scientific reports

OPEN

Comparative study of encoded and alignment-based methods for virus taxonomy classification

Muhammad Arslan Shaukat[1]✉, Thanh Thi Nguyen[2], Edbert B. Hsu[3], Samuel Yang[4] & Asim Bhatti[1]

The emergence of viruses and their variants has made virus taxonomy more important than ever before in controlling the spread of diseases. The creation of efficient treatments and cures that target particular virus properties can be aided by understanding virus taxonomy. Alignment-based methods are commonly used for this task, but are computationally expensive and time-consuming, especially when dealing with large datasets or when detecting new virus variants is time sensitive. An alternative approach, the encoded method, has been developed that does not require prior sequence alignment and provides faster results. However, each encoded method has its own claimed accuracy. Therefore, careful evaluation and comparison of the performance of different encoded methods are essential to identify the most accurate and reliable approach for virus taxonomy classification. This study aims to address this issue by providing a comprehensive and comparative analysis of the potential of encoded methods for virus classification and phylogenetics. We compared the vectors generated for each encoded method using distance metrics to determine their similarity to alignment-based methods. The results and their validation show that K-merNV followed by CgrDft encoded methods, perform similarly to state-of-the-art multi-sequence alignment methods. This is the first study to incorporate and compare encoded methods that will facilitate future research in making more informed decisions regarding selection of a suitable method for virus taxonomy.

COVID-19 infection primarily spreads through respiratory droplets and can cause symptoms ranging from mild such as fever and cough, to severe such as difficulty in breathing and pneumonia[1]. The global impact of COVID-19 cannot be overstated and continues to evolve as new virus variants emerge. Understanding virus taxonomy can aid in virus management by allowing researchers and healthcare providers to identify and track various types of viruses. For example, if a new virus is discovered, establishing its taxonomy might help scientists understand its traits and how it may behave.

Virus taxonomy can support the development of vaccines and treatments. Researchers can assess which types of vaccinations or therapies would be effective by identifying the family and species of a virus. This knowledge may potentially be used to direct the creation of novel therapies that focus on particular viral traits. By using computational techniques like alignment-based methods, it is possible to classify viruses according to their genomic sequences and deduce their evolutionary relationships.

Alignment-based methods for classifying genes rely on finding optimal alignments between sequences using scoring systems. They are often performed using software such as ClustalW or MUSCLE, which can align the sequences and calculate a score that reflects their similarity[2,3]. Once the sequences are aligned, a phylogenetic tree[4] can be constructed using various algorithms, such as the neighbor-joining method[5] or the maximum likelihood method[6]. The resulting tree reflects the evolutionary relationships between the organisms based on their genomic sequences. While accurate, these techniques are computationally expensive, which makes them unsuitable for assessing huge datasets[7]. To overcome these limitations, a variety of alignment-free techniques have been developed in the signal processing domain over the last two decades with promising performance[8].

Alignment-free methods do not employ sequence alignments to compare and classify sequences; instead, they extract features or patterns and use them to compare and classify sequences. Because of their efficiency, scalability, and ability to handle big datasets, alignment-free approaches have grown in popularity[9–11].

[1]Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Victoria, Australia. [2]Faculty of Information Technology, Monash University, Victoria, Australia. [3]Department of Emergency Medicine, Johns Hopkins University, Maryland, USA. [4]Department of Emergency Medicine, Stanford University, California, USA. ✉email: mshaukat@deakin.edu.au

There are several alignment-free (i.e., encoded) approaches available[7,11–14], each with its own stated accuracy. This study comprehensively reviews encoded methods that employ various approaches. We aim to investigate the behaviour of these encoded methods and determine how they address the uncertainty associated with generating reliable results. By identifying dependable and rapid techniques, we seek to highlight promising avenues for future research. Our work offers several important contributions to the field, including:

- Providing a first of its kind, comprehensive and comparative review of seventeen different encoded methods for ten different data-sets of varying lengths.
- Comparing the effectiveness of encoded methods with the well-established non-encoded methods. By doing so, we have identified the strengths and weaknesses of encoded methods and provided insights into their respective performances.
- Identifying the most reliable and fast encoded method, which could be used as an alternative to the computationally expensive alignment method.
- Publishing the datasets and codes used in this study online, to enable other researchers to replicate and build upon the findings of the study.
- Offering supplementary documents that provide phylogenetic trees and metrics results for each method and dataset. This information will be useful to researchers who want to delve further into the data to gain a more detailed understanding of the results.

Materials and methods

This study investigated a total of twenty different methods: four non-encoded multi-aligned methods and seventeen encoded methods. We employed ten different datasets using three separate software tools - Matlab, MEGA, and NGphylogeny (online). The similarity of encoded methods is compared to four state-of-the-art multi-sequence alignment methods. Two of these methods, ClustalW[3] and MUSCLE[15], were implemented on the software package MEGA 11[16]. The other two methods, MAFFT[17] and ClustalOmega[18], were implemented on an online tool called NGphylogeny[19].

This comparison allowed us to rank the encoded methods based on their similarity to the alignment method. It is worth noting that the choice of sequence alignment method may affect the comparison results. However, there are commonly used methods for multiple sequence alignment that are well-respected in the field. Additionally, use of two different software tools (MEGA11 and NGphylogeny) help to ensure that the results are robust and not influenced by a specific software implementation.

We used a funnel approach to evaluate each non-encoded method (Fig. 1). First, we generated distance matrices to record the distances between sequences for each dataset using both encoded and non-encoded methods. In order to generate an evolutionary distance matrix for a non-encoded method, it is essential to first align all the sequences in a given dataset. The Jukes-Cantor model[20] was used to construct the matrix for each alignment (i.e., non-encoded) method to ensure that the results obtained from these methods were comparable. When comparing different encoding methods, one potential issue is that some methods may produce matrices with

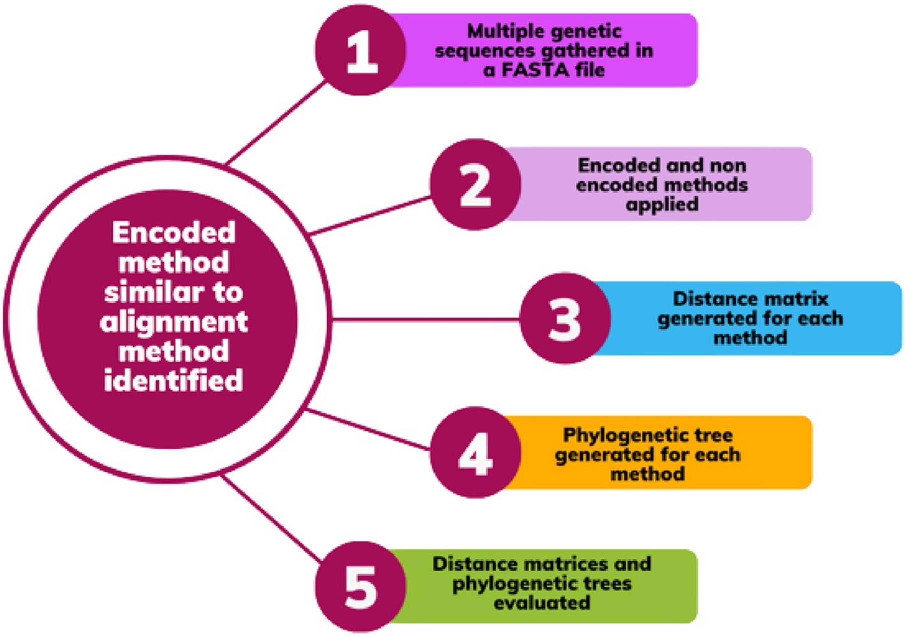

Encoded method similar to alignment method identified
1
Multiple genetic sequences gathered in a FASTA file
2
Encoded and non encoded methods applied
3
Distance matrix generated for each method
4
Phylogenetic tree generated for each method
5
Distance matrices and phylogenetic trees evaluated

Figure 1. Overview of the methodology used to compare encoded and multi-aligned (non-encoded methods). The process involved generating distance matrices, comparing them to rank the encoded methods, filtering out non-similar methods, and analysing phylogenetic trees to validate the results.

large values. In contrast, others may produce matrices with small values. We have normalized our matrix over a range of 0 and 1 where the minimum value is set to 0 and the maximum value is set to 1 since it is likely that the absolute value of the matrix does not reflect the differences between methods . By doing this, the matrix values are transformed into a consistent scale, which allows for fair and unbiased comparisons between different methods.

Next, we compare the distance matrix generated by each encoded method with the distance matrix generated by each non-encoded method using Euclidean distance. The method with the smallest Euclidean distance to the non-encoded method will considered to be the most similar. This comparison allowed us to rank the encoded methods based on their similarity to the alignment method. Finally, in order to validate the results obtained in the previous step, we utilised distance metrics at an arbitrary non-binary phylogenetic tree level. We then excluded any encoded methods that failed to meet the similarity criteria. To provide additional validation, we conducted a visual comparison of the phylogenetic trees. This enabled us to determine which encoded method had the least amount of difference from the multi-sequence alignment method. Through these steps, we were able to determine the effectiveness and accuracy of each encoded method and identify the most similar method to the non-encoded one.

Dataset

In this study, datasets were incorporated from previous studies[12,14,21], where virus genomes were collected from sources such as GenBank[22] and GISAID[23]. To prevent any potential biases from the datasets used in previous studies, a new dataset, Dataset0 is also included in this analysis. Dataset0 has not been previously used in any encoding techniques, ensuring a fair comparison between the different encoding methods being tested. The Table 1 shows the datasets used in the study.

Multi aligned (non-encoded) method

ClustalW

ClustalW uses a progressive alignment to align protein or nucleotide sequences. It involves constructing a guide tree based on pairwise distances between the sequences and then aligning the sequences based on the order of the tree.

ClustalOmega

Clustal Omega uses a combination of progressive and iterative methods to align protein or nucleotide sequences. It constructs a guide tree based on pairwise distances between sequences and then aligns the sequences based on the order of the tree. It then iteratively refines the alignment to improve its accuracy.

MUSCLE

MUSCLE (Multiple Sequence Comparison by Log-Expectation) constructs an initial alignment of the most similar sequences and then iteratively refines the alignment to incorporate additional sequences. It uses a progressive alignment method, which starts with a rough alignment of the most similar sequences and then adds in the remaining sequences one by one.

Name	Description	Total Seqs	Min. length	Max. length
DataSet0	Viruses in the genus AlphaCoV and BetaCoV of coronaviruses, along with their subgenera in BetaCov	59	27165	31526
DataSet1	Viruses from the family Coronaviridae to classify SARS-CoV-2	56	25425	31686
DataSet2	Viruses in the genus BetaCoV to classify SARS-CoV-2 at the genus level	50	29037	31491
DataSet3	Closely related coronaviruses from the seafood market	69	27213	30311
DataSet4	Transmission modes of human coronaviruses originating from animals	106	26883	31473
DataSet5	Virus genomes obtained from human SARS-CoV-2 viruses	141	29674	29882
DataSet6	Genus within the Coronaviridae family, known to induce a range of severe diseases in the respiratory and gastrointestinal systems	34	9646	31357
DataSet7	Influenza A viruses, which are single-stranded, segmented RNA viruses categorized according to their hemagglutinin and neuraminidase viral surface proteins	38	1350	1467
DataSet8	Human rhinoviruses, which is the most common cause of upper respiratory tract	116	6944	7458
DataSet9	HPV (Human Papillomavirus) is a common sexually transmitted DNA virus responsible for cervical cancer and genital warts	400	7814	10424

Table 1. Datasets from previous research are incorporated to evaluate the effectiveness of different encoding methods and their response to various parameters, such as the number of sequences and the maximum sequence length. To prevent any potential biases from the datasets used in previous studies, a new dataset, Dataset0 is also included in this analysis. Dataset0 has not been previously used in any encoding techniques, ensuring a fair comparison between the different encoding methods being tested.

MAFFT

MAFFT (Multiple Alignment using Fast Fourier Transform) uses a variety of algorithms to align sequences, including progressive pairwise alignment, iterative refinement, and consistency-based alignment. It also employs a fast Fourier transform algorithm to improve the accuracy of the alignment.

Encoded methods

Atomic number

The atomic number method of gene encoding assigns each nucleotide base in DNA a corresponding atomic number. The nucleotide bases are represented by A = 70, T = 66, C = 58, and G = 78, which correspond to the number of protons in the nucleus of an atom[24]. These numerical representations offer the ability to perform counting the occurrences of specific sequences within a larger sequence and comparing the similarity between two sequences[25]. This approach has been applied in the analysis of Rubisco protein genes, where a direct mapping using atomic numbers was employed to calculate sequence fluctuation[26].

Electron-ion interaction pseudopotential

The Electron-Ion Interaction Pseudopotential (EIIP) technique is a computational physics method employed to investigate the interactions between electrons and ions in materials. In this technique, specific values (C=0.1340, T=0.1335, A=0.1260, and G=0.0806) represent the relative frequency of the four nucleotides (C, T, A, and G) within a genetic sequence[27]. These values characterize the distribution of energies associated with the pseudopotentials of free electrons across the DNA sequence. The utilization of EIIP values has found practical applications in various domains, including neural networks, wavelet transform, and graph signal processing (GSP)[28]. By reflecting the pseudopotential feature of nucleotide sequences, EIIP values have proven to be valuable in fields such as bioinformatics, genomics, and molecular biology[29].

Molecular mass representation

Molecular mass quantifies the total mass of all atoms present in a molecule. To represent molecular mass numerically, the individual atoms within the molecule are assigned values based on their atomic masses, typically expressed in atomic mass units (amu). In the context of gene sequence encoding, each nucleotide (C, T, A, and G) is assigned a specific numerical code, 110, 125, 134, and 150, respectively[3031]. By employing these encoded nucleotide sequences, various mathematical techniques, such as clustering algorithms, can be applied to analyze and identify patterns or relationships between the sequences.

Frequency-of-occurrence

The fractional occurrence of nucleotides and their frequencies are key parameters used in various bioinformatics analyses of DNA sequences. It can be statistically calculated based on the frequencies of their occurrence in specific regions of the genome, such as exons and introns[32]. The encoding representation of each of the four nucleotides is cytosine = 0.27215, thymine = 0.20576, adenine = 0.24300, and guanine = 0.27909[33].

Pulse amplitude modulation

Pulse Amplitude Modulation (PAM) is a computational method employed in genomics to compare genomic sequence similarity. It is used to compare two genome sequences and quantify their differences in mutations. The specific real numbers assigned to denote the bases, A = -1.5, G = -0.5, T = 1.5, and C = 0.5, are arbitrary and were chosen to represent the differences between the bases in a way that is easily computable[34]. The real numbers used in the PAM scheme are not meant to represent the biological or chemical properties of the bases. Instead, they are meant to provide a convenient way to quantify differences between genomes. PAM has been utilized in comparative genomics, particularly in the analysis of native and synthetic enzymes[35]. It provides a quick and effective way to compare genomes, determine relatedness, and track genome evolution.

Fourier power spectrum

The Fourier power spectrum, represents the power or energy distribution of a signal in the frequency domain. The periodic patterns or repeated motifs that are present in the sequence can be identified using the Fourier Power Spectrum. Due to a variety of aspects, including repeating elements, coding regions, or structural characteristics, DNA sequences can display specific patterns or periodicities. This is achieved by breaking down the sequences into overlapping substrings and calculating the Fourier spectrum for each base[7].

Chaos game representation with discrete Fourier transform

Chaos Game Representation (CGR) is a graphical representation method used in genomics to visualize the structure of genomic sequences proposed by Jeffery[36]. The mathematical equations for Chaos Game Representation of a DNA sequence can be defined as follows:

$$Xo = (1/2, 1/2), Xn = 1/2(Xn - 1 + W), \tag{1}$$

where W is the coordinates of the corner of the unit square.

In the CGR method, a DNA sequence is divided into overlapping triplets of nucleotides (codons), and the positions of the codons are plotted on a two-dimensional plane according to their corresponding amino acids. This allows the DNA sequence structure to be visualized as a pattern on the plane. Regions of the sequence that code for similar amino acids appear as clusters.

The Discrete Fourier Transform (DFT) transforms signals from the time domain to the frequency domain and represents them as a sum of sinusoids of different frequencies. The DFT provides the coefficients of these sinusoids. In genome encoding, DFT can be employed to analyze and compare DNA sequences by transforming them from their original form into a frequency-based representation[37]. The DFT coefficients provide information about the distribution of different frequencies in the sequence. They can be used to generate a spectrogram, which visualizes DNA sequence frequencies.

DFT can also be used to detect motifs and patterns in DNA sequences by looking for peaks in the spectrogram that correspond to specific frequencies. This can provide insights into the underlying structure and functional regions of DNA sequences[38]. CgrDft is a hybrid method that uses Chaos Game Representation with Discrete Fourier Transform (CgrDft) to visualize DNA sequence structure by mapping them onto a two-dimensional plane[12].

Dinucleotide
Dinucleotide DNA encoding is a method of representing DNA sequences using a two-dimensional plot. In this method, sixteen different dinucleotides (two nucleotide pairs) are mapped to a unit circle, with each dinucleotide represented by a distinct position on the circle[39]. The sixteen dinucleotides are AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, and TT. Once the dinucleotides are mapped to the unit circle, neighboring nucleotides in the DNA sequence are encoded as points on a two-dimensional plot. It can also be used for machine learning applications such as predicting genomic features or classifying DNA sequences.

I Ching representation
In gene sequence encoding, the 64 codons that make up the genetic code can be expressed through binary codes. To map these binary codes to the 64 codons, one approach is to use the I Ching[40]. Each hexagram in the I Ching can be assigned its own four-digit binary code based on the arrangement of solid and broken lines. These binary codes represent the 64 codons. Codons are mapped to hexagrams whose binary codes correspond to their own binary representations. For example, the codon AUG, which codes for the amino acid methionine, can be expressed as 0100 in binary. This binary code corresponds to hexagram 23 in the I Ching composed of two solid lines at the bottom and third positions and four broken lines in the middle. Therefore, in this encoding scheme, AUG would be represented as hexagram 23.

Integer number
Integer number encoding is a method of representing data using integers instead of characters or symbols. In the context of DNA sequence encoding, integer number encoding involves mapping each nucleotide (C, T, A, or G) to a corresponding integer value (0, 1, 2, or 3)[41,42]. Integer number encoding can be useful in DNA sequence analysis, as it allows for efficient storage and manipulation of large datasets. It also enables the use of mathematical operations and algorithms designed for working with integers[43]. However, it is imperative to use a consistent encoding scheme to ensure that different systems and programs can interpret the data correctly[44].

Inter-nucleotide distance
In the inter-nucleotide encoding scheme the sequence is represented as a series of inter-nucleotide sequences rather than as a series of nucleotides themselves[45]. One way to encode the distance between pairs of nucleotides that are a fixed length apart in the sequence is with a fixed distance of k[46] i.e., i+k, i+2k,.., i+nk can be used to represent between nucelotides and S(i), S(i+k), can be encodes as k, k1 if the same nucelotides occur at each position. Inter-nucleotide distance encoding can be useful in some applications[47], as it can reduce the dimensionality of the data and facilitate certain types of analysis. However, it may not be appropriate for all types of gene sequence analysis, as it discards some information about the specific nucleotides in the sequence[48].

Minimum entropy
Minimum entropy refers to the minimum amount of information needed to represent the sequence or the level of compression achieved for the sequence[49]. The entropy of a gene sequence can be calculated using the formula:

$$(H(M) = -\sum p(x)logp(x)), \tag{2}$$

where p(x) is the probability of each nucleotide base in the sequence. It also measures the average number of bits required to represent each distance and can be used to compress the gene sequence in a way that preserves the order and relative distances of the nucleotides.

Thermodynamic encoding
Thermodynamic encoding of gene sequences is a method of encoding gene sequences based on enthalpy values. The idea is to use the thermodynamic stability of DNA's double helix structure to encode sequence information in a way that is more robust to errors and noise[50]. The DNA double helix structure is stable due to interactions between nucleotide bases. The base pairs A-T and G-C form hydrogen bonds that stabilize the double helix structure[51]. The enthalpy change of hydrogen bonding interactions between DNA strands can be calculated using thermodynamic principles. The enthalpy values of these interactions can be used to encode sequence information in a way that is more resistant to errors and mutations.

K-mer encoding

K-mer encoding simplifies long DNA sequences into smaller chunks. For example, sequence ATCGAT turned into pieces like AT, TC, CG, GA, and AT when K=2. These chunks serve as DNA fingerprints for various applications, such as database creation or genome assembly. The choice of K affects the trade-off between simplicity and complexity in the chunks. A smaller K yields more fragments that are simpler but may lack detail. A larger K gives fewer, more complex fragments that encapsulate more information about the original sequence.

Triplet encoding

Triplet encoding focuses on encoding DNA sequences using triplets of codons based on the genetic code's 64 possible codons. Unlike K-mer encoding, where the value of K can vary, triplet encoding fixes K at 3, so that the sequence is always read three nucleotides at a time. For example, a DNA sequence like ATCGAT would yield non-overlapping triplets ATC and CGA. Each of the 64 possible codon triplets is assigned a unique identifier. These identifiers are then organized in a specific order, such as alphabetical or numerical order[52]. A repeat function then maps each nucleotide to its corresponding triplet identifier. However, triplet encoding also has limitations. For instance, the method relies on a predetermined list of 64 triplets, which may not be comprehensive enough to cover all possible DNA sequences. Moreover, it is not highly robust to errors or mutations. Even minor changes in the DNA sequence can significantly alter the resulting triplet identifiers, making the encoding less reliable compared to other schemes.

K-mer natural vectors

K-mer natural vectors represent a DNA or RNA sequence that quantifies the composition of k-mers (short contiguous substrings of length k) in the sequence. It is designed to overcome the deficiencies of previous k-mer models and provide a one-to-one mapping between a virus genome and its k-mer natural vector. This representation encodes the sequence into a high-dimensional vector, where each dimension corresponds to the frequency of each k-mer in the sequence[14]. Therefore, while K-mer encoding is simple but limited in its applications, K-mer natural vector provides greater versatility.

Voss representation

Voss representation encodes gene sequences using a series of binary strings[53]. Developed by Jeffrey Voss in the 1990s, it is a way to represent genomic sequence complexity using a simple binary code. A DNA sequence must first be broken up into individual nucleotides to be represented with the Voss representation. Following that, each nucleotide is assigned to the corresponding binary string. Concatenating the resulting binary strings creates a single, lengthy binary string that represents the whole DNA sequence. The Voss representation approach to encoding gene sequences has the benefit of being straightforward and user-friendly. Compact and effective binary codes can express lengthy and complex DNA sequences. It has certain drawbacks, though, namely its sensitivity to errors and mutations in the DNA sequence.

Distance metrics

The similarity between encoded and non-encoded procedures is measured by distance metrics. A small difference between the encoded and non-encoded techniques suggests that they are similar, whereas a significant difference suggests that the encoded method does not capture the necessary characteristics of the non-encoded approach.

Euclidean distance

A genetic sequence can be thought of as a multidimensional vector where a different position in the sequence is represented by a unique dimension. The following equation is used to compute the Euclidean distance

$$d(a,b) = \sqrt{\sum (a[i] - b[i])^2}. \tag{3}$$

The equation calculates the square of the difference between each position in the two sequences. It sums these values, and then takes the square root of the result to get the final distance value. The Euclidean distance[54] is a popular metric for measuring evolutionary relationships and has been widely used in similar domains[55,56].

Tiples metric

The triple metric[57] is a way to evaluate the accuracy of a phylogenetic tree in reconstructing evolutionary relationships between a set of taxa. A phylogenetic tree depicts the evolutionary history of a taxa and phylogentic inferences uses molecular or morphological data to reconstruct the branching pattern of such a tree.. The triplets metric evaluates the ability of a phylogenetic tree to correctly group three taxa together in a branching pattern based on their pairwise distances[58].

Robinson-Foulds (RF) metric

The Robinson-Foulds[59,60] metric is a popular method for comparing topological differences between two phylogenetic trees. The RF distance is based on the number of bipartitions present in one tree but not in the other. A bipartion divides a set of taxas into two groups such that are the taxa in one group are more similar to each other than any other taxas of the other group. The RF distance between tree A and tree B is the sum of bipartions present in tree A and not tree B, divided by two. Division by two is necessary to ensure RF distance is a metric. This means that it satisfies symmetry, non-negativity, and triangle inequality.

Matching pair metric

A matching pair metric[61] is used to calculate the distance or similarity between two paired taxa in a phylogenetic tree. It can also be used to compare evolutionary relationships between two closely related taxas and to identify tree regions where different taxa groups are more related to each other.

Nodal splitted weighted distance metric

In Nodal splitted weighted[62], the tree topology and branch lengths are estimated by minimizing the sum of weighted distances between the observed sequences and the reconstructed sequences at internal nodes of the tree. Each node's weight is dependant on the number of sequences contained in it. The least squares method is employed to estimate branch lengths. In addition to be computionally efficient and scalable to massive datasets, it offers more accuracy than other distance based methods when number of sequences is large or with sequences with considerable evolutionary distances between them.

Matching cluster distance

Matching cluster distance[61] is a distance metric used in phylogenetic tree construction. MCD is based on pairwise distances between clusters of sequences rather than pairwise distances between individual sequences.

MAST

Maximum Agreement Subtree[63], or known as MAST is a distance metric used to compare phylogenetic trees, specifically the similarity between two trees. This metric measures the distance between two trees based on the size of their maximum agreement subtrees. A maximum agreement subtree is a subtree that appears in both trees and has the maximum number of nodes. The MAST distance is the difference between the total number of nodes in the two trees and twice the size of their maximum agreement subtree. In other words, it measures the number of nodes pruned from one tree to obtain the other.

Cophenetic L2

Cophenetic L2[64] distance is a metric used to compare the cophenetic distance matrices of two trees.

$$d = \sqrt{\sum \left(A[i][j] - B[i][j]\right)^2}, \tag{4}$$

where, A[i][j] and B[i][j] represents the points at the i-th row and j-th column, respectively. It is primarily used to assess the quality of different tree reconstruction methods.

Quartet distance

The quartet distance[65] metric is used in unrooted phylogenetic tree reconstruction to evaluate the similarity or dissimilarity between different trees. It is a distance metric that measures the difference between two trees based on the number of quartet trees they share. A quartet tree has four taxas and the branching pattern joins two taxa pairs. For instance a quartet tree with taxa W, X, Y, Z could have ((W,X), (Y,Z)) or ((W,Y), (X,Z)) branching patterns.. The quartet distance metric calculates the number of quartet trees shared between two trees.

Path difference

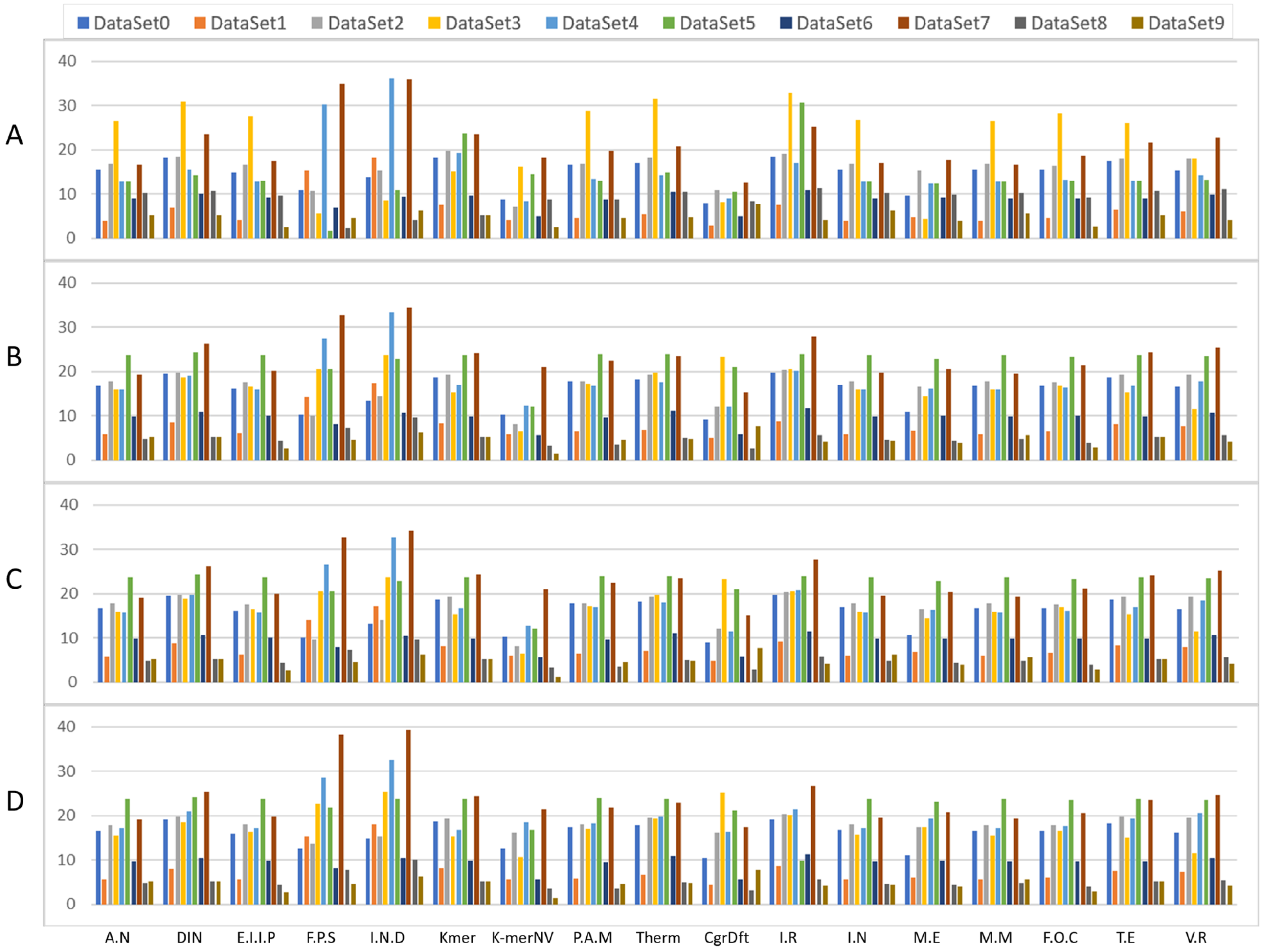
The path difference[66] metric assesses the robustness of branches in a phylogenetic tree. This is done by calculating the path difference between the original tree and a reference tree with a particular branch removed. If the path difference is small, the branch is considered well-supported and robust, while a larger path difference indicates a less robust branch.

Results and discussion

The encoded methods are put to the test using various datasets belonging to viruses like SARS-CoV-2 and influenza, which include both short and long genomes. To evaluate and compare the different genomic data, moment vectors were computed on each method. These vectors represent the statistical properties of the sequence data and are used to compare the similarity between different sequences. A matrix was created showing the pairwise distances between these vectors, which was used to cluster the data into biological groups to construct phylogenetic trees. Further process involves comparing the encoded and non-encoded methods to determine their similarities. This is achieved by creating a normalised matrix for each method and using the Euclidean distance for comparison. Accordingly, this comparison enables understanding of how the encoded and non-encoded methods relate to each other and identifies any similarities between them.

Figure 2 represents the Euclidean distance values for various encoding methods in comparison to non-encoded multi-sequence alignments generated by four different methods: ClustalW, ClustalOmega, MAAFT, and MUSCLE for ten different datasets (DataSet 0–9) where each bar represents the Euclidean distance between the distance matrices of the respective encoded and non-encoded method for a particular dataset.The minimum euclidean distance value for each dataset highlights the method that is most similar to the non-encoded method for that dataset. For each dataset, Chaos Game Representation with Discrete Fourier spectrum (CgrDft), K-mer natural vectors (K-merNV), Fourier Power Spectrum(FPS) were the encoding techniques that had the least Euclidean distances with non-encoded techniques used for comparison in this paper.

Further analysis on the performance of encoded method shows more than 80% of the time, CgrDft and K-merNV methods have similarity with popular non-encoded methods ClustalW (Fig. 2A),ClustalOmega (Fig. 2B),MAAFT (Fig. 2C) and MUSCLE (Fig. 2D). This indicates that for these datasets, the CgrDft and



Atomic Number = A.N, Dinucleotide = DIN, Electron-Ion Interaction Pseudopotential = E.I.I.P, Fourier Power Spectrum = F.P.S, Inter nucleotide Distance = I.N.D, k-mer Natural Vector = K-merNV, Pulse Amplitude Modulation = P.A.M, Thermodynamics = Therm, Chaos Game Representation with Discrete Fourier Transform = CgrDft, Iching Representation = I.R, Integer Number = I.N, Minimum Entropy = M.E, Molecular Mass = M.M, Single nucleotide frequency of occurrence = F.O.C, Triplet Encoding = T.E, Voss Representation = V.R

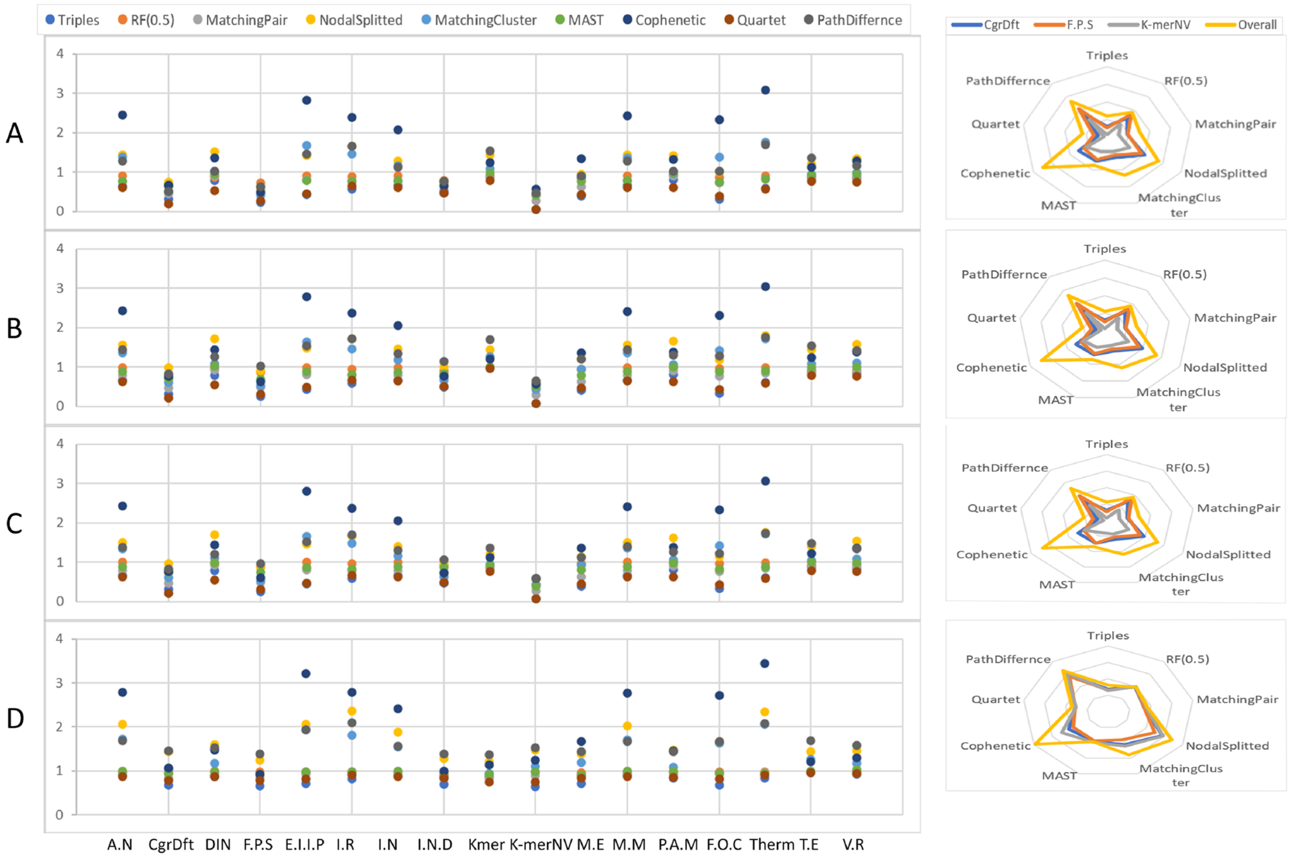
Figure 2. A comparison of Euclidean distances among distance matrices generated by encoding and non encoding techniques across ten distinct datasets (DataSet 0-9). Each bar represents the euclidean distance between distance matrices generated by encoded method (X-axis) and non-encoded multi-sequence alignment (**A**) ClustalW (**B**) ClustalOmega (**C**) MAAFT (**D**) MUSCLE methods.

K-merNV methods are the most effective encoding techniques among the methods listed. In other words, the CgrDft and K-merNV encoding methods have the highest similarity to the non-encoded methods, suggesting that they produce the most accurate results for these datasets. Thus, the CgrDft and K-merNV encoding techniques could be a better choice for encoding sequences when performing sequence analysis on these datasets.

To further validate these results, we applied the visual TreeCmp[67] package for distance metrics tests on arbitrary non-binary phylogenetic trees generated by each encoded method. Here, the reference trees to which each encoded method tree is compared are generated by non-encoded methods(ClustalW, ClustalOmega, MAAFT and MUSCLE). The marked points in Fig. 3 indicate that the values for all mentioned metrics are lower for CgrDft, K-merNV, and FPS encoding methods. Furthermore , Fig. 3 shows the radar chart for values for these methods and the overall average values of all methods. The overall average value is obtained by comparing the 68 values (i.e., 17 × 4) from all encoded techniques (a total of seventeen methods) to all non-encoded methods (a total of four methods). As it can be seen, the position of each distance metric is under the overall values for CgrDft, K-merNV, and FPS encoding methods which validates the results from Euclidean distance. To further validate these findings, methods with the lowest values, such as CgrDft, K-merNV, and FPS were visually compared.

Figures (Supplementary Figs. 1–4) depict phylogenetic trees (based on Dataset0) demonstrating the evolutionary relationships between different virus strains based on their genetic sequences. These trees are created using shortlisted encoded methods from the previous step i.e., K-merNV, CgrDft, FPS, and clustalW. The supplementary material contains phylogenetic tree files that include all methods, both encoded and non-encoded, across all datasets, which can be used to visually compare the performance of different methods.

These phylogenetic trees are a detailed and complete representation of the evolutionary relationships among viruses, including SARS-CoV-2, which is currently the most pathogenic coronavirus strain. It displays a larger number of branches, providing more information about the relationships between strains. The lengths of the



Atomic Number = A.N, Dinucleotide = DIN, Electron-Ion Interaction Pseudopotential = E.I.I.P, Fourier Power Spectrum = F.P.S, Inter nucleotide Distance = I.N.D, k-mer Natural Vector = K-merNV, Pulse Amplitude Modulation = P.A.M, Thermodynamics = Therm, Chaos Game Representation with Discrete Fourier Transform = CgrDft, Iching Representation = I.R, Integer Number = I.N, Minimum Entropy = M.E, Molecular Mass = M.M, Single nucleotide frequency of occurrence = F.O.C, Triplet Encoding = T.E, Voss Representation = V.R

Figure 3. A representation of metrics (Triples, RF(0.5) etc) applied to phylogenetic trees generated by encoding and non-encoding techniques across the dataset (DataSet 0). Each dot represents the metric value between phylogenetic trees generated by encoded method (X-axis) and non-encoded multi-sequence alignment (**A**) ClustalW (**B**) ClustalOmega (**C**) MAAFT (**D**) MUSCLE methods. Radar plot shows the position of distance metrics Triples, RF(0.5), MatchingPair, NodalSplitted, MatchingCluster, MAST, Cophenetic, Quartet, and PathDiffernce) for K-merNV, CgrDft, and FPS.

branches on the tree show how much evolution has occurred since the divergence from a common ancestor. The overall pattern of branching gives meaningful insights into relationships. A short branch does not necessarily signify a lesser connection, nor does a long one indicate a more substantial relationship. It is the way the branches connect that counts.

Comparing the different trees reveals that the K-merNV (Supplementary Fig. 2) and multi-sequence alignment methods (Supplementary Fig. 1) produce consistent and accurate results, reflecting the similar grouping between the virus sequences. In contrast, CgrDft (Supplementary Fig. 3) and FPS (Supplementary Fig. 4) methods show incorrect grouping of some strains. Specifically, CgrDft incorrectly grouped 'Rousettus bat CoV HKU9,' while FPS incorrectly grouped 'TGEV', 'Bat CoV RaTG13' and 'Mink CoV WD1127'. Therefore, the K-merNV phylogenetic tree appears to be the most comprehensive and accurate representation of multi-sequence alignment methods, followed by CgrDft and FPS.

While our study has shown promising results in understanding the genetic relationships among viruses, it mainly looked at a subset of viruses. However, these methods can also be applied to other viruses. For example, recent research has used alignment-free techniques to study phages which have diverse genetic material[68,69]. Moreover, our focus was on some particular techniques for analysing genetic sequences, however, there are newer methods[70–75] that use advanced computer algorithms to possibly get even more accurate results.

Conclusion

This study highlights the potential of encoded methods for classifying viruses and phylogenetics. While previous studies only compared the effectiveness of each method by comparing their phylogenetic trees, this paper compares the vectors generated by encoded and non-encoded methods using distance metrics to determine similarity between them. Through these comparisons, we evaluate how well the encoded methods perform in comparison to existing, widely used alignment methods. By comparing the results of the encoded methods to those of ClustalW and MUSCLE (implemented on MEGA 11) and MAFFT and ClustalOmega (implemented on NGphylogeny), we determine the K-merNV followed by CgrDft encoded methods are similar with the current state of the art-multi sequence alignment methods. To the best of our knowledge, this is a novel approach to incorporation and comparison for encoded methods. In the future, examining the behaviour of encoded methods when tested on other distance vectors, such as those generated using Kimura and Tamura models, might be of interest. Further, it would be interesting to see some advanced algorithms to improve these encoded methods. Another interesting avenue could be to compare alignment-based methods with the latest alignment-free methods to see which ones provide the most accurate results. It would also be insightful to explore the use of encoded methods in other areas of genomics research and compare their performance to existing methods in those domains. Furthermore, making these methods faster and more accurate could be a game-changer for getting important information more quickly.

Data availability

All data generated or analyzed during this study is available on GitHub, https://github.com/marslanshaukat/Encoded-and-Alignment-Based-Methods.git.

Received: 4 July 2023; Accepted: 19 October 2023
Published online: 31 October 2023

References

1. Whitaker, M. *et al.* Persistent Covid-19 symptoms in a community study of 606,434 people in England. *Nat. Commun.* **13**, 1957 (2022).
2. Edgar, R. C. Muscle: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
3. Thompson, J. D., Gibson, T. J. & Higgins, D. G. Multiple sequence alignment using clustalw and clustalx. Current protocols in Bioinformatics 2.3. 1–2.3. 22 (2003).
4. Kapli, P., Yang, Z. & Telford, M. J. Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* **21**, 428–444. https://doi.org/10.1038/s41576-020-0233-0 (2020).
5. Saitou, N. & Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
6. Yang, Z. Paml: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
7. Hoang, T. *et al.* A new method to cluster DNA sequences using Fourier power spectrum. *J. Theor. Biol.* **372**, 135–145 (2015).
8. Vinga, S. & Almeida, J. Alignment-free sequence comparison-a review. *Bioinformatics* **19**, 513–523 (2003).
9. Jing, X., Dong, Q., Hong, D. & Lu, R. Amino acid encoding methods for protein sequences: A comprehensive review and assessment. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **17**, 1918–1931 (2019).
10. Yu, C. *et al.* Real time classification of viruses in 12 dimensions. *PLoS ONE* **8**, e64328 (2013).
11. Yu, N., Li, Z. & Yu, Z. Survey on encoding schemes for genomic data representation and feature learning-from signal processing to machine learning. *Big Data Min. Anal.* **1**, 191–210 (2018).
12. Hoang, T., Yin, C. & Yau, S.S.-T. Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. *Genomics* **108**, 134–142 (2016).
13. Wen, J., Chan, R. H., Yau, S.-C., He, R. L. & Yau, S. S. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene* **546**, 25–34 (2014).
14. Zhang, Y., Wen, J., Li, X. & Li, G. Exploration of hosts and transmission traits for SARS-CoV-2 based on the k-mer natural vector. *Infect. Genet. Evol.* **93**, 104933 (2021).
15. Edgar, R. C. Muscle: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **5**, 113. https://doi.org/10.1186/1471-2105-5-113 (2004).
16. Tamura, K., Stecher, G. & Kumar, S. Mega11: Molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* **38**, 3022–3027. https://doi.org/10.1093/molbev/msab120 (2021).
17. Katoh, K. & Standley, D. M. Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780. https://doi.org/10.1093/molbev/mst010 (2013).
18. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.* **7**, 539 (2011).
19. Lemoine, F. *et al.* Ngphylogeny.fr: New generation phylogenetic services for non-specialists. *Nucleic Acids Res.* **47**, W260–W265. https://doi.org/10.1093/nar/gkz303 (2019).
20. Jukes, T. H. & Cantor, C. R. Evolution of protein molecules. *Mamm. Protein Metab.* **3**, 21–132 (1969).
21. Nguyen, T. T. *et al.* Origin of novel coronavirus causing covid-19: A computational biology study using artificial intelligence. *Mach. Learn. Appl.* **9**, 100328 (2022).
22. Benson, D. A. *et al.* Genbank. *Nucleic Acids Res.* **41**, D36–D42 (2012).
23. Shu, Y. & McCauley, J. Gisaid: Global initiative on sharing all influenza data-from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
24. Holden, T. *et al.* Atcg nucleotide fluctuation of deinococcus radiodurans radiation genes. In *Instruments, Methods, and Missions for Astrobiology X* Vol. 6694 (ed. Holden, T.) 402–411 (SPIE, 2007).
25. Holden, T. *et al.* Nucleotide fluctuation of radiation-resistant halobacterium sp. NCR-1 single-stranded DNA-binding protein (RPA) genes. In *Instruments and Methods for Astrobiology and Planetary Missions XII* Vol. 7441 (ed. Holden, T.) 259–271 (SPIE, 2009).
26. Holden, T. *et al.* Diverse nucleotide compositions and sequence fluctuation in rubisco protein genes. In *Instruments, Methods, and Missions for Astrobiology XIV* Vol. 8152 (ed. Holden, T.) 215–225 (SPIE, 2011).
27. Nair, A. S. & Sreenadhan, S. P. A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). *Bioinformation* **1**, 197 (2006).
28. Mabrouk, M. S. A study of the potential of EIIP mapping method in exon prediction using the frequency domain techniques. *Am. J. Biomed. Eng.* **2**, 17–22 (2012).
29. Adetiba, E., Olugbara, O. O. & Taiwo, T. B. Identification of pathogenic viruses using genomic cepstral coefficients with radial basis function neural network: Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing (NaBIC2015) in Pietermaritzburg, South Africa, held December 01-03, 2015. In *Advances in Nature and Biologically Inspired Computing* (eds Adetiba, E. *et al.*) 281–291 (Springer, 2015).
30. Stanley, H. *et al.* Statistical mechanics in biology: How ubiquitous are long-range correlations?. *Physica A* **205**, 214–253 (1994).
31. Li, W. & Kaneko, K. Long-range correlation and partial 1/fa spectrum in a noncoding DNA sequence. *Europhys. Lett.* **17**, 655 (1992).

32. Akhtar, M., Epps, J. & Ambikairajah, E. On DNA numerical representations for period-3 based exon prediction. In *IEEE International Workshop on Genomic Signal Processing and Statistics* (eds Akhtar, M. *et al.*) 1–4 (IEEE, 2007).
33. Mabrouk, M. Advanced genomic signal processing methods in DNA mapping schemes for gene prediction using digital filters. *Am. J. Signal Process.* **7**, 12–24 (2017).
34. Rosen, G. L. & Moore, J. D. Investigation of coding structure in dna. In IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)., vol. 2, II–361 (IEEE, 2003).
35. Perri, K. A., Manning, S. R., Watson, S. B., Fowler, N. L. & Boyer, G. L. Dark adaptation and ability of pulse-amplitude modulated (pam) fluorometry to identify nutrient limitation in the bloom-forming cyanobacterium, microcystis aeruginosa (kützing). *J. Photochem. Photobiol. B* **219**, 112186 (2021).
36. Jeffrey, H. J. Chaos game representation of gene structure. *Nucleic Acids Res.* **18**, 2163–2170 (1990).
37. Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S. & Ramaswamy, R. Prediction of probable genes by Fourier analysis of genomic sequences. *Bioinformatics* **13**, 263–270 (1997).
38. Fukushima, A. *et al.* Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis. *Gene* **300**, 203–211 (2002).
39. Liu, Z., Liao, B., Zhu, W. & Huang, G. A 2d graphical representation of DNA sequence based on dual nucleotides and its application. *Int. J. Quantum Chem.* **109**, 948–958 (2009).
40. Castro-Chavez, F. Defragged binary i ching genetic code chromosomes compared to nirenberg's and transformed into rotating 2d circles and squares and into a 3d 100% symmetrical tetrahedron coupled to a functional one to discern start from non-start methionines through a stella octangula. Journal of proteome science and computational biology **2012** (2012).
41. Cristea, P. D. Genetic signal representation and analysis. In *Functional Monitoring and Drug-Tissue Interaction* Vol. 4623 (ed. Cristea, P. D.) 77–84 (SPIE, 2002).
42. Hebert, P. D., Cywinska, A., Ball, S. L. & DeWaard, J. R. Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **270**, 313–321 (2003).
43. Ratnasingham, S. & Hebert, P. D. Bold: The barcode of life data system. *Mol. Ecol. Notes* **7**, 355–364 (2007).
44. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
45. Nair, A. S. S. & Mahalakshmi, T. Visualization of genomic data using inter-nucleotide distance signals. Proceedings of IEEE Genomic Signal Processing **408** (2005).
46. Afreixo, V., Bastos, C. A., Pinho, A. J., Garcia, S. P. & Ferreira, P. J. Genome analysis with inter-nucleotide distances. *Bioinformatics* **25**, 3064–3070 (2009).
47. Zhou, L.-Q., Li, R. & Han, G.-S. A method based on the improved inter-nucleotide distances of genomes to construct vertebrates phylogeny tree. In 7th International Conference on Biomedical Engineering and Informatics, 776–780 (IEEE, 2014).
48. Bastos, C. A. *et al.* Inter-dinucleotide distances in the human genome: An analysis of the whole-genome and protein-coding distributions. *J. Integr. Bioinform.* **8**, 31–42 (2011).
49. Galleani, L. & Garello, R. The minimum entropy mapping spectrum of a DNA sequence. *IEEE Trans. Inf. Theory* **56**, 771–783 (2010).
50. Breslauer, K. J., Frank, R., Blöcker, H. & Marky, L. A. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci.* **83**, 3746–3750 (1986).
51. Yu, N., Guo, X., Gu, F. & Pan, Y. Dna as x: An information-coding-based model to improve the sensitivity in comparative gene analysis. In Bioinformatics Research and Applications: 11th International Symposium, ISBRA 2015 Norfolk, USA, June 7-10, 2015 Proceedings 11, 366–377 (Springer, 2015).
52. Zou, S., Wang, L. & Wang, J. A 2d graphical representation of the sequences of DNA based on triplets and its application. *EURASIP J. Bioinf. Syst. Biol.* **2014**, 1–7 (2014).
53. Voss, R. F. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.* **68**, 3805 (1992).
54. Danielsson, P.-E. Euclidean distance mapping. *Comput. Graph. Image Process.* **14**, 227–248 (1980).
55. Singh, M. K., Singh, N. & Singh, A. Speaker's voice characteristics and similarity measurement using Euclidean distances. In *International Conference on Signal Processing and Communication (ICSC)* (eds Singh, M. K. *et al.*) 317–322 (IEEE, 2019).
56. Tantardini, M., Ieva, F., Tajoli, L. & Piccardi, C. Comparing methods for comparing networks. *Sci. Rep.* **9**, 1–19 (2019).
57. Critchlow, D. E., Pearl, D. K. & Qian, C. The triples distance for rooted bifurcating phylogenetic trees. *Syst. Biol.* **45**, 323–334 (1996).
58. Kuhner, M. K. & Yamato, J. Practical performance of tree comparison metrics. *Syst. Biol.* **64**, 205–214 (2015).
59. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
60. Robinson, D. F. & Foulds, L. R. Comparison of weighted labelled trees. In *Combinatorial Mathematics VI: Proceedings of the Sixth Australian Conference on Combinatorial Mathematics, Armidale, Australia* (eds Robinson, D. F. & Foulds, L. R.) 119–126 (Springer, 1978).
61. Bogdanowicz, D. & Giaro, K. On a matching distance between rooted phylogenetic trees. *Int. J. Appl. Math. Comput. Sci.* **23**, 669–684 (2013).
62. Cardona, G., Llabrés, M., Rosselló, F. & Valiente, G. Nodal distances for rooted phylogenetic trees. *J. Math. Biol.* **61**, 253–276 (2010).
63. Farach, M., Przytycka, T. M. & Thorup, M. On the agreement of many trees. *Inf. Process. Lett.* **55**, 297–301 (1995).
64. Cardona, G., Mir, A., Rosselló, F., Rotger, L. & Sánchez, D. Cophenetic metrics for phylogenetic trees, after Sokal and Rohlf. *BMC Bioinform.* **14**, 1–13 (2013).
65. Estabrook, G. Report on eighteenth international numerical taxonomy conference. *Syst. Biol.* **34**, 100–101 (1985).
66. Steel, M. A. & Penny, D. Distributions of tree comparison metrics-some new results. *Syst. Biol.* **42**, 126–141 (1993).
67. Goluch, T., Bogdanowicz, D. & Giaro, K. Visual treecmp: Comprehensive comparison of phylogenetic trees on the web. *Methods Ecol. Evol.* **11**, 494–499 (2020).
68. Song, K. Classifying the lifestyle of metagenomically-derived phages sequences using alignment-free methods. *Front. Microbiol.* **11**, 567769. https://doi.org/10.3389/fmicb.2020.567769 (2020).
69. Bernard, G., Stephens, T. G., González-Pech, R. A. & Chan, C. X. Inferring phylogenomic relationship of microbes using scalable alignment-free methods. *Methods Mol. Biol.* **2242**, 69–76. https://doi.org/10.1007/978-1-0716-1099-2_5 (2021).
70. Ren, R., Yin, C. & S, S. T. Y.,. kmer2vec: A novel method for comparing DNA sequences by word2vec embedding. *J. Comput. Biol.* **29**, 1001–1021. https://doi.org/10.1089/cmb.2021.0536 (2022).
71. Tang, R., Yu, Z. & Li, J. Kinn: An alignment-free accurate phylogeny reconstruction method based on inner distance distributions of k-mer pairs in biological sequences. *Mol. Phylogenet. Evol.* **179**, 107662 (2023).
72. Pei, S., Dong, R., He, R. L. & Yau, S.S.-T. Large-scale genome comparison based on cumulative Fourier power and phase spectra: Central moment and covariance vector. *Comput. Struct. Biotechnol. J.* **17**, 982–994 (2019).
73. Dong, R., He, L., He, R. L. & Yau, S.S.-T. A novel approach to clustering genome sequences using inter-nucleotide covariance. *Front. Genet.* **10**, 234 (2019).
74. Ali, S. et al. A k-mer based approach for sars-cov-2 variant identification. In Bioinformatics Research and Applications: 17th International Symposium, ISBRA 2021, Shenzhen, China, November 26–28, 2021, Proceedings 17, 153–164 (Springer).
75. Kirk, J. M. *et al.* Functional classification of long non-coding rnas by k-mer content. *Nat. Genet.* **50**, 1474–1482 (2018).

Author contributions

Conceptualization, M.A.S., A.B., T.T.N., E.B.H., S.Y.; methodology M.A.S., A.B., T,T.N.; software, M.A.S.; validation, M.A.S.; formal analysis, M.A.S., T,T.N., A.B.; investigation, M.A.S., T.T.N.; resources, T,T.N., A.B., M.A.S.; data creation, M.A.S.; writing–original draft preparation, M.A.S., A.B., T,T.N., E.B.H., S.Y. All authors have read, reviewed and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-45461-0.

Correspondence and requests for materials should be addressed to M.A.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.