Communications in Information and Systems
Volume 22, Number 3, 383–400, 2022

A graph-theoretical approach to DNA similarity analysis

Dong Quan Ngoc Nguyen, Lin Xing,
Phuong Dong Tan Le, and Lizhen Lin

One of the very active research areas in bioinformatics is DNA similarity analysis. There are several approaches using alignment-based or alignment-free methods to analyze similarities/dissimilarities between DNA sequences. In this work, we introduce a novel representation of DNA sequences, using n-ary Cartesian products of graphs for arbitrary positive integers n. Each of the component graphs in the representing Cartesian product of each DNA sequence contain combinatorial information of certain tuples of nucleotides appearing in the DNA sequence. We further introduce a metric space structure to the set of all Cartesian products of graphs that represent a given collection of DNA sequences in order to be able to compare different Cartesian products of graphs, which in turn signifies similarities/dissimilarities between DNA sequences. We test our proposed method on several datasets including Human Papillomavirus, Human rhinovirus, Influenza A virus, and Mammals. We compare our method to other methods in literature, which indicates that our analysis results are comparable in terms of time complexity and high accuracy, and in one dataset, our method performs the best in comparison with other methods.

Keywords and phrases: DNA similarity, graph representations, metric space.

1. Introduction

DNA similarity analysis is one of the main areas in bioinformatics. Two main approaches using in analyzing DNA sequences are *alignment-based methods* and *alignment-free methods*. Among the alignment-based methods, the multiple sequence alignment (MSA) method has the highest accuracy in analyzing similarities/dissimilarities between DNA sequences, but its time complexity increases extremely large for large datasets of DNA sequences whose lengths are sufficiently long. Thus searching for alignment-free methods that is effective in time complexity as well as having a rea-

sonably high accuracy has been a very research problem in DNA similarity analysis. Alignment-free methods, using geometric approaches share a similar strategy that first embed DNA sequences into vectors in a Euclidean space, and then compute the similarity distance matrix, based on the underlying Euclidean distance, whose entries are distances between the representing vectors of DNA sequences. For papers containing alignment-free methods that use this approach, the reader is, for example, referred to [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20].

In this work, an alignment-free method is proposed, which avoids the embedding into Euclidean spaces and the use of numerical encoding of DNA sequences as vectors in a Euclidean space. The main observation in our approach is that every DNA sequence can be viewed as a string of letters–a combinatorial object in which each letter in the string is from the alphabet consisting of four nucleotides A, C, G, T. Using this observation, for an arbitrary positive integer n, and n positive integers $d_1, \ldots, d_n$, our method allows to represent each DNA sequence as an n-ary Cartersian product of graphs, the ith component of which contains combinatorial information of d_i-tuples of nucleotides appearing in the DNA sequence. In this way, the set of all n-ary Cartesian products of graphs can be viewed as an analogue of n-dimensional Euclidean spaces that contain the representing vectors of DNA sequences as in the traditional alignment-free methods. In order to be able to compare n-ary Cartesian products of graphs that represent DNA sequences in our proposed method, a variety of metric space structures on graphs is utilized such as the metric space structures equipped with spectral distance metrics or matrix distance metrics. There are also other types of metric space structures on graphs [21]. Fixing, once and for all, a distance metric for each component in the set of all n-ary Cartesian products of graphs, we can equip this set with a metric space structure by simply taking the maximum of the values of all distances [22]. This procedure converts a given collection of DNA sequences into a metric space consisting of n-ary Cartesian products of graphs that contain combinatorial information of all d_i-tuples of nucleotides in DNA sequences for any $1 \leq i \leq n$, while also carrying a distance metric that allows to compare similarities/dissimilarities between n-ary Cartesian products of graphs that represent DNA sequences. The combinatorial information of tuples of nucleotides in DNA sequences can grow extremely large when allowing n to become sufficiently large. Thus in our approach, suitable values of n are chosen to assure the fast time complexity while maintaining high accuracy in analyzing similarities/dissimilarities between DNA sequences. In comparison with other methods in literature such as the state-of-the-art Clustal Omega [23] – an MSA method, and the

Fourier transform method in [24], an alignment-free method, the method proposed in this paper performs comparable in accuracy and time complexity, and in some datasets (see Section 4), the proposed method performs the best among all the methods that are used to compare.

The structure of our paper is as follows. In Section 2, one-dimensional and high-dimensional graph representations of DNA sequences as well as their metric space structures are introduced, which will be used in experimental analysis. In Section 3, the proposed method is described in detail. In Section 4, the proposed method is applied to test on several real datasets including Human Papillomavirus (HPV) [25, 26], Human rhinovirus (HRV) [27], Influenza A virus [28, 29], and Mammals [30]. In the supplemental file, we tabulate the GenBank[1] accession numbers of DNA sequences contained in the datasets on which we test our method (http://www.intlpress.com/site/pub/files/_supp/cis/2022/0022/0003/cis-2022-0022-0003-s002.pdf).

2. Graph-theoretic representation of DNA sequences

In this section, several representations of DNA sequences using graph theory are described, which allows to equip a given collection of DNA sequences with metric space structures. Such metric space structures of a collection of DNA sequences will be exploited in the proposed method for analyzing similarities/dissimilarities between DNA sequences.

We begin by introducing one-dimensional graph representation of DNA sequences.

Let α denote a DNA sequence of length m of the form $a_1a_2\cdots a_m$, where each a_i is one of the nucleotides A, C, G, T. Let d be a positive integer such that $d < m$. In general, it suffices to choose small values of d between 2 and 10. Let w be a sufficiently small positive integer, and let h be the largest positive integer such that $0 \leq m - (wh + 1) < d - 1$. The last inequality condition assures that there are exactly h d-tuples of nucleotides appearing the sequence α, and the remaining nucleotides appearing after the hth d-tuple do not have enough d letters to form another d-tuple. Using such a pair of integers (d, w), we associate to α a *weighted undirected graph*, denoted by $G_{(d,w)}$ whose nodes are constructed using d consecutive nucleotides in the sequence α, and two nodes form an edge when they are represented by two consecutive sequences of d nucleotides that are exactly w nucleotides apart from each other. For the rest of the paper, w is called the *window* of α that indicates the *distance* one needs to walk along the sequence α to construct nodes in $G_{(d,w)}$. The number of nodes in $G_{(d,w)}$ is at most h.

[1]See https://www.ncbi.nlm.nih.gov/genbank/

The first node, say v_1, in $G_{(d,w)}$ is represented by the ordered d-tuple $a_1a_2\cdots a_d$ consisting of the first d consecutive nucleotides in the DNA sequence α. In order to construct the second node, we start at nucleotide a_{w+1} which is exactly w nucleotides apart from a_1, and form the second node v_2 of the form $a_{w+1}a_{w+2}\cdots a_{w+d}$. In general, by induction, we can define the k-th node, say v_k, with $1 \leq k \leq h$, as the ordered d-tuple $a_{w(k-1)+1}a_{w(k-1)+2}\cdots a_{w(k-1)+d}$. Since each a_i is one of the nucleotides A, C, G, T, there are only finitely many choices for each node v_k. In fact, there are exactly 4^d choices for each v_i. So it may occur that the construction can result in $v_i = v_j$ for some $i \neq j$, which implies that the set of nodes of $G_{(d,w)}$ consists of all the distinct elements in the multiset $\{v_1, v_2, \ldots, v_h\}$.

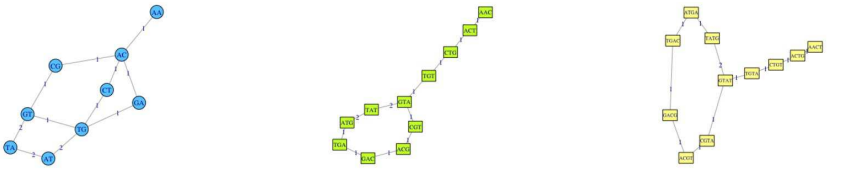
Two distinct nodes v_i, v_j form an edge $e = (v_i, v_j)$ in $G_{(d,w)}$ if there exists an integer $1 \leq k \leq h$ such that either of the following is true.

(i) $v_i = a_{w(k-1)+1}a_{w(k-1)+2}\cdots a_{w(k-1)+d}$ and $v_j = a_{wk+1}a_{wk+2}\cdots a_{wk+d}$.
(ii) $v_j = a_{w(k-1)+1}a_{w(k-1)+2}\cdots a_{w(k-1)+d}$ and $v_i = a_{wk+1}a_{wk+2}\cdots a_{wk+d}$.

In other words, $e = (v_i, v_j)$ is represented by two consecutive ordered d-tuples appearing in the DNA sequence α. The *weight* of e is defined to be the number of times that the pair (v_i, v_j) that represents the edge e, appears as two consecutive ordered d-tuples in the sequence α.

Example 2.1. Let α denote the DNA sequence AACTGTATGACGTATG of length $m = 16$. We illustrate the above construction to represent α as a weighted undirected graph $G_{(2,1)}$, where $d = 2$ and $w = 1$. Such a graph representation is called a *dinucleotide* representation with window 1. Using the above construction, we obtain that $v_1 = AA$, $v_2 = AC$, $v_3 = CT$, $v_4 = TG$, $v_5 = GT$, $v_6 = TA$, $v_7 = AT$, $v_8 = TG$, $v_9 = GA$, $v_{10} = AC$, $v_{11} = CG$, $v_{12} = GT$, $v_{13} = TA$, $v_{14} = AT$, and $v_{15} = TG$. Thus the set of nodes of $G_{(2,1)}$ consists of all distinct elements in the multiset $\{v_1, v_2, \ldots, v_{15}\}$, which implies that the set of nodes of $G_{(2,1)}$ is AA, AC, AT, CG, CT, GA, GT, TA, TG. Note that GT and TG are distinct nodes since we consider the ordered 2-tuples appearing in α. The graph $G_{(2,1)}$ that represents α is illustrated in Figure 1(A) in which the positive integer appearing on each edge indicates its weight. For example, the edge (AC, CG) appears in $G_{(2,1)}$ with weight 1 since AC, CG appear as two consecutive ordered 2-tuples in α exactly one time.

When (d, w) is $(3, 1)$ (resp., $(4, 1)$), we, in a similar way as above, obtain the *trinucleotide* representation with window 1 (resp. the *tetranucleotide* representation with window 1) of α. See Figure 1(B) and (C) for the graphs $G_{(3,1)}$ and $G_{(4,1)}$.



(a) Dinucleotide representation $G_{(2,1)}$

(b) Trinucleotide representation $G_{(3,1)}$

(c) Tetranucleotide representation $G_{(4,1)}$

Figure 1: Graph representations of DNA sequence AACTGTATGACGTATG.

There is a natural generalization of one-dimensional graph representation to high-dimensional graph representations of DNA sequences as follows. To each DNA sequence α we associate a Cartersian product of weighted undirected graphs. Let α be a DNA sequence of length m. Let $(d_1, w_1), \ldots, (d_n, w_n)$ be n pairs of positive integers, where each d_i satisfies $d_i < m$, and the w_i are chosen to be sufficiently small. In our experimental analysis, we choose $w_i = 1$, which is fast in time complexity as well as high in accuracy. For each (d_i, w_i) with $1 \leq i \leq n$, we associate to α the weighted undirected graph $G_{(d_i,w_i)}$. We then combine all the graphs $G_{(d_i,w_i)}$, and associate to the DNA sequence α the n-ary Cartersian product $\prod_{i=1}^{n} G_{(d_i,w_i)}$, which can be viewed as a high-dimensional graph representation of α.

As illustration, let α be the DNA sequence as in Example 2.1. Let $(d_1, w_1) = (2, 1)$, $(d_2, w_2) = (3, 1)$, and $(d_3, w_3) = (4, 1)$. Using the above construction, we obtain the 3-ary Cartersian product $G_{(2,1)} \times G_{(3,1)} \times G_{(4,1)}$ that represents the sequence α, where $G_{(2,1)}$, $G_{(3,1)}$, and $G_{(4,1)}$ are the graphs in Example 2.1 and Figure 1.

Once a graph representation of DNA sequences is chosen in the above method, such a Cartesian product representation of DNA sequences is equipped with a metric to convert a given collection of DNA sequences into a metric space. Let $\mathcal{C}$ be a finite set of DNA sequences, the minimum of whose lengths is m. Let $(d_1, w_1), \ldots, (d_n, w_n)$ be n pairs of positive integers for some integer $n \geq 1$, where each d_i satisfies $d_i < m$, and the w_i are chosen to be sufficiently small. Let $\mathcal{G}_n$ be the set of all n-ary Cartesian products of weighted undirected graphs. As described above, for each DNA sequence $\alpha \in \mathcal{C}$, there is an n-ary Cartersian product $\prod_{i=1}^{n} G^{\alpha}_{(d_i,w_i)}$ of weighted undirected graph that represents α. Thus we obtain a map $\Gamma_n : \mathcal{C} \to \mathcal{G}_n$ that sends each DNA

sequence α to its associated n-ary Cartesian product $\prod_{i=1}^{n} G^{\alpha}_{(d_i,w_i)}$. The map Γ_n is called an *n-ary Cartersian product representation of $\mathcal{C}$*.

It is well-known that there are many metric space structures for a collection of graphs [21], i.e. for a given collection of graphs $\mathcal{X}$, there exist several distance metrics $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ that satisfy the following conditions:

(D1) (**symmetry**) $d(G_1, G_2) = d(G_2, G_1)$ for any graphs G_1, G_2 in $\mathcal{X}$; and
(D2) (**triangle inequality**) $d(G_1, G_2) \leq d(G_1, G_3) + d(G_2, G_3)$ for any graphs G_1, G_2, G_3 in $\mathcal{X}$.

In general, we do not require that $d(G_1, G_2) = 0$ if and only if $G_1 = G_2$, as in the traditional notion of a distance metric in mathematics [22]. In practice, and in our experimental analysis, we rarely encounter two graphs G_1, G_2 that represent two DNA sequences such that $d(G_1, G_2) = 0$. It suffices to obtain that G_1 is very similar to G_2 from the fact that $d(G_1, G_2)$ is sufficiently small from which we wish to deduce that the DNA sequences that G_1, G_2 represent are similar.

For such a distance metric d satisfying (D1) and (D2) as above, we can equip the collection $\mathcal{X}$ of graphs with a metric space structure associated to d, which allows us to compare similarities/dissimilarities between graphs contained in $\mathcal{X}$. For the proposed method in this paper, the **edit distance** [21][2] is used to equip the collection of graphs that represent a given collection of DNA sequences, with a metric space structure. Note that the edit distance for DNA similarity analysis used in this paper scales linearly or near-linearly in the size in the graphs associated to the DNA sequences.

Using the edit distance as a distance metric on a collection of graphs, a distance metric for a collection of n-ary Cartersian products of graphs is naturally constructed as follows. Let $\mathcal{G}_n$ denote a collection of n-ary Cartersian products of graphs of the form $\prod_{i=1}^{n} G_i$, where the G_i are weighted undirected graphs. For each $1 \leq i \leq n$, let $\mathcal{X}_i$ denote the collection of weighted undirected graphs G_i appearing in the i-th component of the n-ary Cartersian products of graphs in $\mathcal{G}_n$. There is a natural bijection between $\mathcal{G}_n$ and the Cartesian product $\prod_{i=1}^{n} \mathcal{X}_i$.

Let d denote the edit distance which is a distance metric on each $\mathcal{X}_i$. We can define a distance metric $d_{\max} : \mathcal{G}_n \times \mathcal{G}_n \to [0, \infty)$ by setting

$$d_{\max}\Big(\prod_{i=1}^{n} G_i, \prod_{i=1}^{n} G'_i\Big) = \max\left(d(G_1, G'_1), \cdots, d(G_n, G'_n)\right). \tag{1}$$

[2]The Python library that implements the above distances used in this work can be accessed from the GitHub of Peter Wills (see https://github.com/peterewills/netcomp).

It is not difficult to verify that $d_{\max}$ satisfies the requirements of a distance metric, say (D1) and (D2) above.

3. Geometric graph representation method (GGRT)

In this section, an alignment-free method is proposed for analyzing similarities/dissimilarities between DNA sequences. The proposed method for reconstructing a phylogenetic tree of DNA sequences, using graph representations described in Section 2, is described in the following algorithm:

Algorithm 1: Geometric Graph Representation Method (GGRT)

Input : A collection $\mathcal{C}$ consisting of m DNA sequences $\alpha_1, \ldots, \alpha_m$.
Output: Phylogenetic tree of the DNA sequences

1. Choose n tuples of positive integers (d_i, w_i) for $1 \leq i \leq n$, where the d_i is less than the minimum of the lengths of $\alpha_1, \ldots, \alpha_m$, and the w_i are sufficiently small. In practice, and in our experimental analysis, we choose $w_i = 1$ for all i.
2. Construct high-dimensional graph representation (HDGR) of each DNA sequence α_i to obtain a finite collection $\mathcal{G}_n$ of n-ary Cartesian products of graphs $\mathcal{PG}_{\alpha_1}, \ldots, \mathcal{PG}_{\alpha_m}$, where the $\mathcal{PG}_{\alpha_i}$ is the n-ary Cartesian product of graphs, corresponding to the DNA sequence α_i
3. Associate the distance metric $d_{\max}$ on $\mathcal{G}_n$ as in Section 2, where each component distance metric is the edit distance d.[3]
4. Compute the distance matrix of dimensions $m \times m$ whose (i, j)-entry is the distance $d(\mathcal{PG}_{\alpha_i}, \mathcal{PG}_{\alpha_j})$.
5. Construct the phylogenetic tree of the DNA sequences from the distance matrix in Step 3, using UPGMA algorithm [31].

4. Experimental analysis

In this section, we describe our experimental analysis, and the results obtained from applying our proposed method in Section 3 to several real datasets including Human Papillomavirus (HPV) [25, 26], Human rhinovirus (HRV) [27], Influenza A virus [28, 29], and Mammals [30]. The GenBank[4] accession numbers of DNA sequences contained in these datasets are listed in the supplemental file. Computations in this research are implemented on a PC with configuration of Intel Core i7, CPU 2.50 GHz, and 16 GB 1600 MHz DDR3.

[4]See https://www.ncbi.nlm.nih.gov/genbank/

Each of the above datasets is represented by a finite collection $\mathcal{C}$ of DNA sequences $\alpha_1, \ldots, \alpha_l$ for various lengths l. We apply the proposed method in Section 3 for the collection $\mathcal{C}$. More precisely, in *Step 1*, we fix, once and for all, $n = 3$, $(d_1, w_1) = (2, 1)$, $(d_2, w_2) = (3, 1)$, and $(d_3, w_3) = (4, 1)$. Thus in *Step 2*, we, for each α_i, obtain a 3-ary Cartersian product of graphs $G_{(2,1),\alpha_i} \times G_{(3,1),\alpha_i} \times G_{(4,1),\alpha_i}$ that represents α_i. Here the construction of graphs $G_{(2,1),\alpha_i}, G_{(3,1),\alpha_i}, G_{(4,1),\alpha_i}$ and their 3-ary Cartesian products follow Section 2. In other words, $G_{(2,1),\alpha_i}$, $G_{(3,1),\alpha_i}$, and $G_{(4,1),\alpha_i}$ are dinucleotide, trinucleotide, and tetranucleotide representations with window 1 of α_i, respectively. Thus one obtains a collection $\mathcal{G}$ of exactly l 3-ary Cartesian products of graphs of the above form that represent all the α_i. See Example 2.1 for an explicit example of a DNA sequence, and its graph representations.

In carrying out computations, several representations of DNA sequences as in Section 2 are used, but the results are approximately similar, and the above representation provides us with the *fastest time complexity* and *highest accuracy* in analyzing similarities/dissimilarities between DNA sequences. In addition to edit distance, spectral distance [21] is also used in *Step 3* of the proposed GGRT method to compare with the GGRT method using edit distance in this paper. Using the time complexity summary in [21], edit distances have faster time complexity than spectral distances. The proposed method is compared with other methods in literature such as the Fourier transform method developed in [24] and the state-of-the-art alignment method for DNA similarity analysis called Clustal Omega in [23]. The proposed method using edit distance performs the best in comparison with the two methods in [23] and [24] when applying to test some of the above datasets in terms of time complexity as well as accuracy. Tables 1 and 2 summarize time complexity and accuracy in applying our proposed methods, using both edit distances and adjacency spectral distances, to the datasets HPV, HRV, Influenza A virus, and Mammals. The tables also list time complexity and accuracy, using the methods in [23] and [24]. The computations and reconstruction of phylogenetic trees using Clustal Omega were produced in [24].

4.1. Influenza A virus

We first consider the dataset of Influenza A viruses. Influenza A viruses are very dangerous because they have a diverse range of hosts including birds, horses, swine, and humans. These viruses have been a serious health threat to humans and animals [32], and are known to have high degree of genetic and antigenic variability [28, 29]. Some subtypes of Influenza A

Table 1: Time complexity using our proposed methods in comparison with those in [24] and [23]

Method	Influenza	HPV	Mammals	HRV
GGRT (Edit distance)	0.51s (see Fig. 2)	43.31s (see Fig. 3)	4.34s (see Fig. 4)	8.15s (see Fig. 5)
GGRT (Spectral distance)	54.67s	2h 29min	42.60s	10min
Clustal Omega in [23]	9s	2h 17min	10min	19min 35s
Fourier transform method in [24]	< 1s	< 30s	4s	7s
Number of DNA sequences	38	400	31	116
Lengths of DNA sequences	$1350 - 1467$	$7814 - 10424$	$16338 - 17447$	$6944 - 7458$

Table 2: Number of misclassified sequences using our proposed methods in comparison with those in [23] and [24]

Method	Influenza	HPV	Mammals	HRV
GGRT (Edit distance)	1 (A/turkey/VA/H5N1)[2] (see Fig. 2)	1 (HPV11-14) (see Fig. 3)	0 (see Fig. 4)	0 (see Fig. 5)
GGRT (Spectral distance)	0	1 (HPV11-14)	2 (Rabbit, Pig)	1 (C c025*)
Clustal Omega in [23]	1 (A/turkey/VA/H5N1)[2]	0	1 (Squirrel)	0
Fourier transform method in [24]	1 (A/duck/Guangxi/ H1N1)	1 (one from HPV11)	1 (Squirrel)	0

viruses are even lethal including H1N1, H2N2, H5N1, H7N3, and H7N9. The GGRT method is tested on the dataset consisting of 38 Influenza A virus genomes. From Figure 2, and Tables 1 and 2, the GGRT method using edit distance incorrectly identifies one subtype of Influenza A viruses. In terms of time complexity, the GGRT method using edit distance is comparable with the Fourier transform method in [24], and both methods perform the best in terms of time complexity. Figures 2 illustrates the phylogenetic trees of Influenza A viruses, based on the GGRT method.

[2]a part of H1N1 viruses are incorrectly grouped with H5N1 subtype.

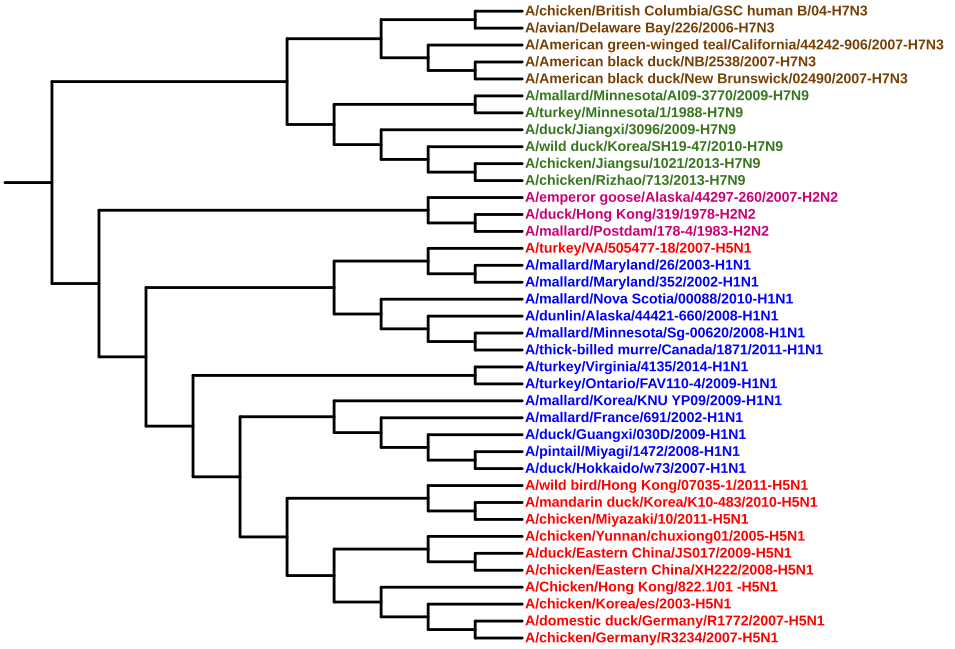


Figure 2: Phylogenetic tree of Influenza based on the GGRT method.

4.2. Human Papillomavirus (HPV)

In this subsection, we consider the dataset of Human Papillomavirus (HPV). Human Papillomavirus is mostly responsible for cervical cancer which is the second most common cancer among women [25]. The GGRT method is tested on the data set of 400 HPV genomes. In terms of time complexity and accuracy, the GGRT method using edit distance is similar to the Fourier transform method in [24]. The GGRT method incorrectly identifies one HPV genome HPV11-14. In terms of time complexity. See Figures 3 for the phylogenetic trees of HPV, using the GGRT method.

4.3. Mammals

It is known that there is a rapid mutation rate in the mitochondrial genome. In 2011, Deng et al. [30] classified a complete mitochondrial DNA dataset of 31 mammalian genome sequences from GenBank. The dataset was classified into 7 groups consisting of *Carnivore, Perissodactyla, Cetacea and Artiodactyla, Lagomorpha, Rodentia, Primates, and Erinaceomorpha*. In this

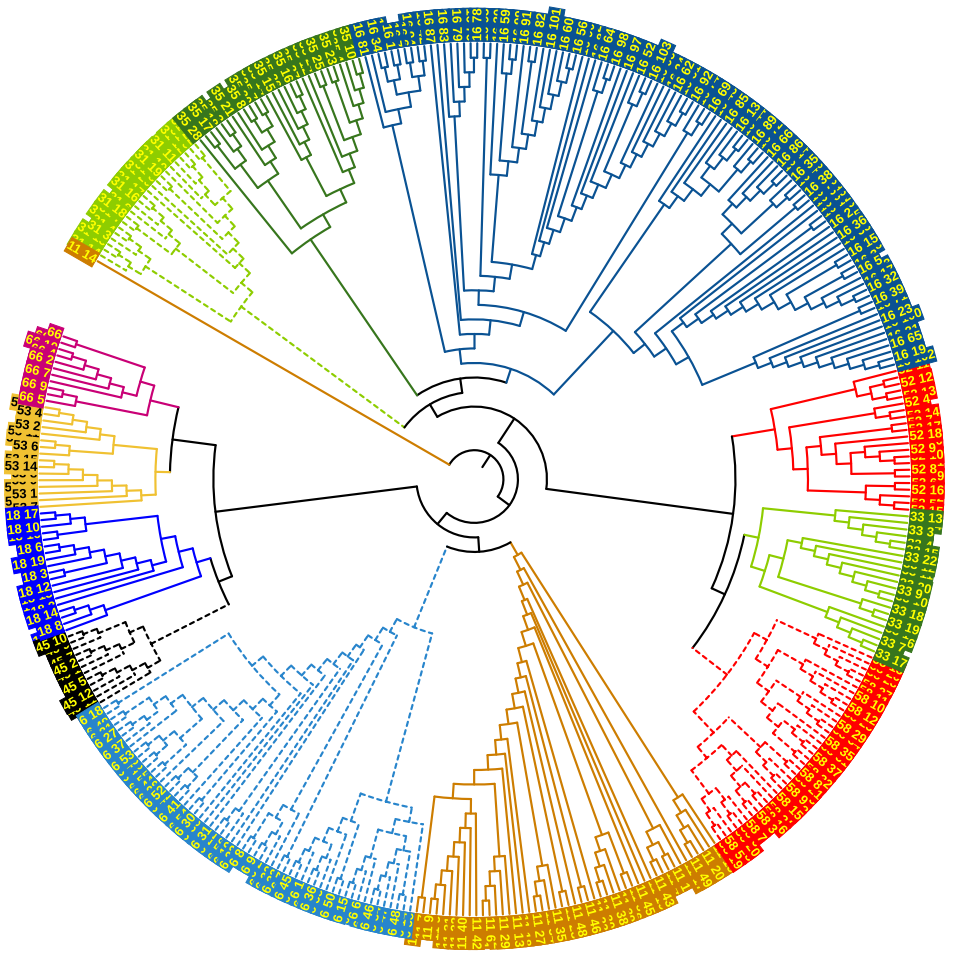


Figure 3: Phylogenetic tree of HPV based on the GGRT method.

subsection, the GGRT method is tested on the same datatse. The GGRT method (see Figure 4 and Tables 1 and 2) correctly groups 31 mammalian genome sequences into their corresponding 7 groups. Both Clustal Omega and the Fourier transform method in [24] have a misplacement. It took Clustal Omega 10 minutes and 9 seconds for the classification, and the Fourier transform method in [24] 4 seconds for the classification. It took the GGRT method using edit distance 4.34 seconds to correctly classify 31 mammalian genome sequences.

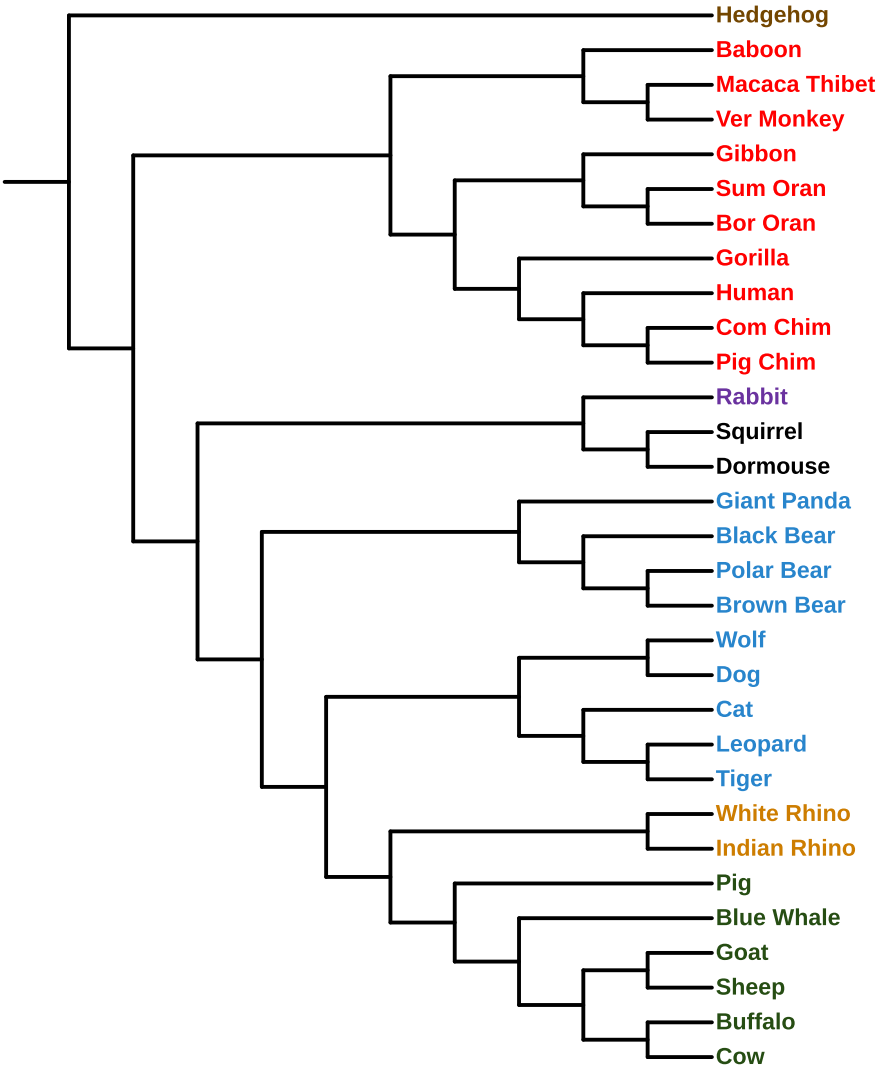


Figure 4: Phylogenetic tree of Mammals based on the GGRT method.

4.4. Human rhinovirus (HRV)

Human rhinovirus (HRV) is the most common viral infectious agent in humans, and is the main cause of the common cold [27]. Using multiple sequence alignment, Palmenberg et al. [27] correctly classified the complete HRV genomes into three genetically distinct groups within the genus *En-*

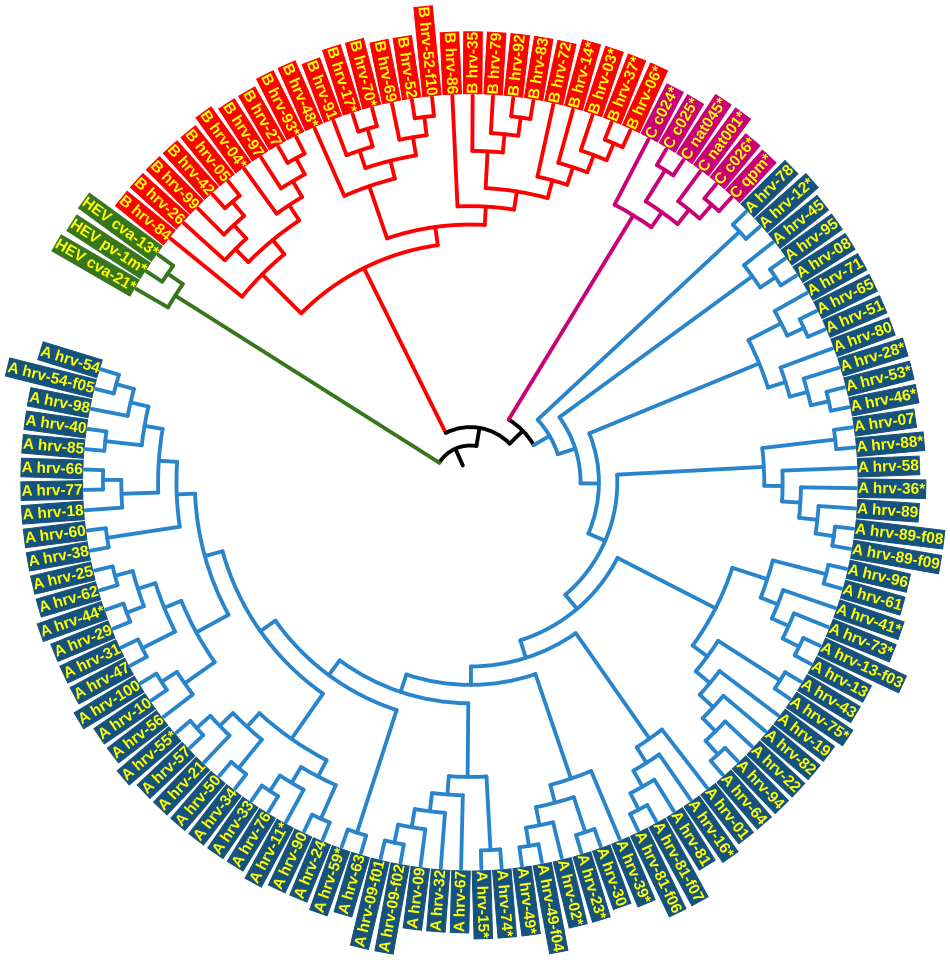


Figure 5: Phylogenetic tree of HRV based on the GGRT method.

terovirus (HEV) and the family *Picornaviridae*. The dataset used in [27] consists of three groups HRV-A, HRV-B, HRV-C including 113 genomes, and three outgroup sequences HEV-C. The GGRT method (see Tables 1 and 2 and Figure 5) for the phylogenetic tree of HRV genomes correctly classifies the complete HRV genomes into the corresponding genetically distinct groups in about 8.15 seconds. In terms of time complexity, the GGRT method is comparable to the Fourier transform method in [24], which performs the computation in 7 seconds.

5. Conclusions and discussions

In this paper, an alignment-free method for DNA similarity analysis, called the GGRT method is proposed, using a combination of graph theory-for representing DNA sequences as Cartesian products of graphs, and metric space structures on graphs, to view a given collection of DNA sequences as points in a metric space. In order to analyze similarities/dissimilarities, distances between such points are computed,which indicates similarities/dissimilarities between the corresponding DNA sequences. Throughout the paper, the edit distance on graphs is used due to its well-performed features in time complexity, accuracy as well as simplicity in computations. The GGRT method is tested on several standard datasets in literature, and compared with other available methods such as Fourier transform method [24] and Clustal Omega [23]. In some dataset, the GGRT performs the best in terms of time complexity and accuracy, and in several datsets, the GGRT method is comparable with Fourier transform method. In future work, we plan to improve accuracy and time complexity of the GGR method. Furthermore, we plan to study variants of the GGRT method, in combination with algebraic topology such as persistence diagrams from topological data analysis to propose more alignment-free methods for DNA similarity analysis.

Acknowledgements

Lizhen Lin and Dong Quan Ngoc Nguyen acknowledge the generous support of NSF grant DMS-2113642.

References

[1] X. Jin, R. Nie, D. Zhou, S. Yao, Y. Chen, J. Yu, and Q. Wang, A novel DNA sequence similarity calculation based on simplified pulse-coupled neural network and human coding. *Physica A: Statistical Mechanics and its Applications* **461** (2016), 325–338. MR3519878

[2] J. F. Yu, J. H. Wang, and X. Sun, Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation. *MATCH Commun. Math. Comput. Chem* **63** (2010), 493–512. MR2654804

[3] S. Wang, F. Tian, Y. Qiu, and X. Liu, Bilateral similarity function: A novel and universal method for similarity analysis of biological sequences. *Journal of Theoretical Biology* **265** (2010), 194–201. MR2981547

[4] N. Jafarzadeh and A. Iranmanesh, C-curve: a novel 3D graphical representation of DNA sequence based on codons. *Mathematical Biosciences* **241** (2013), 217–224. MR3019709

[5] B. Liao, Y. Zhang, K. Ding, and T. M. Wang, Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation. *Journal of Molecular Structure: THEOCHEM* **717** (2005), 199–203. MR2562816

[6] E. Hamori and J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *Journal of Biological Chemistry* **258** (1983), 1318–1327.

[7] B. Liao, M. Tan, and K. Ding, A 4D representation of DNA sequences and its application. *Chemical Physics Letters* **402** (2005), 380–383. MR2248778

[8] J. Wang and Y. Zhang, Characterization and similarity analysis of DNA sequences grounded on a 2D graphical representation. *Chemical Physics Letters* **423** (2006), 50–53.

[9] X. Q. Liu, Q. Dai, Z. Xiu, and T. Wang, PNN-curve: A new 2D graphical representation of DNA sequences and its application. *Journal of Theoretical Biology* **243** (2006), 555–561. MR2306345

[10] X. Q. Qi, J.Wen, and Z. H. Qi, New 3D graphical representation of DNA sequence based on dual nucleotides. *Journal of Theoretical Biology* **249** (2007), 681–690. MR2930186

[11] C. Li, X. Yu, and N. Helal, Similarity analysis of DNA sequences based on codon usage. *Chemical Physics Letters* **459** (2008), 172–174.

[12] N. Jafarzadeh and A. Iranmanesh, A novel graphical and numerical representation for analyzing DNA sequences based on codons. *Match-Communications in Mathematical and Computer Chemistry* **68** (2012), 611. MR3025639

[13] F. Kabli, H. R. Mohamed, and A. Abdelmalek, Similarity analysis of DNA sequences based on the chemical properties of nucleotide bases: frequency and position of group mutations. *Comput. Sci. Inf. Technol.* **6** (2016),1–10.

[14] Q. Dai, X. Liu, and T. Wang, A novel 2D graphical representation of DNA sequences and its application. *Journal of Molecular Graphics and Modelling* **25** (2006), 340–344.

[15] Y. H. Yao, X. Y. Nan, and T. M. Wang, A new 2D graphical representation–classification curve and the analysis of similarity/ dissimilarity of DNA sequences. *Journal of Molecular Structure: THEOCHEM* **764** (2006), 101–108.

[16] B. Liao, Q. Xiang, L. Cai, and Z. Cao, A new graphical coding of DNA sequence and its similarity calculation. *Physica A: Statistical Mechanics and its Applications* **392** (2013), 4663–4667. MR3083122

[17] P. A. He and J. Wang, Characteristic sequences for DNA primary sequence. *Journal of Chemical Information & Modeling* **42** (2002), 1080–1085.

[18] W. Hou, Q. Pan, and M. He, A novel representation of DNA sequence based on CMI coding. *Physica A* **409** (2014), 87–96. MR3213756

[19] R. Zhang and C. Zhang, Z curves, an intutive tool for visualizing and analyzing the DNA sequences. *Journal of Biomolecular Structure and Dynamics* **11** (1994), 767–782.

[20] R. Zhang and C. Zhang, A brief review: The z-curve theory and its application in genome analysis. *Curr Genomics* **15** (2014), 78–94.

[21] P. Wills and F. Meyer, Metrics for graph comparison: A practitioner's guide. *PLOS ONE* **15** (2020), 1–54

[22] J. A. Dieudonne, Foundations of modern analysis. *Pure and applied mathematics (Academic Press)*, New York: Academic Press. (1960). MR0120319

[23] F. Sievers, A. Wilm, D. Dineen, T. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, and e. a. J. Soding, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7** (2011), 539.

[24] T. Hoang, C. Yin, and S. S.-T. Yau, Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. *Genomics* **108** (2016), 134–142.

[25] M. Arbyn, X. Castellsague, S. D. Sanjose, L. Bruni, M. Saraiya, F. Bray, and J. Ferlay, Worldwide burden of cervical cancer in 2008. *Ann. Oncol.* **22** (2011), 2675–2686.

[26] J. Smith, L. Lindsay, B. Hoots, J. Keys, S. Franceschi, R. Winer, and G. Cliord, Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: a meta-analysis update. *Int. J. Cancer* **121** (2007), 621–632.

[27] A. C. Palmenberg, D. Spiro, R. Kuzmickas, S. Wang, A. Djikengand, J. A. Ratheand, C. M. Fraser-Liggett, and L. S. B, Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution. *Science (American Association for the Advancement of Science)* **324** (2009), 55–59.

[28] R. Garten, C. Davis, C. Russell, B. Shu, S. Lindstrom, A. Balish, W. Sessions, E. S. X. Xu, and e. a. V. Deyde, Antigenic and genetic characteristics of swine-origin 2009 A (H1N1) influenza viruses circulating in humans. *Science* **325** (2009), 197–201.

[29] P. Palese and J. Young, Variation of influenza A, B, and C viruses. *Science* **215** (1982), 1468–1474.

[30] C. Yu, M. Deng, and S. S.-T. Yau, DNA sequence comparison by a novel probabilistic method. *Information Sciences* **181** (2011), 1484–1492. MR2770365

[31] S. Kumar, G. Stecher, and K. Tamura, MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution* **33** (2016), 1870–1874.

[32] D. Alexander, A review of avian influenza in different bird species. *Vet. Microbiol.* **74** (2000), 3–13.

Dong Quan Ngoc Nguyen
Department of Applied and Computational
Mathematics and Statistics
University of Notre Dame
Notre Dame, IN 46556
USA
E-mail address: dongquan.ngoc.nguyen@nd.edu
URL: https://www3.nd.edu/~dnguyen15/

Lin Xing
Department of Applied and Computational
Mathematics and Statistics
University of Notre Dame
Notre Dame, IN 46556
USA
E-mail address: lxing@nd.edu

Phuong Dong Tan Le
Department of Applied Mathematics
University of Waterloo
Waterloo, Ontario
Canada N2L 3G1
E-mail address: pdle@uwaterloo.ca

Lizhen Lin
Department of Applied and Computational
Mathematics and Statistics
University of Notre Dame
Notre Dame, IN 46556
USA
E-mail address: lizhen.lin@nd.edu

Received February 21, 2022