GRaphical Footprint Based Alignment-Free Method (GRAFree) for Classifying the Species in Large-Scale Genomics

Aritra Mahapatra(✉) and Jayanta Mukherjee

Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur 721302, India
aritra.mhp@iitkgp.ac.in, jay@cse.iitkgp.ac.in

Abstract. In our study, we propose to use novel features from mitochondrial genomic sequences reflecting their evolutionary traits by a novel GRaphical footprint based Alignment-Free method (GRAFree). These features are used to classify a set of species to different classes. A novel distance measure in the feature space is also proposed to measure the proximity of these species in the evolutionary processes. The distance function is found to be a metric. Further we model the evolutionary relationships of these classes by forming a phylogenetic tree. Experimentations were carried out with 157 species covering four different classes such as, Insecta, Actinopterygii, Aves, and Mammalia. We apply our proposed distance function on the selected feature vectors for three different graphical representations of genome. The inferred trees corroborate accepted evolutionary traits. This demonstrates that our proposed distance function and feature representation can be applied to classify different species and to capture the evolutionary relationships among their classes.

Keywords: Classification · Phylogeny · Mitochondrial genome · Graphical footprint · k-nearest neighbor classifier · Hierarchical clustering

1 Introduction

The phylogeny can be considered as the clustering problem. In studying phylogeny of different species using molecular data, mostly the homologous segments of DNA sequences are observed by computational biologists. This approach is sensitive to the selection of segments (e.g. genes, coding segments, etc.) of the sequence. The mitochondrial genomes (mtDNA) are haploid, inherited maternally in most animals, and recombination is very rare event in it [4]. So the changes of mtDNA sequence occur mainly due to mutations. Hence, the evolutionary traits are more conserved in mtDNA. The mtDNA usually consists of a few numbers of non-overlapping fragments called genes.

B. Deka et al. (Eds.): PReMI 2019, LNCS 11942, pp. 105–112, 2019.
https://doi.org/10.1007/978-3-030-34872-4_12

During evolution, the genes of mtDNA very often change their order within the mtDNA and also get fragmented [2]. This violates the collinearity of homologous regions. Apart from this fact, the complexity and versatility of the data make it difficult to develop any simple method in comparative genomics [13]. Conventional methods compute the distance between sequences through computationally intensive process of multiple sequence alignment, which remains a bottleneck in using whole genomic sequences for constructing phylogeny [8]. As a result, there exist a few works to discover evolutionary features in the non-homologous regions of using alignment-free methods [17].

The existing alignment-free methods can be categorized into four types, namely, k-mer frequency based method [1], Substring based method [11], Information theory based method [6], and Graphical representation based method [21].

But they are not suitable to deal with a large number of taxa, and the size of the input sequences is also limited [3]. 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 with mathematical encoding (both numerically and graphically) [15]. It has been hypothesized that each species carries unique patterns over their DNA sequence which makes a species different from others [10]. There exist various representations of the large genome sequences through line graph by mapping the nucleotides to numeric representations. Considering a genome sequence as a signal (called the **genomic signal**), these methods analyze respective sequences using different signal processing techniques. Several techniques have been proposed to represent DNA sequences graphically from 2D space to higher dimensional space [16]. The graphical representation has a serious limitation of overlapping paths which cause loss of information [16]. GRAFree takes care of this problem 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 to classify the species according to their class.

Experimentations were carried out with a dataset of total 157 species from four different classes, namely, Insecta (insect), Actinopterygii (ray-finned fish), Aves (bird), and Mammalia (Mammal). Our proposed method, GRAFree, classifies the species based on their classes with a high accuracy.

2 Materials and Methods

2.1 Feature Space

Definition 1 (Graphical Foot Print (GFP)). *Let a sequence,* $\mathcal{S} \in \Sigma^+$, $\Sigma = \{A, T, G, C\}$. *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 or GFP-RY (Φ_{RY}): for Purine (R)/Pyrimidine (Y) [14]

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

Case-2 or GFP-SW (Φ_{SW}): for Strong H-bond (S)/Weak H-bond (W) [7]

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

Case-3 or GFP-MK (Φ_{MK}): for Amino (M)/Keto (K) [12]

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

Definition 2 (Drift of GFP). *For a length L, drift at the i^{th} position is, $\delta_i^{(L)} = \phi_{i+L}(\mathcal{S}) - \phi_i(\mathcal{S})$, where $(i+L) \leq |\mathcal{S}|$ and $\phi_i(\mathcal{S})$ the i^{th} coordinate of $\phi(\mathcal{S})$*

Considering the drifts for every $i^{th}, i = 1, 2, \ldots, n$, 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 $(m + L) = |\mathcal{S}|$

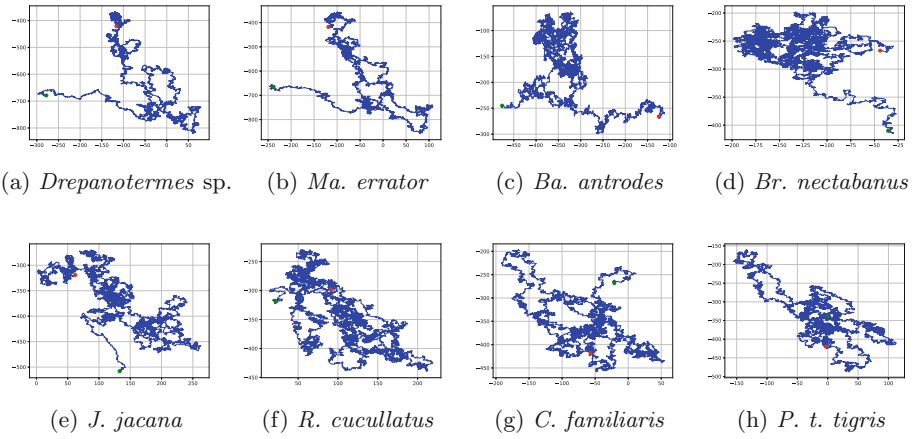
For three different cases, we denote drifts as $\Delta_{RY}^{(L)}$, $\Delta_{SW}^{(L)}$, and $\Delta_{MK}^{(L)}$, respectively.

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. Some typical examples in Fig. 1 show that the species from class insect (Fig. 1(a, b)) have the similar pattern in their drift sequences but the pattern of the drift sequence of fish (Fig. 1(c)) is different from them. This intuitively indicates that the intraclass species are closer than the interclass species.

We represent spatial distribution of these points of Δ by 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 between the eigen vector corresponding to Λ and x-axis. We make $\mathcal{F}$ number of non-overlapping equal length fragments from Δ and represent each fragment using the 5D feature descriptor.

2.2 Distance Function and Its Properties

For two sequences $\mathcal{P}$ and $\mathcal{Q}$ with the feature descriptors of $i^{th}, i \leq \mathcal{F}$ 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 $\mu_{\mathcal{P}i} = (\mu_{x\mathcal{P}i}, \mu_{y\mathcal{P}i})$



(a) *Drepanotermes* sp. (b) *Ma. errator* (c) *Ba. antrodes* (d) *Br. nectabanus*

(e) *J. jacana* (f) *R. cucullatus* (g) *C. familiaris* (h) *P. t. tigris*

Fig. 1. The Δ_{RY} for $L = 2000$ of (a, b) the insects namely, *Drepanotermes* sp. and *Macrognathotermes errator*, respectively. (c, d) the fishes namely, *Bathygadus antrodes* and *Bregmaceros nectabanus*, respectively. (e, f) the birds namely, *Jacana jacana* and *Raphus cucullatus*, respectively. (g, h) the mammals namely, *Canis familiaris* and *Panthera tigris tigris*, respectively.

and $\mu_{\mathcal{Q}i} = (\mu_{x\mathcal{Q}i}, \mu_{y\mathcal{Q}_i})$, we propose the following distance function,

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

The distance D between two sequences is found to be a metric.

2.3 Taxon Sampling and Acquiring Mitochondrial Genome

We have prepared our dataset of 157 species by selecting their mitochondrial genomes sequenced by various researchers. We consider 30, 59, 32, and 36 species from four classes such as mammal [9,22], bird [5], ray-finned fish [18,23], and insect [20], respectively. 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 indicates that the quality of data is good.

2.4 Classification and Phylogenetic Inference

We randomly select ten species from each class and consider the mean of them as the representative of the corresponding class. We consider five such representatives for each class. Finally, we apply k-nearest neighbor classifier based on

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

the representatives to classify the species. In k-nearest neighbor classification we apply our proposed distance function (Eq. (4)) for a given value of L, $\mathcal{F}$, and α.

We also apply the same distance function to compute pairwise distances among representatives of four classes.. By applying a hierarchical clustering technique, UPGMA [19], over this distance matrix, we get a phylogenetic tree. We compute phylogenetic trees each separately using representation schemes, such as, GFP-RY, GFP-SW, and GFP-MK (refer to Eqs. (1), (2), (3), respectively).

3 Results and Discussion

3.1 Comparison

For a given number of fragments, $\mathcal{F}$, our proposed features vector is $5\mathcal{F}$ dimensional. We compare our feature vector with an existing k-mer based feature representation [1].

As our input data is genomic sequence which contains only four characters, A, T, G, C, hence, for k length of word, the length of the k-mer based feature is 4^k. So, the k-mer feature vector takes a significantly larger memory space than our proposed feature vector. We have applied Euclidean, Canberra, and Chebyshev distance functions on both k-mer based feature vector and our proposed feature vector for different parameters value. It is observed that, our proposed feature representation using GFP outperforms the Canberra and Chebyshev methods using k-mer based representation and shows a comparable result with k-mer based Euclidean method in classifying the species.

We also compare the performance of our proposed distance function with the same distance functions. It is observed that all of these distance functions perform differently for different parameters value. However, the best performance of the proposed distance function is comparable to the performance of Euclidean and Canberra. The Chebyshev distance function does not perform well in this task of classification.

Apart from that, the phylogenetic trees derived from the proposed distance function are consistent and supports the mostly accepted hypotheses, whereas, the trees derived from Euclidean, Canberra, and Chebyshev distance functions give different relationships among the four selected classes for GFP-RY, GFP-SW, and GFP-MK.

3.2 Observations

Here, we present the clusters generated by GRAFree using the whole mitochondrial genome sequences of the selected species. We enumerate the value of L, $\mathcal{F}$, and α from 50 to 5000, 1 to 200, and 0 to 1, respectively. We consider three different cases, such as GFP-RY, GFP-SW, and GFP-MK (please refer to Eq. (1), (2) and (3), respectively). It is observed that for a given value of the parameters (L, $\mathcal{F}$, and α) the classifier performs differently for GFP-RY, GFP-SW, and GFP-MK. This fact reveals that for GFP-RY, GFP-SW, and GFP-MK, the species contains different signatures in their corresponding drifts.

However, the best accuracy we get 95.5%, 96.2%, and 96.2% for GFP-RY (with $L = 100$, $\mathcal{F} = 150$, and $\alpha = 0.50$), GFP-SW (with $L = 300$, $\mathcal{F} = 55$, and $\alpha = 1$), and GFP-MK (with $L = 50$, $\mathcal{F} = 50$, and $\alpha = 0.50$), respectively. The confusion matrices for the best cases are shown in Table 1. This accuracy is moderately better compared to those obtained using the Euclidean and Chebyshev distance functions and comparable to the Canberra distance function. For the same set of parameters of each case, the accuracy of the Euclidean are 96.82%, 78.34%, and 94.27%, respectively. The accuracy of the Canberra for the same set of parameters are 96.82%, 97.45%, and 98.10%, respectively. The accuracy of the Chebyshev is as low as <70% for all of the three cases.

Table 1. Confusion matrix of four classes Mammal (Mamm), Bird (Bird), Fish (Fish), and Insect (Inse) for GFP-RY, GFP-SW, and GFP-MK

Original	Prediction											
	GFP-RY				GFP-SW				GFP-MK			
	Mamm	Bird	Fish	Inse	Mamm	Bird	Fish	Inse	Mamm	Bird	Fish	Inse
Mamm	28	2	0	0	30	0	0	0	30	0	0	0
Bird	0	58	1	0	4	55	0	0	0	59	0	0
Fish	0	0	32	0	0	0	32	0	1	4	27	0
Inse	2	2	0	32	0	0	2	34	1	0	0	35

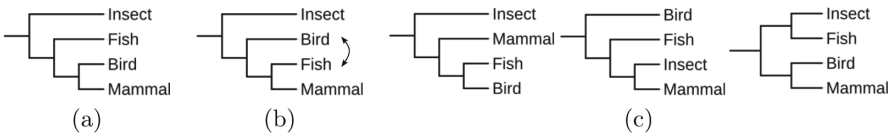


Fig. 2. Derived trees after applying the UPGMA on the representatives of classes by using (a) GRAFree, (b) both Euclidean and Canberra, and (c) Chebyshev. The arrow in Fig. (b) indicates that the position of bird and fish are interchangeable in different topologies derived from both Euclidean and Canberra.

For a particular window of $L = [50, 300]$, $\mathcal{F} = [50, 150]$, and $\alpha = [0.50, 1]$, the accuracy of our method are found to be $(84.36 \pm 4.28)\%$, $(83.42 \pm 3.15)\%$, and $(84.8 \pm 2.91)\%$ for GFP-RY, GFP-SW, and GFP-MK, respectively. So it can be noticed that the proposed technique is very sensitive to the parameter values.

To derive the phylogenetic relationships among the selected classes, we consider all (here five) representatives of each class. We apply our proposed distance measure to derive the pairwise distances followed by the UPGMA. For a given value of L, $\mathcal{F}$, and α the trees are generated for three different cases (GFP-RY, GFP-SW, and GFP-MK). It is observed that for both GFP-RY, GFP-SW, and GFP-MK all the representatives of a particular class placed under a same clade. It is also observed that the interclass relationships of the three trees carry the same topology (Fig. 2) which shows that the GRAFree is robust in deriving the phylogeny among different classes.

Table 2. Time and space complexity to compute distance between two sequences

Methods	Time complexity		Space complexity
	Deriving features	Computing distance	
GRAFree	$\mathcal{O}(M-L+1)$	$\mathcal{O}((M-L+1)\mathcal{F})$	$\mathcal{O}(\mathcal{F})$
Euclidean	$\mathcal{O}(k(M-k+1))$	$\mathcal{O}(4^k)$	$\mathcal{O}(4^k)$
Canberra	$\mathcal{O}(k(M-k+1))$	$\mathcal{O}(4^k)$	$\mathcal{O}(4^k)$
Chebyshev	$\mathcal{O}(k(M-k+1))$	$\mathcal{O}(4^k)$	$\mathcal{O}(4^k)$

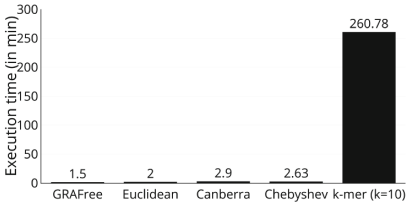


Fig. 3. The execution time for different methods. All the methods are executed in the same system. System configuration: 16 GB RAM, Intel Core i5 processor.

3.3 Complexity Analysis

We have derived both time and space complexities of GRAFree and the other reference methods. It is found that the GRAFree is most time and space economic method among all the reference methods (refer to Table 2). It can also be noticed that the execution time of GRAFree is less than all the other methods (Fig. 3)

4 Conclusion

We have proposed a $5\mathcal{F}$-dimensional feature space and a new metric for classifying the species using large scale genomic features in the method GRAFree. GRAFree uses the graphical representation of the genome. In this study we have selected three graphical representations of a genome considering residues independently. This method can classify the species based on their class (taxonomy rank). The selection of the value of the parameters used in the GRAfree needs further study. We observe presence of evolutionary traits among the selected classes in the proposed feature descriptor extracted from the whole mitochondrial sequences. These exhibit the effectiveness of the proposed feature representation along with the metric for measuring the pairwise distances of species.

References

1. Bernard, G., Ragan, M.A., Chan, C.X.: Recapitulating phylogenies using k-mers: from trees to networks. F1000Research **5**, 2789 (2016)
2. Bernt, M., Braband, A., Schierwater, B., Stadler, P.F.: Genetic aspects of mitochondrial genome evolution. Mol. Phylogenet. Evol. **69**(2), 328–338 (2013)

3. Bourque, G., Pevzner, P.A.: Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Res. **12**, 26–36 (2002)
4. Eyre-Walker, A., Awadalla, P.: Does human mtDNA recombine? J. Mol. Evol. **53**(4), 430–435 (2001)
5. Fulton, T.L., Wagner, S.M., Fisher, C., Shapiro, B.: Nuclear DNA from the extinct Passenger Pigeon (Ectopistes migratorius) confirms a single origin of New World pigeons. Ann. Anat. - Anatomischer Anzeiger **194**(1), 52–57 (2012). Special Issue: Ancient DNA
6. Gao, Y., Luo, L.: Genome-based phylogeny of dsDNA viruses by a novel alignment-free method. Gene **492**(1), 309–314 (2012)
7. Gates, M.: A simple way to look at DNA. J. Theor. Biol. **119**(3), 319–328 (1986)
8. Huang, Y., Wang, T.: Phylogenetic analysis of DNA sequences with a novel characteristic vector. J. Math. Chem. **49**(8), 1479–1492 (2011)
9. Kumar, V., et al.: The evolutionary history of bears is characterized by gene flow across species. Sci. Rep. **7**, 46487 (2017)
10. Langille, M.G.I., Hsiao, W.W.L., Brinkman, F.S.L.: Detecting genomic islands using bioinformatics approaches. Nat. Rev. Microbiol. **8**(5), 373–382 (2010)
11. Leimeister, C.A., Morgenstern, B.: Kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. Bioinformatics **30**(14), 2000–2008 (2014)
12. Leong, P., Morgenthaler, S.: Random walk and gap plots of DNA sequences. Bioinformatics **11**(5), 503–507 (1995)
13. Moret, B.M.E.: Phylogenetic analysis of whole genomes. In: Chen, J., Wang, J., Zelikovsky, A. (eds.) ISBRA 2011. LNCS, vol. 6674, pp. 4–7. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21260-4_3
14. Nandy, A.: A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. Curr. Sci. **66**(4), 309–314 (1994)
15. Nandy, A., Harle, M., Basak, S.C.: Mathematical descriptors of DNA sequences: development and applications. ARKIVOC **2006**(9), 211–238 (2006)
16. Randić, M., Novič, M., Plavšić, D.: Milestones in graphical bioinformatics. Int. J. Quantum Chem. **113**(22), 2413–2446 (2013)
17. Ren, J., et al.: Alignment-free sequence analysis and applications. Ann. Rev. Biomed. Data Sci. **1**, 93–114 (2018)
18. Shi, X., Tian, P., Lin, R., Huang, D., Wang, J.: 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)
19. Sneath, P.H.A., Sokal, R.R.: Numerical Taxonomy. W. H. Freeman and Company, San Francisco (1973)
20. Song, S.N., Tang, P., Wei, S.J., Chen, X.X.: Comparative and phylogenetic analysis of the mitochondrial genomes in basal hymenopterans. Sci. Rep. **6**, 20972 (2016)
21. Xie, G.S., Jin, X.B., Yang, C., Pu, J., Mo, Z.: Graphical representation and similarity analysis of DNA sequences based on trigonometric functions. Acta Biotheoretica **66**(2), 113–133 (2018)
22. Zhang, W., Zhang, M.: Complete mitochondrial genomes reveal phylogeny relationship and evolutionary history of the family Felidae. Genet. Mol. Res. **12**, 3256–3262 (2013)
23. Zhao, L., Gao, T., Lu, W.: Complete mitochondrial DNA sequence of the endangered fish (Bahaba taipingensis): mitogenome characterization and phylogenetic implications. ZooKeys **546**, 181 (2015)