset. We measured each distances with the word (k-mer) length of 10. The resultant trees are shown in Fig. 8. It is found that D_2^* and Co-phylog separate four major clades accurately (with the average deformity index of 0), Canberra separate four major clades with minor errors (with the average deformity index of 0.11) in their inferred trees whereas the other methods, FFP and Chebyshev, cannot identify the four classes as clades. The bootstrap support of the clades of the inferred tree from D_2^*, Co-phylog, and Canberra methods are also very high. The D_2^* and Co-phylog infer the tree having insects as the oldest clade followed by fishes, birds, and mammals. Mammals and birds are inferred as the sister clades in the derived trees from both D_2^* and Co-phylog methods. So the it is accepted that the D_2^*, Co-phylog, and Canberra infer better clades than that of FFP and Chebyshev.

Complexity analysis

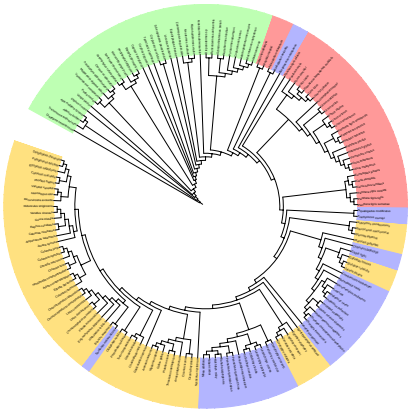
To compute the complexity of GRAFree, we consider M as the length of the genome sequences of two species, $\mathcal{S}_1$ and $\mathcal{S}_2$, the length of the window to compute drift is L, and the number fragments of the drift is $\mathcal{F}$. The GRAFree consists of three major steps.



(a) $\mathcal{T}_{RY}$, for $L = 800$, $\mathcal{F} = 160$, and $\alpha = 1$



(b) $\mathcal{T}_{SW}$, for $L = 800$, $\mathcal{F} = 165$, and $\alpha = 1$



(c) $\mathcal{T}_{MK}$, for $L = 800$, $\mathcal{F} = 165$, and $\alpha = 1$

Figure 6: The inferred trees with the selected value of L, $\mathcal{F}$, and α for GFP-RY, GFP-SW, and GFP-MK. The mammals, fishes birds, and insects are represented by red, blue, yellow, and green, respectively.

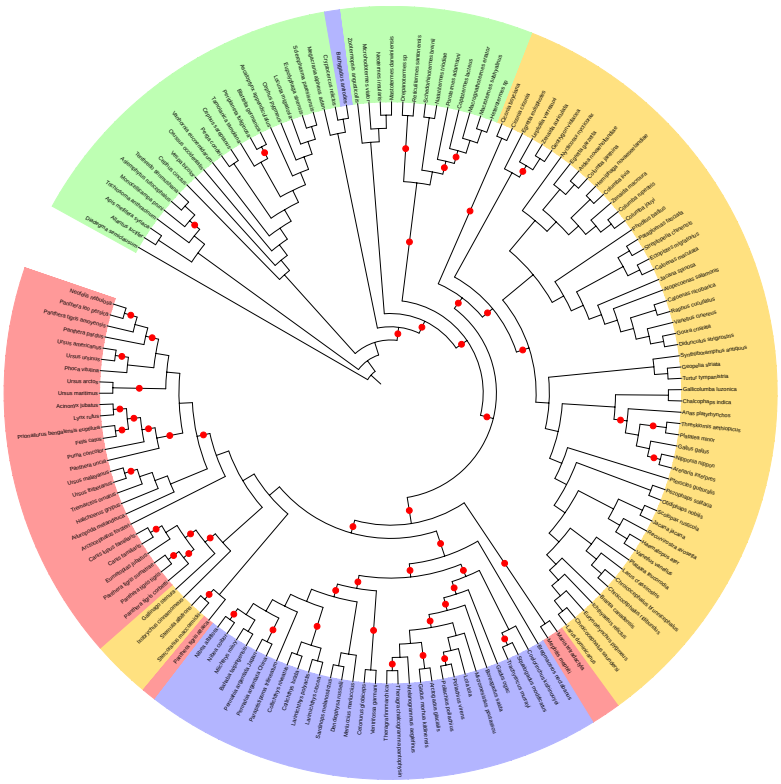


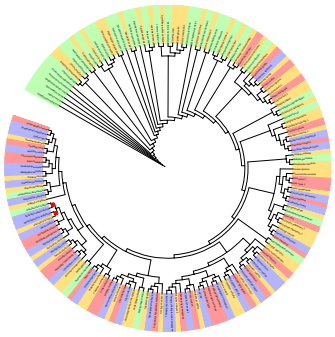
Figure 7: Inferred tree, $\mathcal{T}$ after merging $\mathcal{T}_{RY}$, $\mathcal{T}_{SW}$, and $\mathcal{T}_{MK}$. Red colored group represents the class mammals, blue colored group represents the class ray-finned fishes, yellow colored group represents the class birds, and green color represents the species from insect. The red dot on the branch represents the bootstrap score of that clade is greater than 75%.

Table 2: Deformity index for different trees

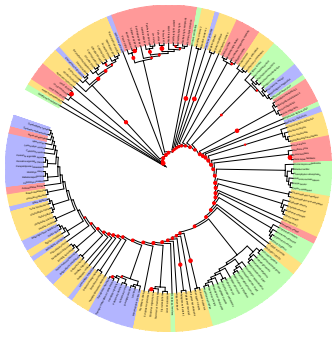
Method	**Class**			
	Mammalia	**Avian**	**Fish**	**Insecta**
GFP-RY	1.87	0.39	0.25	3.92
GFP-SW	2.30	4.54	0.56	0.22
GFP-MK	1.57	4.54	6.50	7.75
COSPED	0.50	4.56	0.41	2.61
Reference Methods				
FFP	11.75	12.86	12.59	102.25
Chebyshev	8.88	24.76	14.00	7.94
Canberra	0.00	0.00	0.31	0.00
D_2^*	0.00	0.00	0.00	0.00
Co-phylog	0.00	0.00	0.00	0.00

Method	**Order**						
	Carnivora (Order of Mammal)	**Charadriiformes (Order of Bird)**	**Columbiformes (Order of Bird)**	**Gadiformes (Order of Fish)**	**Perciformes (Order of Fish)**	**Hymenoptera (Order of Insect)**	**Blattodea (Order of Insect)**
GFP-RY	1.79	6.89	5.96	1.16	0.25	3.43	3.63
GFP-SW	2.10	5.58	5.96	4.16	0.42	1.86	4.00
GFP-MK	1.41	9.42	5.67	2.89	0.17	4.29	3.00
COSPED	0.38	5.47	7.04	1.68	0.33	2.93	3.19
Reference Methods							
FFP	11.63	132.00	10.14	12.11	74.67	295.64	201.38
Chebyshev	7.47	13.80	16.43	12.16	3.83	8.79	5.63
Canberra	0.13	1.75	0.00	0.47	0.00	0.00	3.13
D_2^*	0.20	0.00	0.00	0.11	0.00	6.21	2.50
Co-phylog	0.07	1.75	0.00	0.11	0.00	0.00	0.31

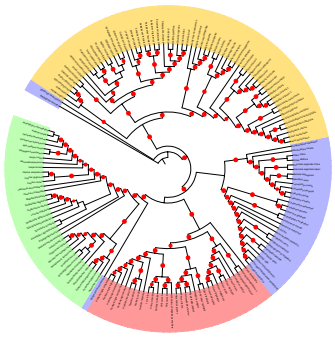
Method	**Family**			
	Felidae (Family of Carnivora)	**Threskiornithidae (Family of Pelecaniformes)**	**Gadidae (Family of Gadiformes)**	**Sciaenidae (Family of Perciformes)**
GFP-RY	4.21	43.50	0.40	0.73
GFP-SW	3.43	3.75	1.00	2.09
GFP-MK	3.29	3.75	1.70	0.45
COSPED	3.43	3.50	0.40	1.91
Reference Methods				
FFP	10.14	276.50	11.00	82.45
Chebyshev	3.14	649.00	4.30	4.09
Canberra	0.29	0.00	0.00	0.00
D_2^*	0.86	0.00	0.00	0.00
Co-phylog	0.29	0.00	0.00	0.00



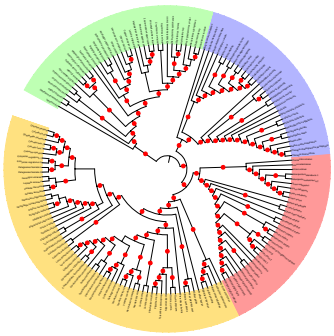
(a) Derived Tree from FFP



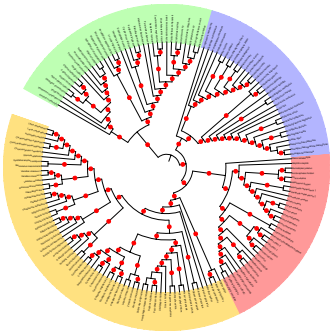
(b) Derived Tree from Chebyshev



(c) Derived Tree from Canberra



(d) Derived Tree from D_2^*



(e) Derived Tree from Co-phhylog

Figure 8: Derived tree from the reference methods. The mammals, fishes birds, and insects are represented by red, blue, yellow, and green, respectively. The clades having the bootstrap scores more than 75% are denoted by the red dots on the branch.

Computing drift

As drift is computed considering two point coordinates on the GFP, so for each species the time complexity to compute drift is $\mathcal{O}(M-L+1)$. The drift sequence contains the 2D coordinate of total $M-L+1$ points. So the space complexity of drift sequence of a species should be $\mathcal{O}(M-L+1)$.

Computing feature vector

Each fragment of the drift is represented by μ, Λ, λ, and θ (please refer to Subsection 1). The time complexity of Λ and λ are depending on the covariance matrix of drift. Since, we consider the 2D coordinate points in drift, the time complexity of computing Λ and λ for each species is $\mathcal{O}(M-L+1)$. Time complexity of computing μ and θ are linearly related to the length of drift sequence. Hence, for each species the time complexity of computing the feature vector for $\mathcal{F}$ number of fragments is $\mathcal{O}((M-L+1)\mathcal{F})$. Similarly, the space complexity of feature vector for each species is $\mathcal{O}(\mathcal{F})$.

Computing distance between a pair of species

GRAFree considers the distance function which computes the distances for all $\mathcal{F}$ number of fragments in a constant time, means $\mathcal{O}(1)$. So the time complexity of computing distance between a pair of species is $\mathcal{O}(\mathcal{F})$.

Hence, both time and space complexity of GRAFree is $\mathcal{O}(M-L+1)\mathcal{F}$.

Complexity analysis of FFP

FFP considers the frequencies of all k-mers. For the M length of sequence, the time complexity to compute the feature frequency of all k-mers is $\mathcal{O}(k(M-k+1))$. FFP uses Jensen-Shannon Divergence (JSD) for computing the distance between two feature frequency profiles. As JSD consider the entropy for deriving the distance, hence for $(M-k+1)$ length of sequences the time complexity for computing the JSD is $\mathcal{O}((M-k+1)^2)$. So the total time complexity of FFP for computing distance between two sequences is $\mathcal{O}(M(M-k+1))$. Since, FFP considers all k-mers to compute the feature frequency profile of the sequence, the total space complexity for nucleotide is not more than 4^k. Hence, the space complexity of FFP for two sequences is $\mathcal{O}(4^k)$.

Complexity analysis of Chebyshev

Chebyshev distance function considers the number of occurrences k-mers in the sequences. So for M length of sequences, the time complexity of computing the occurrence of a particular k-mers is $\mathcal{O}(k(M-k+1))$. The maximum among the absolute value of the difference between each k-mers of two sequences is considered as the Chebyshev distance. So the time complexity of Chebyshev for computing the distance between two sequences is $\mathcal{O}(4^k)$. Since, Chebyshev considers the occurrence of all k-mers, the space complexity of Chebyshev for two sequences is $\mathcal{O}(4^k)$.

Complexity analysis of Canberra

Similar to the Chebyshev, Canberra distance function also considers the number of occurrences of k-mers in the sequences. Hence, the time complexity to compute the occurrence k-mers is $\mathcal{O}(k(M-k+1))$. Canberra

Table 3: Time and space complexity to compute distance between two sequences by different methods

Methods	**Time Complexity**		**Space Complexity**
	Deriving features	**Computing distance**	
GRAFree	$\mathcal{O}(M-L+1)$	$\mathcal{O}((M-L+1)\mathcal{F})$	$\mathcal{O}(\mathcal{F})$
FFP	$\mathcal{O}(k(M-k+1))$	$\mathcal{O}((M-k+1)^2)$	$\mathcal{O}(4^k)$
Chebyshev	$\mathcal{O}(k(M-k+1))$	$\mathcal{O}(4^k)$	$\mathcal{O}(4^k)$
Canberra	$\mathcal{O}(k(M-k+1))$	$\mathcal{O}(4^k)$	$\mathcal{O}(4^k)$
D_2^*	$\mathcal{O}(k^2(M-k+1))$	$\mathcal{O}(4^k)$	$\mathcal{O}(4^k)$
Co-phylog [72]	$\mathcal{O}(M)$	$\mathcal{O}(M)$	$\mathcal{O}(kM)$

distance is considered as the summation of the ratio between the absolute value of the difference between each k-mer and the total occurrence of that particular k-mer within two sequences. Hence, the time complexity for computing the Canberra distance between two sequences is $\mathcal{O}(4^k)$. Similarly the space complexity of Canberra for two sequences is $\mathcal{O}(4^k)$.

Complexity analysis of D_2^*

Similar to the FFP, D_2^* considers occurrence of k-mers in the sequences. So for M length of sequences, the time complexity of computing the occurrence of a particular k-mers is $\mathcal{O}(k(M-k+1))$. D_2^* takes the probability of the k-mers within the combined sequence of the two sequences. For combined sequence of $2M$ length, the time complexity of computing the probability of each k-mer is $\mathcal{O}(k)$. Using these two values, D_2^* computes the distance for the particular k-mer. Finally, the sum of the distances for all k-mers is considered as the distance between two sequences. Hence, the time complexity of D_2^* for computing the distance between two sequences is $\mathcal{O}(kM4^k)$. As it stores the occurrence of all k-mers, the space complexity is $\mathcal{O}(4^k)$.

For the large scale genomic study the time and space complexities are one of the important things to be remembered. We compare the time and space complexity of GRAFree with some of the existing methods in Table 3. It can be observed that GRAFree is significantly efficient than all the selected existing methods in order of the time and space complexity. It is noted that D_2^* and Co-phylog are efficient in quality of reconstruction of tree. The execution time of different methods are shown in Fig. 9.

Conclusion

We have proposed a $5\mathcal{F}$-dimensional feature space and a new metric for capturing evolutionary relationship using large scale genomic features in the method GRAFree. GRAFree uses the graphical representation of the genome. In this study we have selected three very naive graphical representations of a genome considering residues independently. We have also proposed a novel measure to evaluate the performance of the techniques. The resultant tree accumulates most of the monophyletic clades with minor deviations. In spite of these limitations, we could observe presence of evolutionary traits in the proposed feature descriptor extracted from the whole mitochondrial sequences. The tree has a high bootstrap support for a good number of clades. These demonstrate the effectiveness of the

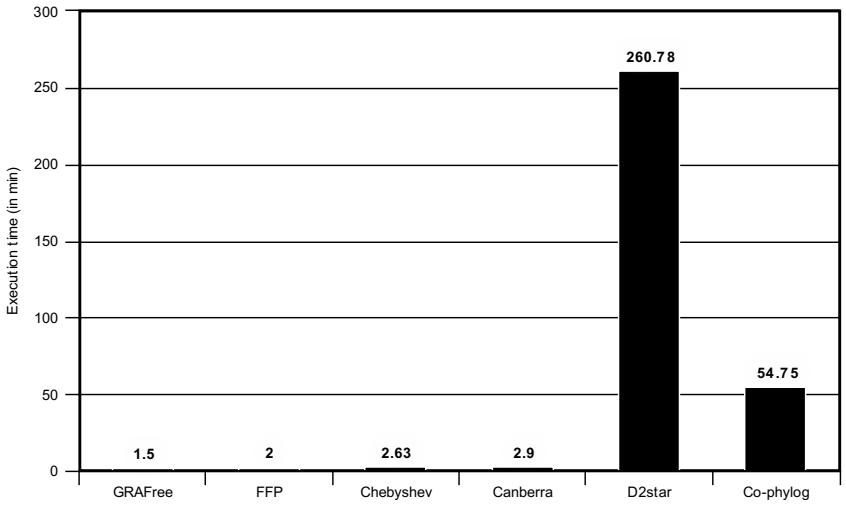


Figure 9: The execution time for different methods. All the methods are executed in the same system. The configuration of the system is 16GB RAM, Intel Core i5 processor, and it had 64 bit Ubuntu 17.10

proposed feature representation, as well as the metric for measuring the pairwise distances of species.

References

[1] J. S. Almeida. Sequence analysis by iterated maps, a review. *Briefings in Bioinformatics*, 15(3):369–375, 2013.

[2] M. Bernt, A. Braband, B. Schierwater, and P. F. Stadler. Genetic aspects of mitochondrial genome evolution. *Molecular Phylogenetics and Evolution*, 69(2):328–338, 2013.

[3] S. Bhattacharyya and J. Mukherjee. COSPEDTree: COuplet Supertree by Equivalence Partitioning of Taxa Set and DAG Formation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(3):590–603, May 2015.

[4] S. Bhattacharyya and J. Mukhopadhyay. COSPEDTree-II: Improved Couplet Based Phylogenetic Supertree. In *International Conference on Bioinformatics and Biomedicine (BIBM)*, Shenzhen, China, 2016. IEEE.

[5] C. E. Bird, S. A. Karl, P. E. Smouse, and R. J. Toonen. Detecting and measuring genetic differentiation. *Phylogeography and Population Genetics in Crustacea*, 19(3), 2011.

[6] M. Blanchette, G. Bourque, and D. Sankoff. Breakpoint phylogenies. *Genome Informatics*, pages 25–34, 1997.

[7] G. Bourque and P. A. Pevzner. Genome-Scale Evolution: Reconstructing Gene Orders in the Ancestral Species. *Genome Research*, 12:26–36, 2002.

[8] S. L. Cameron, N. Lo, T. Bourguignon, G. J. Svenson, and T. A. Evans. A mitochondrial genome phylogeny of termites (Blattodea: Termitoidae): Robust support for interfamilial relationships and molecular synapomorphies define major clades. *Molecular Phylogenetics and Evolution*, 65(1):163–173, 2012.

[9] Z. Cao, B. Liao, and R. Li. A group of 3D graphical representation of DNA sequences based on dual nucleotides. *International Journal of Quantum Chemistry*, 108(9):1485–1490, 2008.

[10] I. Charon, L. Denoeud, A. Guénoche, and O. Hudry. Maximum transfer distance between partitions. *Journal of Classification*, 23(1):103–121, 2006.

[11] R. Chi and K. Ding. Novel 4D numerical representation of DNA sequences. *Chemical Physics Letters*, 407(1-3):63–67, 2005.

[12] J. M. Cummins, T. Wakayama, and R. Yanagimachi. Fate of microinjected spermatid mitochondria in the mouse oocyte and embryo. *Zygote*, 5(4):301308, 1997.

[13] W. H. Day. The complexity of computing metric distances between partitions. *Mathematical Social Sciences*, 1(3):269–287, 1981.

[14] B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetictrees. *Proceedings of the National Academy of Sciences (PNAS)*, 93(23):13429–13434, 1996.

[15] C. F. Ehret and G. D. Haller. Origin, development, and maturation of organelles and organelle systems of the cell surface in Paramecium. *Journal of Ultrastructure Research*, 9:1–42, 1963.

[16] A. Eyre-Walker and P. Awadalla. Does human mtdna recombine? *Journal of Molecular Evolution*, 53(4):430–435, 2001.

[17] J. Felsenstein. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, 39(4):783–791, 1985.

[18] J. Felsenstein. *Inferring phylogenies*. Sinauer Associates, 2003.

[19] T. L. Fulton, S. M. Wagner, C. Fisher, and B. Shapiro. Nuclear DNA from the extinct Passenger Pigeon (Ectopistes migratorius) confirms a single origin of New World pigeons. *Annals of Anatomy - Anatomischer Anzeiger*, 194(1):52–57, 2012. Special Issue: Ancient DNA.

[20] Y. Gao and L. Luo. Genome-based phylogeny of dsDNA viruses by a novel alignment-free method. *Gene*, 492(1):309–314, 2012.

[21] M. Gates. A Simple Way To Look at DNA. *Journal of Theoretical Biology*, 119(3):319–328, 1986.

[22] C. Gissi, F. Iannelli, and G. Pesole. Evolution of the mitochondrial genome of metazoa as exemplified by comparison of congeneric species. *Heredity*, 101(4):301, 2008.

[23] K. Hatje and M. Kollmar. A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. *Frontiers in Plant Science*, 3:192, 2012.

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

[25] Y. Huang and T. Wang. Phylogenetic analysis of DNA sequences with a novel characteristic vector. *Journal of Mathematical Chemistry*, 49(8):1479–1492, 2011.

[26] Y. Huang and T. Wang. New graphical representation of a DNA sequence based on the ordered dinucleotides and its application to sequence analysis. *International Journal of Quantum Chemistry*, 112(6):1746–1757, 2012.

[27] N. C. Jones and P. A. Pevzner. *An Introduction to Bioinformatics Algorithms (Computational Molecular Biology)*. The MIT Press, 2004.

[28] J. Joseph and R. Sasikumar. Chaos game representation for comparison of whole genomes. *BMC Bioinformatics*, 7(1):243, 2006.

[29] M. R. Kantorovitz, G. E. Robinson, and S. Sinha. A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, 23(13):i249, 2007.

[30] P. Kolekar, M. Kale, and U. Kulkarni-Kale. Alignment-free distance measure based on return time distribution for sequence analysis: applications to clustering, molecular phylogeny and subtyping. *Molecular Phylogenetics and Evolution*, 65(2):510–522, 2012.

[31] V. Kumar, F. Lammers, T. Bidon, M. Pfenninger, L. Kolter, M. A. Nilsson, and A. Janke. The evolutionary history of bears is characterized by gene flow across species. *Scientific Reports*, 7:46487, 2017.

[32] T. Kuramoto, H. Nishihara, M. Watanabe, and N. Okada. Determining the Position of Storks on the Phylogenetic Tree of Waterbirds by Retroposon Insertion Analysis. *Genome Biology and Evolution*, 7(12):3180–3189, 2015.

[33] B. F. Lang, M. W. Gray, and G. Burger. Mitochondrial genome evolution and the origin of eukaryotes. *Annual Review of Genetics*, 33(1):351–397, 1999.

[34] M. G. I. Langille, W. W. L. Hsiao, and F. S. L. Brinkman. Detecting genomic islands using bioinformatics approaches. *Nature Reviews Microbiology*, 8(5):373–382, 2010.

[35] C.-A. Leimeister and B. Morgenstern. Kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. *Bioinformatics*, 30(14):2000–2008, 2014.

[36] F. Lemoine, J.-B. D. Entfellner, E. Wilkinson, D. Correia, M. D. Felipe, T. Oliveira, and O. Gascuel. Renewing Felsensteins phylogenetic bootstrap in the era of big data. *Nature*, page 1, 2018.

[37] P. Leong and S. Morgenthaler. Random walk and gap plots of DNA sequences. *Bioinformatics*, 11(5):503–507, 1995.

[38] S. Levy et al. The Diploid Genome Sequence of an Individual Human. *PLOS Biology*, 5(10):1–32, 09 2007.

[39] B. Liao and K. Ding. A 3D graphical representation of DNA sequences and its application. *Theoretical Computer Science*, 358(1):56–64, 2006.

[40] B. Liao, R. Li, W. Zhu, and X. Xiang. On the Similarity of DNA Primary Sequences Based on 5-D Representation. *Journal of Mathematical Chemistry*, 42(1):47–57, 2007.

[41] B. Liao and T.-m. Wang. Analysis of Similarity/Dissimilarity of DNA Sequences Based on Nonoverlapping Triplets of Nucleotide Bases. *Journal of Chemical Information and Computer Sciences*, 44(5):1666–1670, 2004.

[42] B. Liao, Y. Zhang, K. Ding, and T. ming Wang. Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation. *Journal of Molecular Structure: THEOCHEM*, 717(1-3):199–203, 2005.

[43] L. Liu, D. K. Pearl, R. T. Brumfield, and S. V. Edwards. Estimating Species Trees using Multiple-Allele DNA Sequence Data. *Evolution*, 62(8):468–477, 2008.

[44] Y. Y. Lu, K. Tang, J. Ren, J. A. Fuhrman, M. S. Waterman, and F. Sun. CAFE: aCcelerated Alignment-FrEe sequence analysis. *Nucleic Acids Research*, 45(W1):W554–W559, 2017.

[45] B. Moret, S. Wyman, D. Bader, T. Warnow, and M. Yan. A new implementation and detailed study of breakpoint analysis. In *Proc 6th Pacific Symp Biocomputing*, pages 583–594. Hawaii, 2001.

[46] B. M. E. Moret. Phylogenetic Analysis of Whole Genomes. In *Bioinformatics Research and Application: 7th International Symposium, ISBRA*, pages 4–7, Changsha, China, 2011. Springer.

[47] A. Nandy. A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. *Current Science*, 66(4):309–314, 1994.

[48] A. Nandy, M. Harle, and S. C. Basak. Mathematical descriptors of DNA sequences: development and applications. *ARKIVOC*, 2006(9):211–238, 2006.

[49] S. Ohno. So much 'junk' DNA in our genome. *Evolution in Genetic Systems*, 23:366–370, 1972.

[50] H. H. Otu and K. Sayood. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*, 19(16):2122–2130, 2003.

[51] A. F. Palazzo and E. S. Lee. Non-coding RNA: what is functional and what is junk? *Frontiers in Genetics*, 6, 2015.

[52] R. Peng, B. Zeng, X. Meng, B. Yue, Z. Zhang, and F. Zou. The complete mitochondrial genome and phylogenetic analysis of the giant panda (Ailuropoda melanoleuca). *Gene*, 397(1):76–83, 2007.

[53] J.-M. Pons, A. Hassanin, and P.-A. Crochet. Phylogenetic relationships within the Laridae (Charadriiformes: Aves) inferred from mitochondrial markers. *Molecular Phylogenetics and Evolution*, 37(3):686–699, 2005.

[54] R. A. Pyron. Post-molecular systematics and the future of phylogenetics. *Trends in Ecology & Evolution*, 30(7):384–389, 2015.

[55] J. Qi, H. Luo, and B. Hao. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Research*, 32(suppl_2):W45–W47, 2004.

[56] M. Randić, M. Novič, and D. Plavšić. Milestones in Graphical Bioinformatics. *International Journal of Quantum Chemistry*, 113(22):2413–2446, 2013.

[57] M. Randić, M. Vračko, A. Nandy, and S. C. Basak. On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization. *Journal of Chemical Information and Computer Sciences*, 14(5):1235–1244, 2000.

[58] M. Randić, M. Vračko, J. Zupan, and M. Novič. Compact 2-D graphical representation of DNA. *Chemical Physics Letters*, 373(5-6):558–562, 2003.

[59] G. Reinert, D. Chew, F. Sun, and M. S. Waterman. Alignment-free sequence comparison (I): statistics and power. *Journal of Computational Biology*, 16(12):1615–1634, 2009.

[60] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[61] X. Shi, P. Tian, R. Lin, D. Huang, and J. Wang. Characterization of the Complete Mitochondrial Genome Sequence of the Globose Head Whiptail Cetonurus globiceps (Gadiformes: Macrouridae) and Its Phylogenetic Analysis. *PLOS One*, 11(4):688–704, 2016.

[62] G. E. Sims, S.-R. Jun, G. A. Wu, and S.-H. Kim. Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proceedings of the National Academy of Sciences (PNAS)*, 106(40):17077–17082, 2009.

[63] G. E. Sims and S.-H. Kim. Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs). *Proceedings of the National Academy of Sciences (PNAS)*, 108(20):8329–8334, 2011.

[64] P. H. A. Sneath and R. R. Sokal. *Numerical Taxonomy*. W. H. Freeman and Company, San Francisco, 1973.

[65] S.-N. Song, P. Tang, S.-J. Wei, and X.-X. Chen. Comparative and phylogenetic analysis of the mitochondrial genomes in basal hymenopterans. *Scientific Reports*, 6:20972, 2016.

[66] J. A. Tenreiro Machado. Shannon entropy analysis of the genome code. *Mathematical Problems in Engineering*, 2012, 2012.

[67] I. Ulitsky, D. Burstein, T. Tuller, and B. Chor. The average common substring approach to phylogenomic reconstruction. *Journal of Computational Biology*, 13(2):336–350, 2006.

[68] L. Wan, G. Reinert, F. Sun, and M. S. Waterman. Alignment-free sequence comparison (II): theoretical power of comparison statistics. *Journal of Computational Biology*, 17(11):1467–1490, 2010.

[69] H. Wang, Z. Xu, L. Gao, and B. Hao. A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC Evolutionary Biology*, 9(1):195, 2009.

[70] J. Wang et al. The diploid genome sequence of an Asian individual. *Nature*, 456(7218):60–65, 2008.

[71] Y.-h. Yao and T.-m. Wang. A class of new 2-D graphical representation of DNA sequences and their application. *Chemical Physics Letters*, 398(4-6):318–323, 2004.

[72] H. Yi and L. Jin. Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Research*, 41(7):e75–e75, 2013.

[73] L. Yu, Y.-W. Li, O. A. Ryder, and Y.-P. Zhang. Analysis of complete mitochondrial genome sequences increases phylogenetic resolution of bears (Ursidae), a mammalian family that experienced rapid speciation. *BMC Evolutionary Biology*, 7(1):198, 2007.

[74] W. Zhang and M. Zhang. Complete mitochondrial genomes reveal phylogeny relationship and evolutionary history of the family Felidae. *Genetics and Molecular Research*, 12:3256–3262, 2013.

[75] Z. J. Zhang. DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences. *Bioinformatics*, 25(9):1112–1117, 2009.

[76] L. Zhao, T. Gao, and W. Lu. Complete mitochondrial DNA sequence of the endangered fish (Bahaba taipingensis): Mitogenome characterization and phylogenetic implications. *ZooKeys*, (546):181, 2015.

[77] A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, 18(1):186, 2017.

[78] G. Zuo and B. Hao. CVTree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy. *Genomics, Proteomics & Bioinformatics*, 13(5):321–331, 2015.