NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., 2025; 14(2), 516-531

Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences

Araştırma makalesi / Research article

www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

Using transfer learning models for DNA sequence similarity via fCGR method

fCGR yöntemi ile DNA dizi benzerliği için transfer öğrenme modellerinin kullanılması

Emre Delibaş[1,*]

[1] *Sivas Cumhuriyet Üniversitesi, Bilgisayar Mühendisliği Bölümü, 58140, Sivas Türkiye*

Abstract

Similarity analysis of DNA sequences is a critical issue for understanding evolutionary relationships and identifying genetic mutations. Since traditional alignment-based methods have high computational costs, this study investigated the applicability of transfer learning models for alignment-independent DNA similarity analysis. DNA sequences were visualized with the Frequency Chaos Game Representation (fCGR) method and feature extraction was performed with ResNet50, EfficientNetB0, and MobileNet models. Three similarity metrics such as cosine similarity, Euclidean distance, and correlation and four different hierarchical clustering methods were compared. The results show that cosine similarity metric reflects genetic similarities better. MobileNet provided the highest accuracy rate with its lightweight structure and efficient feature extraction. Feature vectors visualized with PCA exhibited strong clustering tendencies and were in agreement with reference phylogenetic trees. The study demonstrates the applicability of transfer learning in genetic analyses and shows that scalable and biologically meaningful analyses can be performed.

Keywords: DNA sequence similarity, Transfer learning, Automatic feature extraction, fCGR

Öz

DNA dizilerinin benzerlik analizi, evrimsel ilişkilerin anlaşılması ve genetik mutasyonların belirlenmesi açısından kritik bir konudur. Geleneksel hizalama tabanlı yöntemler yüksek hesaplama maliyetine sahip olduğundan, bu çalışmada hizalamadan bağımsız DNA benzerlik analizi için transfer öğrenme modellerinin uygulanabilirliği incelenmiştir. DNA dizileri, Frequency Chaos Game Representation (fCGR) yöntemiyle görselleştirilmiş ve ResNet50, EfficientNetB0, MobileNet modelleriyle özellik çıkarımı yapılmıştır. Cosine similarity, Euclidean distance ve correlation gibi üç benzerlik metriği ve dört farklı hiyerarşik kümeleme yöntemi karşılaştırılmıştır. Sonuçlar, cosine similarity metriğinin genetik benzerlikleri daha iyi yansıttığını göstermektedir. MobileNet, hafif yapısı ve verimli özellik çıkarımıyla en yüksek doğruluk oranını sunmuştur. PCA ile görselleştirilen özellik vektörleri güçlü kümelenme eğilimleri sergilemiş ve referans filogenetik ağaçlarla uyum göstermiştir. Çalışma, transfer öğrenmenin genetik analizlerde uygulanabilirliğini ortaya koyarak ölçeklenebilir ve biyolojik olarak anlamlı analizler yapılabileceğini göstermektedir.

Anahtar Kelimeler: DNA dizi benzerliği, Transfer öğrenme, Otomatik özellik çıkarımı, fCGR

1 Introduction

The analysis of DNA sequence similarity represents a foundational challenge within a plethora of bioinformatics applications, including those pertinent to evolutionary biology, the investigation of genetic diseases, and the identification of novel genes [1]. Alignment-based methods, including the BLAST algorithm and the Needleman-Wunsch algorithm, have been instrumental in this field [2,3]. However, these methods have high computational costs when working with large datasets or long sequences. This limitation has led to the development of alignment-free methods to analyse similarities between sequences [4]. These alignment-free methods have garnered significant attention, particularly due to their capacity to expedite the processing of voluminous genomic data. These approaches have been shown to overcome the computational difficulties of alignment-based methods by using various strategies, such as k-mer frequency analysis, chaos game representation (CGR) and information theory-based metrics [4]. Frequency Chaos Game Representation (fCGR) is a particularly promising technique that facilitates rapid and efficient similarity analysis by converting DNA sequences into a format that is both visualizable and analyzable [5].

While alignment-free methods have made significant contributions, deep learning has revolutionized the field of bioinformatics [6]. Transfer learning is a subfield of deep learning that allows models pre-trained on large datasets to extract meaningful features from new datasets. Such approaches have shown great success in DNA sequence classification problems where labeled datasets are available [7]. However, clustering-based problems, such as the one in this study, present unique challenges due to the lack of labeled data.

The necessity for alignment-free methods has arisen in order to overcome the computational costs of alignment-

* Sorumlu yazar / Corresponding author, e-posta / e-mail edelibas@cumhuriyet.edu.tr (E. Delibaş)
Geliş / Received: 29.10.2024 Kabul / Accepted: 05.02.2025 Yayımlanma / Published: 15.04.2025
doi: 10.28948/ngumuh.1575701

based approaches, particularly in the context of large-scale genomic data. These methods provide bioinformatics workflows that are both fast and scalable by analysing sequence similarities without the necessity of direct alignment [8, 9]. However, the integration of alignment-free techniques with deep learning methods has not been sufficiently explored, especially for clustering-based problems. Deep learning methods have predominantly focused on supervised learning scenarios, where the training process is guided by labelled data [10]. In scenarios where such labelled data is not available, transfer learning emerges as a viable alternative. The present work aims to establish a bridge that combines alignment-free methods and deep learning in DNA sequence similarity analysis by investigating the usability of pre-trained models such as ResNet50, EfficientNet, and MobileNet for feature extraction.

Within the scope of alignment-free sequence analysis, many different fields of study have been included in this subject based on similarity calculation. Alignment-free methods have been studied in a wide range and several basic groups of methods have emerged in the literature: graphical representation-based approaches, image/texture based similarity analysis methods, information-theory-based algorithms, k-mer-based analyses and methods based on chaos theory. Each of them aims to provide fast and efficient similarity analysis by using various features of DNA sequences [11, 12].

Comparison of DNA sequences with different encoding methods based on graphical representation offers a comparison approach based on matching multidimensional curves on the coordinate plane [13]. These approaches have been developed based on 2D graphical representation [13], 3D graphical representation [14, 15] and larger dimensional representations [16-18]. On the other hand, examples of analyzing DNA sequences by converting them to different data formats are not limited to this. DNA sequences converted to image format with different coding techniques were vectorized with texture comparison methods and subjected to similarity calculations [19, 20]. When we look at the methods based on information theory, the amount of information shared between the analyzed sequences is calculated and explained. While the calculations are made with complexity calculations such as Kolmogrov or Lempel-Ziv [21, 22], it has been seen that methods based on vectorizing the transition distances of the nucleotides in the sequence with different metrics are also used [23]. DNA sequences have been evaluated as digital signal sequences in some studies and have found a wide field of study from the signal processing perspective [24, 25]. Some of the above-mentioned studies have tried to use the information they may contain by grouping them as dinucleotides, trinucleotides or by their physicochemical properties, as in [26] and [27].

Biomolecular sequences consisting of texts with a four-letter alphabet have also been an important area of study for word-based alignment-free algorithms [4, 8, 11]. K-mer-based analyses are one of the most common alignment-free methods. This method allows comparing similarities between sequences using the frequency of fixed-length subsequences (k-mers) in DNA sequences. This approach has been widely adopted due to its simplicity and computationally efficient nature [8]. In addition, a similar approach, n-gram-based methods, are among the word-based methods that bring different coding approaches in phylogenetic analyses [28, 29]. The use of statistical calculations of the obtained words [30], the use of Markov models found with k-mers or a combination of these two [31], similarity analysis with Cavnar and Trenkle distance calculation based on reduced n-grams [32], genome analysis combining the information theory approach [33] are examples of studies carried out in this field. Additionally, different approaches have been proposed by hybridizing CGR and word-based algorithms and the ranged word matches approach with maximum likelihood-based algorithms [34, 35].

Chaos game representation (CGR) is represented as a square matrix of nucleotide counts of a genome sequence [36, 37]. This technique has been used in various applications. CGR is also used in protein classification [38], Safoury et al. [37] proposed a DNA sequence classification method using convolutional neural network (CNN) from CGR image. Similarly, Rizzo et al. [39] developed a DNA sequence classification based on CGR image. In these studies, CGR has been shown to be an efficient method for rapid comparison of DNA sequences. However, a frequency-based extension of CGR, Frequency Chaos Game Representation (fCGR), provides a more detailed analysis by also taking into account k-mer frequencies in the sequences [5]. fCGR has been demonstrated to be efficacious in a variety of bioinformatics applications, including genome analysis, the inference of phylogenetic relationships, and protein classification. To illustrate this, consider the example of fCGR, a frequency-based extension of CGR, which facilitates expeditious calculation of similarities between sequences with k-mer representation [40].

The utilisation of transfer learning methodologies confers considerable advantages in the domains of DNA sequence analysis and classification. The utilisation of transfer learning facilitates the reutilisation of pre-trained models in novel tasks, thereby ensuring the attainment of superior outcomes with limited datasets [41, 42]. Transfer learning applications are critical to improve the overall performance of the model, especially when working with small numbers of examples. For instance, Bredesen and Rehmsmeier demonstrated that DNA sequence models enhance their capacity to generalise to independent Polycomb response elements [43]. Such studies underscore the impact of transfer learning on DNA sequence analysis. The potential of transfer learning is particularly evident in supervised learning tasks. Models such as ResNet50, EfficientNetB0 and MobileNet, pre-trained on extensive datasets like ImageNet, have been adapted for DNA sequence classification, where their weights are fine-tuned. However, the applicability of these models in clustering scenarios where fine-tuning is not possible due to the lack of labelled data has not yet been sufficiently investigated.

This study proposes a novel methodology for DNA sequence analysis. The objective of this study is to combine

alignment-free methods and transfer learning techniques in order to vectorize DNA sequences and evaluate the clustering performance of these vectors. The study utilises Frequency Chaos Games Representation (fCGR) as a fundamental technique to derive structured data from DNA sequences. This method has been shown to significantly reduce data size while simultaneously providing a high level of detail in the representation of DNA features. Transfer learning models ResNet50, EfficientNet and MobileNet are used for feature extraction on fCGR outputs. The classification layers of these models are then extracted, resulting in high-quality feature vectors of DNA sequences. Subsequently, an array of similarity measures, including cosine similarity, Euclidean distance, and correlation, are evaluated on these vectors. In addition, a range of clustering methods, namely average linkage, Ward's method, single linkage, and complete linkage, are employed to analyse the data. To assess the efficacy of the proposed methodology, a comparison with reference phylogenetic trees obtained from the MEGA11 program, a widely utilised reference tool in the domain of computational biology, was undertaken. The novelty of this work lies in the introduction of a transfer learning-based feature extraction approach for the analysis of DNA sequences using alignment-free methods, with the objective of optimising clustering performance and establishing a new framework for bioinformatics applications.

2 Material and Methods

2.1 Dataset

In this study, three distinct data sets were utilised to illustrate the generalisability of the proposed methodology on DNA sequences of varying lengths and complexity levels. The selection of data sets was guided by the following criteria:

- Widely Used in the Literature: These data sets are frequently used in the scientific literature and are accepted in the comparison of methods. Consequently, the findings of this study can be meaningfully compared to existing research in the scientific literature.
- Current and Scientific Importance: Data sets such as Influenza and Coronavirus indicate important biological problems that are frequently updated and always valid in the scientific world. These data sets are frequently used as data sources, especially in genetic analyses.
- Generalizability: Datasets of different lengths were chosen to test the generalizability of the proposed methodology and provide an opportunity to evaluate whether the method is effective on various types of DNA sequences.

The DNA sequences utilised in this study were obtained directly from the NCBI GenBank database, with accession numbers assigned for each sequence. These sequences pertain to specific gene regions and have been extensively cited in the extant literature, thus demonstrating their reliability. The obtained sequences are presented in FASTA format, in accordance with the standards established by GenBank, and are devoid of any missing data or low-quality regions. As the data set belongs to a specific gene region, there is no requirement for additional alignment or preprocessing, such as anomaly detection. Consequently, the raw data was utilised directly in the analysis process, a method that is widely accepted in the literature.

Summary information about the datasets detailed below is presented in Table 1.

Table 1. Datasets used in the analysis

Data Set	Source	Length (bp)	Number of Spec.
Influenza A	NCBI	1350-1467	38
mtDNAs	NCBI	16295-17019	18
Coronavirus	NCBI	9000-31000	36

2.1.1 Influenza A virus dataset

A DNA dataset of Influenza A viruses was used to test the proposed method. This dataset has been used in the literature to test many methods and consists of 38 viruses including H1N1, H2N2, H5N1, H7N3 and H7N9 subtypes [44- 47]. NCBI accession numbers of sequences with lengths ranging from 1350 to 1467 bases are given in Table S1.

2.1.2 Whole mitochondrial genomes of 18 eutherian mammals

Whole mitochondrial genomes contain a wealth of genetic information from 18 eutherian mammals and have been widely used in recent years [19, 23, 48, 49]. All sequences were obtained from the NCBI database and are listed in Table S2, ranging in length from 16,295 to 17,019 bases.

2.1.3 Coronavirus dataset

Coronavirus belongs to the Coronavirinae subfamily of the Coronaviridae family in the Nidovirales order. , In this article, a dataset containing 36 Coronaviruses was also used. The genomic size of coronaviruses, which is widely used in the literature, varies between approximately 9 thousand and 31 thousand bp and has an average of 27,567 nucleotides [50]. The dataset containing viruses of different genome lengths is considered important in terms of generalizability in the analysis of the method. Details of the genome information of the viruses and access information are given in Table S3.

2.2 Frequency chaos game representation on DNA

The first application of CGR on DNA was made by Jeffrey in 1990 [51]. He applied CGR with a square instead of a triangle and used its four vertices to represent the four nucleotides, adenine (A), cytosine (C), guanine (G) and thymine (T) (Figure 1). The figure on the left shows the CGR algorithm and coordinates for four vertices labeled A, C, G, and T. In CGR, the center has coordinates (0,0) and CGR extends from (−1,−1) to (1,1). The figure on the right shows the partitioning of the CGR space iteratively [5].

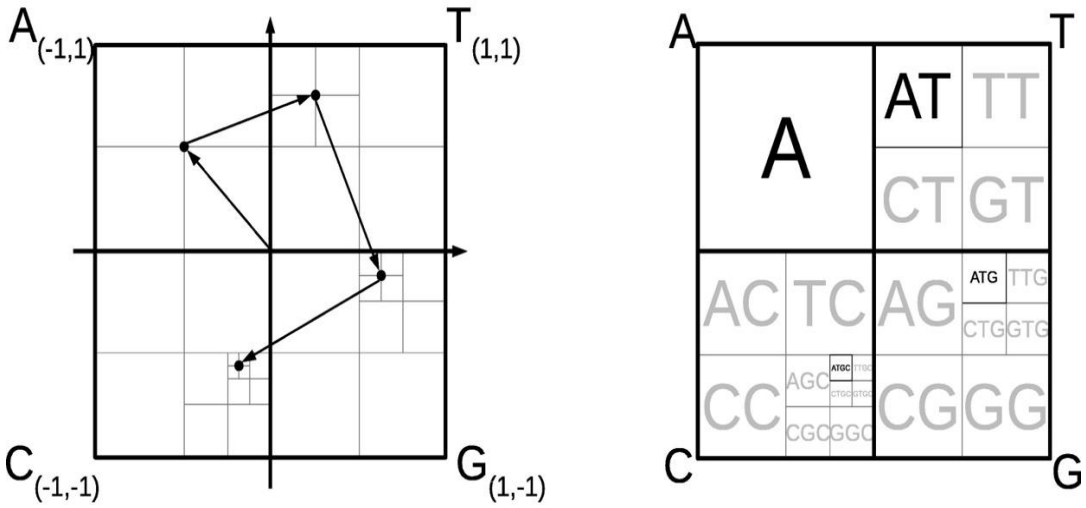

Figure 1. Application of CGR to DNA

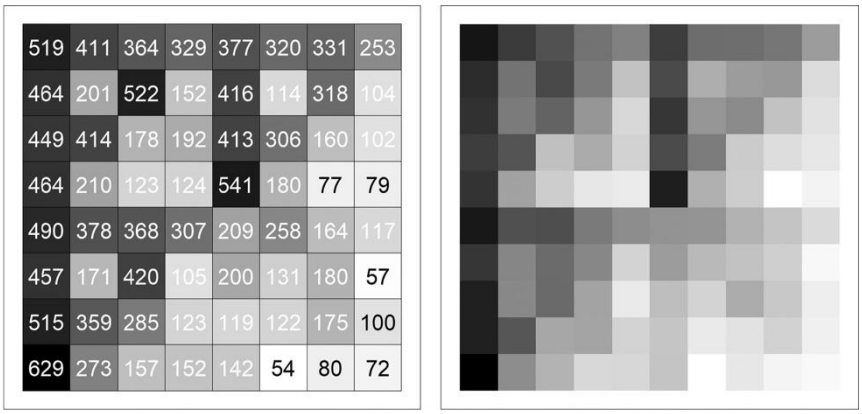

Figure 2. FCGR with count matrix for the complete mitochondrial genome sequence of the Tyrolean Iceman (GenBank:EU810403.1) and FCGR as a grayscale visualization [36].

While the CGR algorithm uses exact coordinates for each point, a discritization called frequency chaos game representation (FCGR) has enabled the use of a coarse-grained and less noisy CGR abstraction for sequences. FCGR is based on the frequency of CGR points according to a predefined grid according to the mapping in CGR. In Figure 2, CGR is divided by a grid (here 8x8 cells) and the number of points in each cell is counted. This count is presented as a matrix representing the frequency of k-mers (here 3-mers) in the corresponding map region and a grayscale image normalized between 0 and 255 [5].

2.3 Transfer learning models

Transfer learning constitutes a methodology that facilitates the reutilisation of deep learning models that have been pre-trained on extensive and diverse datasets on smaller and specialised datasets. This method facilitates the successful application of deep learning models that necessitate substantial data for training on smaller, more specialised datasets [52]. In bioinformatics and genetic analyses, transfer learning offers significant advantages in terms of time and resource efficiency during the visualisation and analysis of large genomic datasets.

In this study, three distinct transfer learning models were employed to enhance the extraction of information from visualised representations of genetic data (fCGR) and to assess the generalisability of the proposed method across diverse datasets: ResNet50, EfficientNetB0, and MobileNet. The selection of these models was made based on their suitability for the needs of the study, their effectiveness in the literature, and the advantages provided by transfer learning.

- ResNet50 has been shown to demonstrate superior performance in the learning of complex relationships by preventing information loss in deep learning processes thanks to residual connections. Its efficacy in complex tasks, such as medical image analysis, has been well-documented in the literature. Notably, it is particularly adept at extracting intricate features from high-resolution fCGR images.
- EfficientNetB0 offers a balanced approach between computational efficiency and accuracy with the Compound Scaling method. Its capacity to provide high accuracy while optimising the processing cost in medium-sized datasets

enables it to provide an efficient solution in genetic data. The EfficientNet model has been documented as successful in a wide range of applications in areas such as biomedical image analysis (see the relevant literature for further details).

- MobileNet, on the other hand, is a lightweight architecture that provides expeditious results with minimal computational expense. It is particularly effective in cases where limited processing power is a constraint, such as in small and medium-sized datasets, including fCGR images extracted from short DNA sequences. MobileNet has been documented in the literature as providing high accuracy at low cost in the context of medical imaging and classification.

The three models under consideration permit effective analysis of fCGR images obtained from DNA sequences by taking advantage of the strong generalisation capacity provided by transfer learning. In addition, it has been stated that the selected models offer lower processing cost and a wider application area when compared to other models in the literature (such as DenseNet, VGGNet, InceptionNet). For these reasons, these models were selected for the purposes of evaluating the performance of our study on different datasets and measuring the comparative effectiveness of the models [53].

2.3.1 ResNet50

ResNet50 is a 50-layer convolutional neural network (CNN) that stands out with its residual connections used in deep learning models [54]. Residual connections allow deeper networks to be trained by preventing the vanishing gradient problem. ResNet50 has been frequently used in medical image analysis and bioinformatics applications due to its capacity to extract complex features from high resolution images. In this study, ResNet50 is used to extract high resolution detailed features from fCGR images. By removing the last classification layer of the model (include_top = False), 2.048-dimensional feature vectors were obtained from each image with the Global Average Pooling (GAP) method. This structure was especially advantageous in learning complex relationships in fCGR images created from long DNA sequences.

2.3.2 EfficientNetB0

EfficientNetB0 is a CNN architecture that provides both accuracy and computational efficiency by scaling the depth, width and resolution dimensions of the model in a balanced way with the Compound Scaling method. EfficientNetB0 is designed to achieve high accuracy in limited processing power and datasets [53]. In this study, EfficientNetB0 is used to optimize the processing costs during feature extraction from fCGR images and at the same time to provide high accuracy. The classification layer of the model is removed and 1,280-dimensional feature vectors are extracted from each image with the GAP method. EfficientNetB0 contributes to the generalizability of the study by preserving the accuracy while reducing the processing time in images created from medium-length DNA sequences.

2.3.3 MobileNet

MobileNet is a CNN architecture that provides lightness and speed in deep learning applications. With a method called depthwise separable convolution, the number of model parameters and computational cost are significantly reduced. MobileNet is known for its use especially in mobile devices or systems with limited processing power [55]. In this study, MobileNet is used to quickly extract information from fCGR images. The last classification layer of the model is removed and 1,024-dimensional feature vectors are obtained from each image with the GAP method. MobileNet strengthens the performance of the study on different datasets by quickly and effectively extracting features in fCGR images derived from short DNA sequences.

2.4 Details of the model

2.4.1 Model structure

In this study, three distinct transfer learning models are utilised to extract information from fCGR (Chaos Game Representation) images derived from DNA sequences. These models are distinguished from each other by their technical features.

ResNet50 is a 50-layer deep CNN architecture with residual connections. The incorporation of residual connections within deep CNN architectures has been demonstrated to enhance the trainability of such networks, facilitate the transmission of input information to subsequent layers, and thereby mitigate information loss during the analysis of complex images. In this study, ResNet50 is optimised to learn detailed features of fCGR images derived from long DNA sequences.

EfficientNetB0 is an architecture that scales the depth, width and resolution dimensions in a balanced way with the Compound Scaling method. This structure, which provides high accuracy with fewer parameters, has made efficient information extraction by reducing the processing cost in medium-sized datasets such as fCGR images.

MobileNet is an architecture that reduces the number of parameters and processing cost with depthwise separable convolution. This structure facilitates rapid inference from fCGR images, a capability that is particularly effective in short DNA sequences.

The efficacy of each model in enhancing the study's scope is noteworthy, particularly in regard to its capacity to process genetic information from fCGR images, a capability that is well-suited to the unique characteristics of this data type.

2.4.2 Feature extraction

In this study, fCGR (Chaos Game Representation) images derived from DNA sequences were converted into feature vectors using three deep learning-based transfer learning models (ResNet50, EfficientNetB0, and MobileNet). In this process, the final classification layers of the transfer learning models were removed and global feature extraction was performed from the images. This approach enabled the effective modelling of biological information specific to DNA sequences from fCGR images.Furthermore, fCGR images were rescaled to 224x224 pixels to ensure compatibility with the input layers of all models. The images were processed in RGB format and converted into feature vectors with a specific pooling

strategy for each model. During the extraction of features, a fixed-size feature vector was obtained from each image by applying the Global Average Pooling (GAP) method:

• ResNet50: The features extracted from the images formed a 2,048-dimensional vector. This high dimensionality reflects the capacity of ResNet50 to extract detailed information in complex and high-resolution fCGR images.

• EfficientNetB0, optimised by the Compound Scaling structure, generated a 1,280-dimensional feature vector from each image. This balanced scaling is employed with the objective of reducing the computational cost of the model.

• MobileNet: Utilising depth-separated convolution technology, MobileNet extracted a 1,024-dimensional vector, thereby achieving lightweight and fast feature extraction.

It is noteworthy that all models were executed with pre-trained weights on ImageNet for feature extraction, and no fine-tuning was performed. The extracted feature vectors provided biological representations that effectively represented genetic information and were made suitable for subsequent analysis steps.

2.4.3 Network Weights

All transfer learning models utilised in this study (ResNet50, EfficientNetB0, and MobileNet) were executed with weights that had previously been trained on the ImageNet dataset. ImageNet is a substantial and diverse dataset consisting of natural images, which increases the generalisation capacity of the models and enables effective

results to be obtained in different data types. In this study, these pre-trained weights were utilised to represent genetic information in fCGR images derived from DNA sequences.

The final classification layers of the models were removed, and feature extraction was performed only with the pre-trained weights. This approach enhanced the model's capacity to extract information in new data types by leveraging the generalization advantage provided by the transfer learning method in DNA datasets of limited size. Each model extracted feature vectors of differing dimensions from fCGR images, characterised by its distinct weight structure:

- **ResNet50:** Achieved high accuracy in complex images with 25.6 million parameters.
- **EfficientNetB0:** Achieved balanced accuracy and computational cost efficiency with 5.3 million optimized parameters.
- **MobileNet:** It has achieved fast and low-cost inference with its lightweight structure and 4.2 million parameters.

It is important to note that the weights were not fine-tuned, and the models were run with fixed weights. This strategy reduced the risk of overfitting on existing datasets, making it possible to extract accurate and generalisable features from visualised representations of genetic data.

2.4.4 Preparation of fCGR images

The conversion of DNA sequences into visual representations was achieved by employing the Frequency Chaos Game Representation (fCGR) method. This method involves the calculation of the frequencies of specific nucleotide subsequences (k-mers) of a given length within the sequence, and the subsequent visualisation of this information as a two-dimensional matrix.The fCGR matrices were created according to the k-mer length that had been determined for each individual DNA sequence. These matrices were then saved as grayscale images. Subsequent to this, the images were rescaled to 224x224 pixels and converted to RGB format, ensuring compatibility with deep learning models. This format was selected to ensure compatibility with the input format employed in the training of models with ImageNet data, thereby facilitating efficient model operation.

Determination of k-length: The determination of k-mer length constitutes a critical step that exerts a direct influence on the discrimination power and computational cost of fCGR. A plethora of approaches have been proposed in the extant literature to determine the optimal k-mer length; however, these recommendations are contingent on the features of the methods utilised. Shorter k-mer lengths have been shown to reduce the discrimination power of subsequences, while longer k-mer lengths have been shown to increase the discrimination power, albeit at the cost of significantly increased computational complexity. In this study, the logL(n) calculation was utilised as a starting point, an approach that has been accepted in the literature [56]. In this method, n denotes the average length of the sequences to be compared, while L signifies the size of the DNA alphabet (e.g. 4: {A, C, G, T}). In order to better determine the optimal k-mer length, analysis was performed at different k-mer lengths using one value below and one value above this value. This approach yielded a method that provided an approximate optimum limit in k-mer length selection, and the effect of the determined k-mer length on the analysis results was evaluated.

2.4.5 Normalisation of data

Prior to the integration of fCGR images into the models' inputs, a process of normalisation was implemented. This process entailed the subtraction of a specific mean from the pixel values of all images, with these values then being scaled between -1 and 1. The application of this process rendered the models more stable and efficient, whilst concomitantly reducing potential disparities in images from disparate data sources.

This process was implemented for all transfer learning models utilised in the study (ResNet50, EfficientNetB0, and MobileNet), in accordance with the standards that the models were trained with, namely the ImageNet dataset. The application of this process to the normalized images enabled the models to extract meaningful biological information from fCGR images and accurately represent the biological properties of DNA sequences. Furthermore, the normalization process ensured that the general visual features inherent in each model contributed effectively to the acquisition of biological information from fCGR images.

The diagram showing the steps of our method, the details of which will be given in the next subsections, is given in Figure 3.

2.5 *Clustering and reconstruction of the Pylogenetic Tree*

In this study, the authors evaluated the phylogenetic relationships between DNA sequences and reconstructed phylogenetic trees using feature vectors extracted from fCGR images. To this end, three different distance metrics (cosine similarity, Euclidean distance and correlation distance) and four different hierarchical clustering methods (average linkage, single linkage, complete linkage, and ward linkage) were applied to measure the similarities between feature vectors. The employment of a range of metrics and methods was driven by the objective of offering a more comprehensive perspective on phylogenetic analyses and comparing the performance of the methods.

2.5.1 *Distance metrics*

Distance metrics serve as the fundamental tools for determining relationships between feature vectors. In this study, the following three metrics were utilised:

- **Cosine Similarity:** This metric quantifies the angular relationship between two vectors and is well-suited for similarity analyses independent of size.The efficacy of this metric in evaluating the orientation-based similarities of fCGR feature vectors extracted from DNA sequences was demonstrated (Equation (1)).

$$\cos(\theta) = \frac{x.y}{\|x\|\|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

- **Euclidian distance:** A conventional metric that quantifies the geometric distance between two vectors. This metric is well-suited for the analysis of distances resulting from genetic mutations or minor alterations (Equation (2)).

$$\mathrm{d}(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \quad (2)$$

- **Correlation distance:** A metric that quantifies the linear relationship between two vectors, particularly employed for the evaluation of the density and linear correlations of genetic information (Equation (3)).

$$\mathrm{correlation}(A,B) = 1 - \frac{cov(A,B)}{\sigma_A.\sigma_A} \quad (3)$$

Each distance metric has enhanced the generalisability of analyses by offering a distinct perspective on genetic data.

2.5.2 *Clustering Methods*

Hierarchical clustering methods are utilised to construct phylogenetic trees from the relationships obtained from distance metrics. In this study, four distinct methods were employed:

- **Average Linkage:** This method provides a balanced structure by considering the average distance between groups.
- **Single Linkage:** This method is based on the closest distance between two groups and creates more separated clusters.
- **Complete Linkage:** This method provides a tighter clustering structure by measuring the farthest distance between two groups.

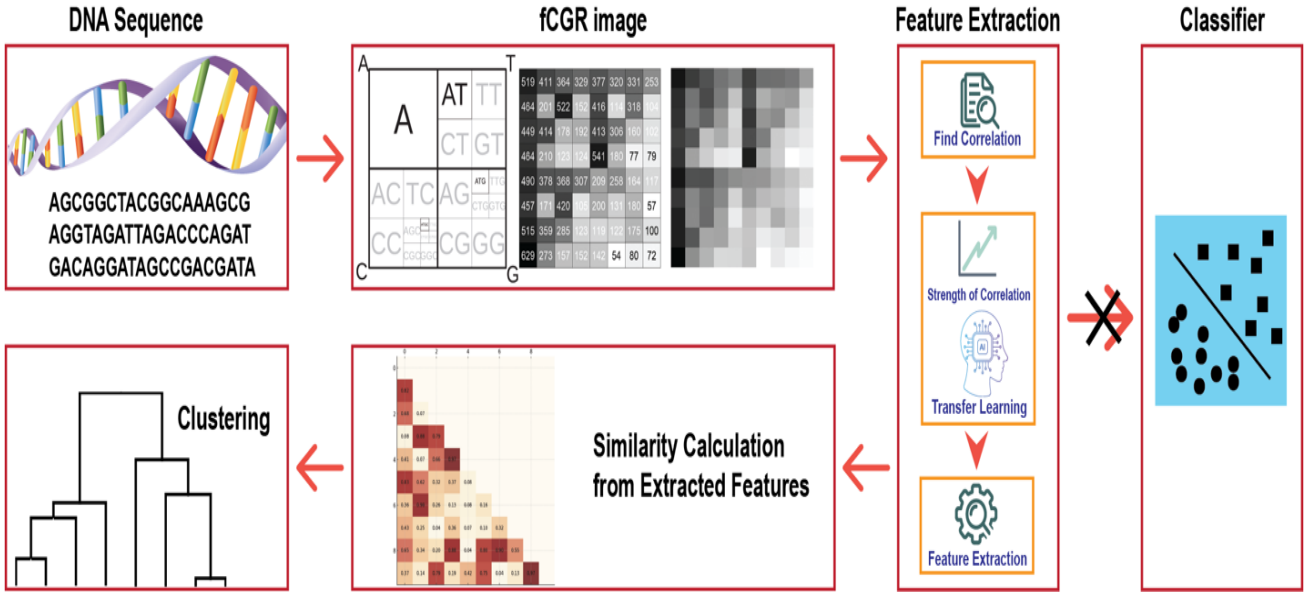

Figure 3. Diagram of the proposed method

- **Ward Linkage:** This method preserves the diversity in the data set by minimizing the sum of the squares of the differences.

The application of these methods in combination enabled the analysis of phylogenetic relationships from both a detailed and a broad perspective.

The performance of the various metrics and methods employed in this study was evaluated using the Normalized Robinson-Foulds (nRF) score, by comparing them with the reference phylogenetic trees. This score measured the similarity of the reconstructed phylogenetic tree to the reference tree and revealed the accuracy of the method. Furthermore, a comparative analysis of the nRF scores derived from various metric and method combinations was conducted, leading to the identification of the most efficacious combinations. The nRF metric was utilised to compare the generated phylogenetic trees with a reference tree, thereby evaluating the efficacy of the employed method. The calculation of accuracy is outlined in Equation (4):

$$Accuracy = 1 - nRF * 100 \quad (4)$$

As a result, this diversification provided a more comprehensive analysis of the phylogenetic relationships of DNA sequences and provided the opportunity to better evaluate the biological significance extracted from genetic data.

3 Results

This study comprehensively evaluated the automatic extraction and reconstruction of phylogenetic trees from DNA sequences using fCGR images. A range of transfer learning models (ResNet50, EfficientNetB0, and MobileNet) were analysed with various combinations of k-mer lengths, distance metrics (cosine similarity, Euclidean distance, correlation), and hierarchical clustering methods (average, single, complete, ward linkage). The results demonstrate that the proposed method can provide a generalisable solution and that transfer learning is an effective tool in the analysis of genetic data. The accuracy rates presented in the results were determined through the alignment process facilitated by the MEGA11 [57] tool, a widely utilised and referenced bioinformatics instrument, and the comparison to the reference tree generated by the UPGMA method.

Table 2. Accuracy rates of model, method and parameter combinations (%)

			Cosine				Euclidean				Correlation			
			Avg.	**Ward**	**Single**	**Compl.**	**Avg.**	**Ward**	**Single**	**Compl.**	**Avg.**	**Ward**	**Single**	**Compl.**
Corona	**Resnet50**	**k=6**	40.63	40.63	40.63	**50.00**	40.63	40.63	40.63	43.75	40.63	40.63	40.63	**50.00**
		k=7	34.38	31.25	37.50	34.38	34.38	31.25	34.38	34.38	34.38	31.25	37.50	34.38
		k=8	31.25	28.13	31.25	28.13	31.25	25.00	31.25	28.13	31.25	28.13	31.25	28.13
	Mobilenet	**k=6**	59.38	**65.63**	56.25	62.50	62.50	62.50	59.38	62.50	62.50	65.63	56.25	62.50
		k=7	56.25	56.25	59.38	56.25	59.38	56.25	59.38	53.13	56.25	56.25	59.38	59.38
		k=8	34.38	31.25	34.38	31.25	31.25	25.00	34.38	31.25	34.38	31.25	34.38	31.25
	EfficientNet	**k=6**	**53.13**	46.88	**53.13**	46.88	**53.13**	46.88	**53.13**	46.88	**53.13**	46.88	**53.13**	46.88
		k=7	50.00	46.88	50.00	46.88	46.88	50.00	50.00	46.88	50.00	46.88	50.00	46.88
		k=8	31.25	31.25	31.25	31.25	31.25	31.25	31.25	31.25	28.13	28.13	28.13	28.13
Primates	**Resnet50**	**k=6**	26.66	33.33	33.33	26.66	33.33	40.00	**46.66**	26.66	26.66	33.33	33.33	26.66
		k=7	40.00	40.00	40.00	40.00	33.33	33.33	33.33	33.33	40.00	40.00	40.00	40.00
		k=8	33.33	33.33	40.00	33.33	33.33	26.66	26.66	33.33	33.33	33.33	40.00	33.33
	Mobilenet	**k=6**	**73.34**	60.00	53.33	60.00	73.33	46.66	53.33	60.00	**73.34**	60.00	53.33	46.66
		k=7	26.67	26.67	26.67	26.67	40.00	26.67	26.67	26.67	26.67	26.67	26.67	26.67
		k=8	40.00	33.34	33.34	33.34	40.00	40.00	40.00	40.00	40.00	33.34	33.34	33.34
	EfficientNet	**k=6**	26.67	26.67	20.00	20.00	33.34	33.34	26.67	33.34	26.67	26.67	20.00	20.00
		k=7	66.67	**73.34**	66.67	66.67	66.67	60.00	66.67	66.67	66.67	**73.34**	66.67	66.67
		k=8	40.00	33.34	46.66	33.34	40.00	33.34	46.66	33.34	40.00	33.34	46.66	33.34
Influenza	**Resnet50**	**k=4**	40.00	31.43	37.14	34.29	40.00	34.29	37.14	34.29	40.00	31.43	37.14	34.29
		k=5	**42.86**	40.00	40.00	37.14	37.14	37.14	37.14	40.00	**42.86**	40.00	40.00	37.14
		k=6	34.29	31.43	31.43	37.14	31.43	28.57	28.58	25.71	34.29	34.29	31.43	37.14
	Mobilenet	**k=4**	57.14	**62.86**	57.14	51.43	57.14	57.14	51.43	51.43	57.14	**62.86**	57.14	57.14
		k=5	40.00	45.71	42.86	42.86	40.00	42.86	40.00	45.71	40.00	45.71	42.86	42.86
		k=6	40.00	37.14	37.14	40.00	40.00	37.14	34.29	40.00	40.00	37.14	37.14	40.00
	EfficientNet	**k=4**	25.71	28.57	31.43	28.57	28.57	31.43	31.43	28.57	25.71	28.57	31.43	28.57
		k=5	34.29	34.29	34.29	34.29	37.14	37.14	37.14	**40.00**	34.29	34.29	34.29	34.29
		k=6	34.29	34.29	34.29	37.14	37.14	28.57	34.29	42.86	34.29	37.14	34.29	37.14

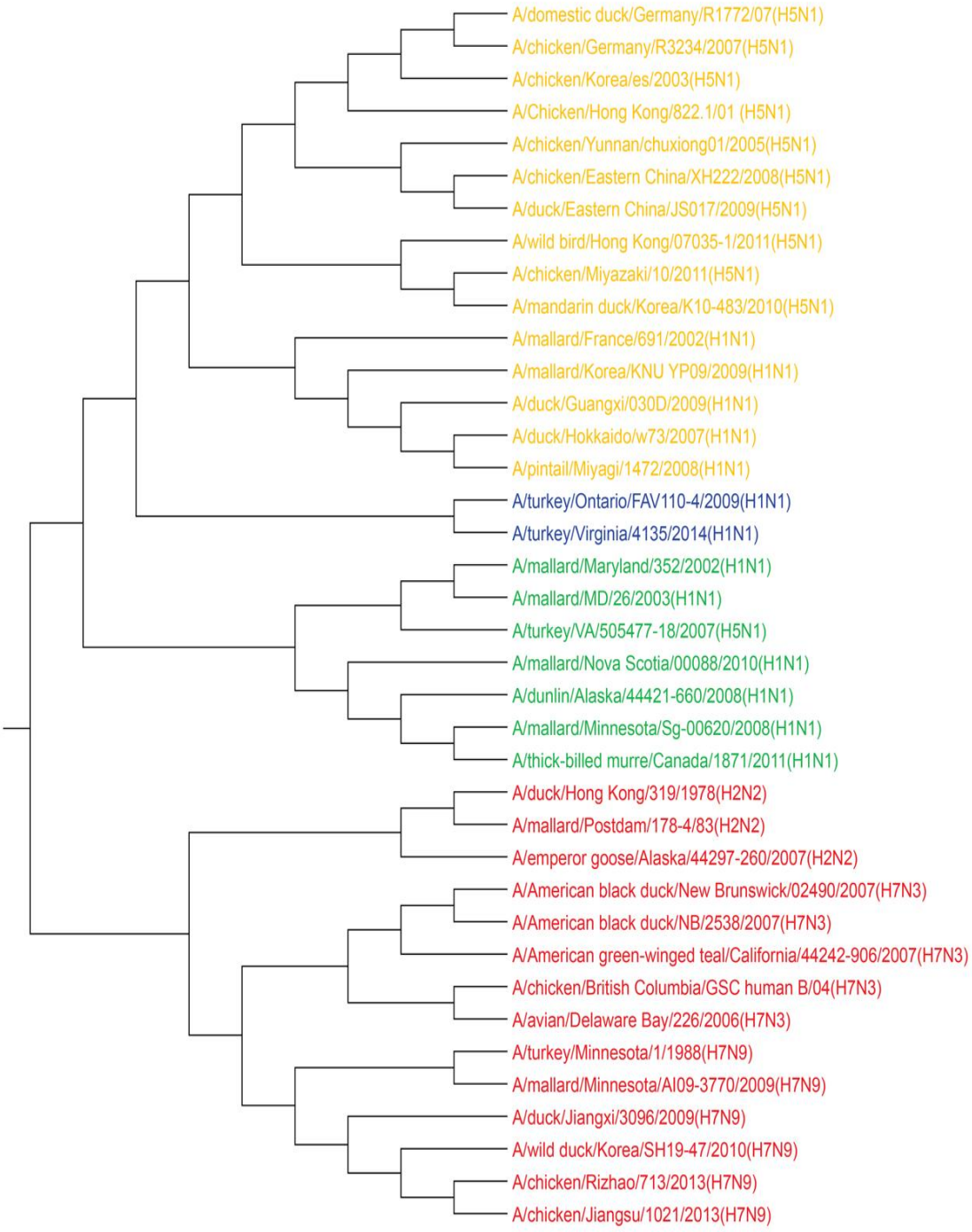

Figure 4. Phylogenetic tree constructed by alignment with ClustalW method and UPGMA method using MEGA11 software

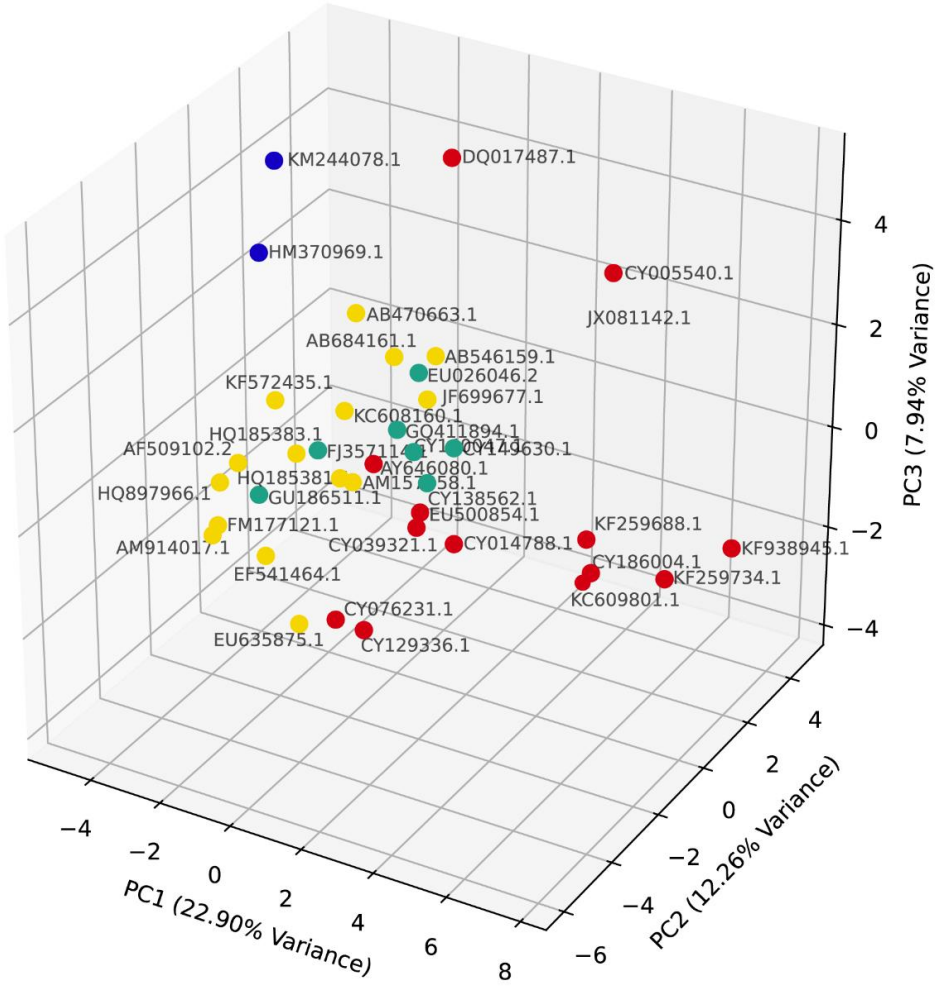

Figure 5. Projection of 3D space consisting of PC1, PC2 and PC3 principal components for features obtained from 38 Influenza A virus DNAs**.**

3.1 Comperative analysis of accuracy across combinations

The accuracy of various combinations was analysed using all datasets. Table 2 presents the accuracies of combinations of three models, three k-mer lengths, three distance metrics, and four clustering methods. This variation has been designed to evaluate both the methodological flexibility and the applicability of transfer learning-based automatic feature extraction to different data types.

3.2 PCA based visualizations of feature spaces

The species in the phylogenetic tree displayed in Figure 4 were grouped and coloured. The feature vectors obtained from the transfer learning models according to the parameter and model preference with the highest accuracy score were projected into a three-dimensional space using PCA (Principal Component Analysis). The graph presented in Figure 5 illustrates that the proportion of information encompassed within the three principal components exhibiting the highest degree of variance is equivalent to 43.10%. This visualisation method enables the observation of the clustering structure of the vectors belonging to the data set and the compatibility of these structures with phylogenetic trees. This analysis demonstrates the potential of feature vectors in reflecting genetic similarities and underscores the efficacy of the proposed method in elucidating relationships within biological data sets.

The phylogenetic tree displayed in Figure 6 has been grouped and coloured according to the hierarchical clustering of species in the 18 eutherian mammals dataset. The construction of this tree was achieved through the utilisation of feature vectors derived from transfer learning models, with the parameter and model combination that yielded the maximum accuracy score serving as the foundation. In order to further examine the structure of the data, the feature vectors were reduced to a three-dimensional space using PCA (Principal Component Analysis). The graph shown in Figure 7 reveals that the three principal components exhibiting the highest variance cover 60.55% of the total information. This approach enables the observation of the clustering tendencies of the vectors belonging to the dataset and the compatibility of these structures with phylogenetic trees. The findings emphasise the potential of feature vectors to reflect genetic similarities and the effectiveness of the proposed method in explaining relationships in biological datasets.

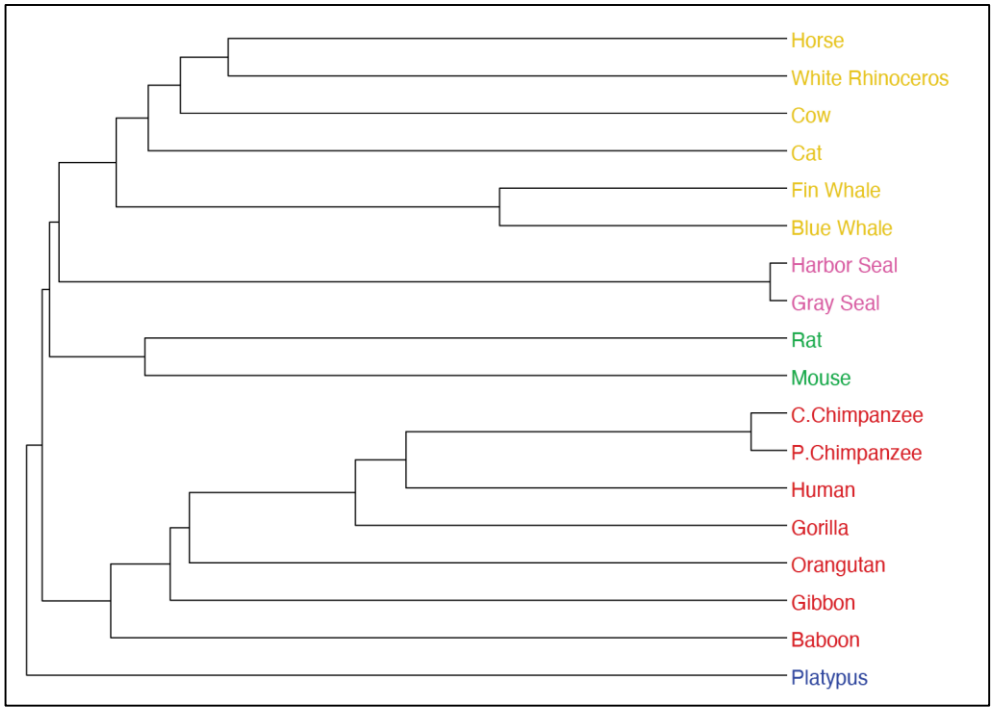

Figure 2. Phylogenetic tree constructed by alignment with ClustalW method and UPGMA method using MEGA11 software

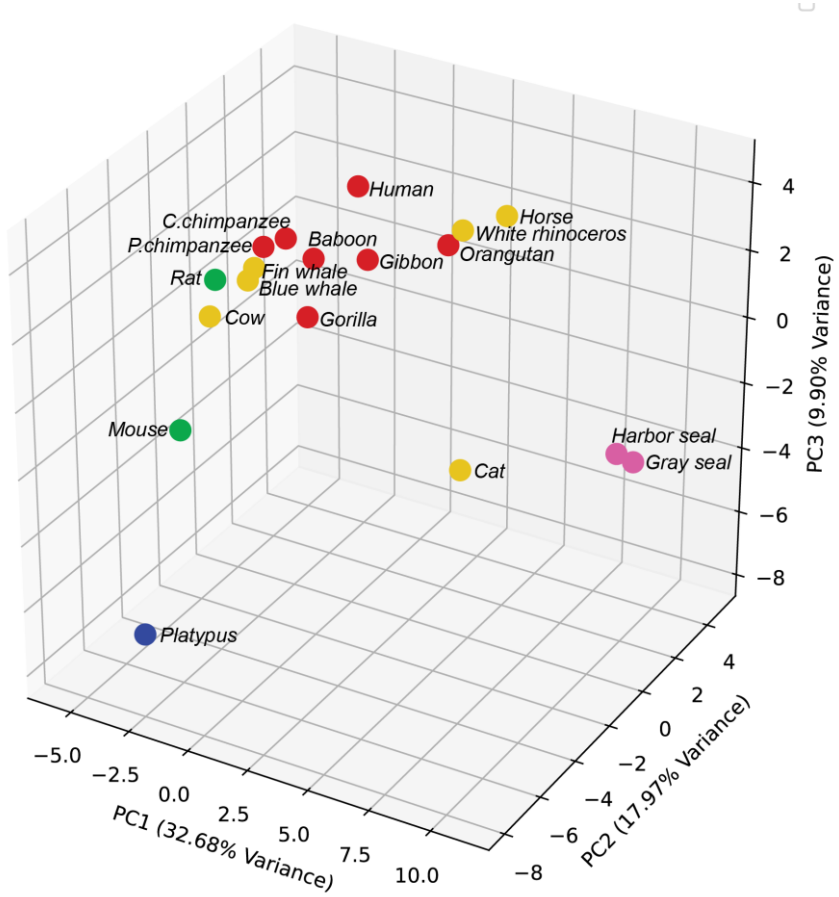

Figure 3. Projection of 3D space consisting of PC1, PC2 and PC3 principal components for features obtained from 18 mtDNAs of Mammals

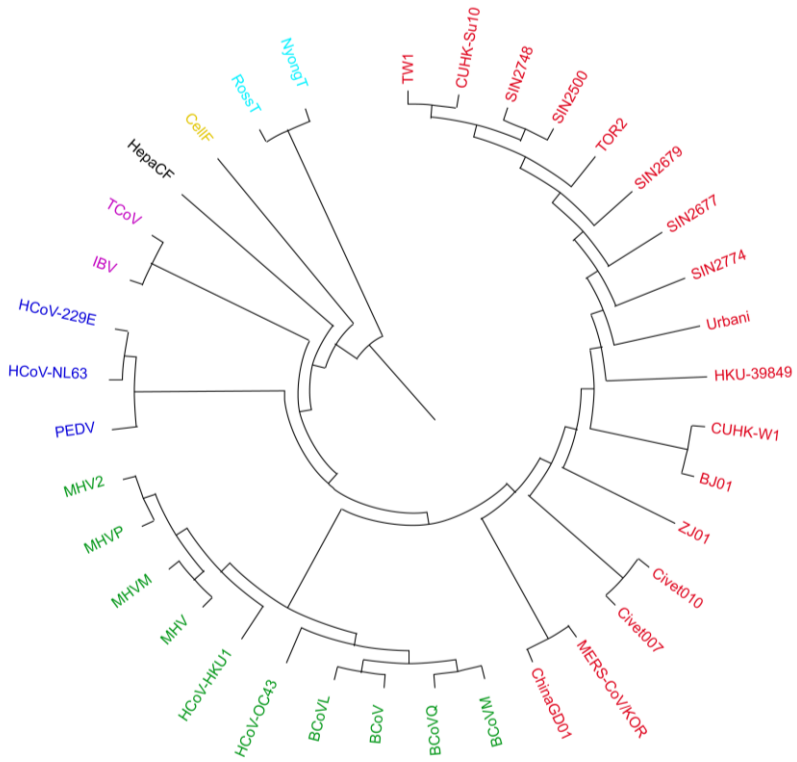

Figure 4. Phylogenetic tree constructed by alignment with ClustalW method and UPGMA method using MEGA11 software

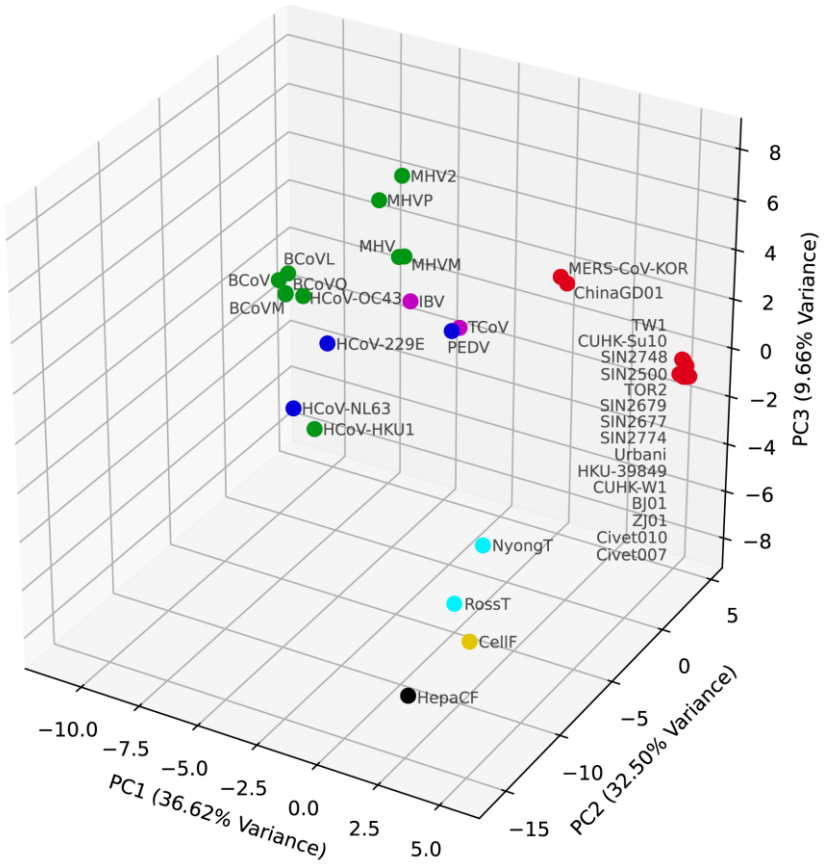

Figure 5. Projection of 3D space consisting of PC1, PC2 and PC3 principal components for features obtained from Coronavirus Dataset

The phylogenetic tree displayed in Figure 8 has been grouped and coloured according to the hierarchical clustering of species in the Coronavirus dataset. The construction of this tree was achieved through the utilisation of feature vectors derived from transfer learning models, with the parameter and model combination that yielded the maximum accuracy score serving as the foundation. In order to further examine the structure of the data, the feature vectors were reduced to a three-dimensional space using PCA (Principal Component Analysis). The graph shown in Figure 9 reveals that the three principal components exhibiting the highest variance cover 78.78% of the total information. This approach enables the observation of the clustering tendencies of the vectors belonging to the dataset and the compatibility of these structures with phylogenetic trees. The findings emphasise the potential of feature vectors to reflect genetic similarities and the effectiveness of the proposed method in explaining relationships in biological datasets.

3.3 Contributions and future perspectives

This study proposes a novel methodology for the automatic extraction of features from fCGR images, utilising transfer learning models to represent DNA sequences independently of alignment. A comprehensive comparison between three transfer learning models (ResNet50, EfficientNet, MobileNet) and three different distance metrics (Cosine, Euclidean, Correlation) and four different clustering methods (Average, Ward, Single, Complete) used in the study is presented. This comparison demonstrates the generalizability of the method and its performance on different datasets in detail.

The results presented in Table 2 demonstrate that the cosine metric is more effective than other metrics in reflecting genetic similarities between feature vectors obtained from DNA sequences. This finding aligns with the extant literature, which supports the capacity of the cosine metric to evaluate direction-independent similarities in DNA analyses. It is well-documented in the existing literature that the cosine metric has a proven track record in reflecting density-based and angular similarities in genetic datasets. In this context, the preference for the cosine metric provides a robust foundation for the proposed method, particularly in the analysis of intricate genetic relationships.

In the context of transfer learning models, it has been observed that the MobileNet model consistently achieves the highest accuracy rates. This is attributable to the MobileNet model's low computational cost and lightweight structure, which facilitates efficient information extraction in limited datasets. The fast and efficient computational capacity of MobileNet supports the high performance obtained, especially on fCGR images derived from short DNA sequences. While ResNet50 and EfficientNet models have been shown to achieve commendable results on specific datasets, MobileNet's balanced approach, which strikes a judicious equilibrium between computational cost and accuracy, is particularly noteworthy.

One of the most significant contributions of this study is to demonstrate the applicability of transfer learning models in alignment-independent DNA similarity analysis. Furthermore, the methodology outlined herein demonstrates the capacity for genetic relationships to be visualised and modelled via fCGR images. The use of PCA visualisations provides clear evidence for the distribution of feature vectors in a coherent manner in clusters on phylogenetic trees. This finding lends further support to the efficacy of the proposed method in reflecting meaningful genetic patterns in biological datasets.

- Methodological Diversity: In this study, the diversity of metric and clustering method combinations is designed to comprehensively evaluate how the method performs on different datasets. This methodological diversity serves to enhance the generalisability of the proposed method and its applicability in bioinformatic analyses.
- Originality and Innovation: The study presents a significant innovation in the use of transfer learning models in DNA similarity analysis. The integration of the fCGR method, a widely used technique in the relevant literature, with deep learning models suggests a new analysis paradigm compared to traditional approaches.
- Success Rates: The accuracy rates obtained demonstrate the potential for enhancement of the method in its present state. The enhancement of these rates can be achieved through model optimisation, the utilisation of more extensive datasets, and the refinement of the fCGR method.

In conclusion, this study demonstrates the usability of transfer learning models in similarity analysis of DNA sequences and provides a solid foundation for future studies. Future studies may expand the scope to enhance the accuracy and biological significance of these methods.

4 Conclusion

This study comprehensively examined the usability of the Frequency Chaos Game Representation (fCGR) method for alignment-independent analysis of DNA sequences by integrating it with transfer learning models. A comparative analysis was performed on three transfer learning models (ResNet50, EfficientNetB0 and MobileNet) used in the study, three different distance metrics (Cosine, Euclidean, Correlation) and four clustering methods (Average, Ward, Single, Complete). The results demonstrated the generalizability of the method and its applicability in biological datasets.

The findings indicated that the cosine metric exhibited enhanced consistency and efficacy in reflecting genetic similarities between DNA sequences in comparison to other metrics. In accordance with extant literature, the capacity of the cosine metric to reflect direction-independent similarities increased the effectiveness of feature vectors obtained from fCGR images in modelling genetic relationships. Furthermore, MobileNet demonstrated a notable advantage over other transfer learning models by exhibiting low

computational cost and high accuracy rates. The lightweight architecture of MobileNet yielded particularly impressive results in the context of data obtained from short DNA sequences.

This study demonstrates the potential of transfer learning models in bioinformatics analysis and proves that processing DNA sequences with visual representations is an effective method for genetic similarity analysis. Nevertheless, the findings indicate that the method has potential for enhancement, and future studies may concentrate on the following areas:

The following tests were carried out:

- An evaluation of model performance with larger datasets;
- An optimisation of transfer learning models with fine-tuning techniques;
- An evaluation of new architectures and innovative fCGR variations in the feature extraction process.

The findings demonstrate that the proposed approach provides a powerful tool for understanding similarities and relationships in genetic data. The study provides a new perspective on the role of transfer learning models in bioinformatics analysis and provides a solid foundation for future research in this area.

Conflict of Interest

The authors declare that there is no conflict of interest.

Similarity Rate (Turnitin): 9%

References

[1] Z. D. Stephens et al., Big Data: Astronomical or Genomical?, PLoS Biol, 13, 7, p. e1002195, 2015. https://doi.org/10.1371/JOURNAL.PBIO.1002195.

[2] S. B. Needleman and C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J Mol Biol, 48, 3, 443–453, Mar. 1970. https://doi.org/10.1016/0022-2836(70)90057-4.

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, Basic local alignment search tool, J Mol Biol, 215, 3, 403–410, Oct. 1990. https://doi.org/10.1016/S0022-2836(05)80360-2.

[4] S. Vinga and J. Almeida, Alignment-free sequence comparison—a review, Bioinformatics, 19, 4, 513–523, Mar. 2003. https://doi.org/10.1093/BIOINFORMATICS/BTG005

[5] H. F. Löchel and D. Heider, Chaos game representation and its applications in bioinformatics, Comput Struct Biotechnol J, 19, 6263–6271, Jan. 2021. https://doi.org/10.1016/J.CSBJ.2021.11.008.

[6] M. Yousef and J. Allmer, Deep learning in bioinformatics, Turkish Journal of Biology, 47, 6, p. 366, 2023. https://doi.org/10.55730/1300-0152.2671.

[7] H. Gunasekaran, K. Ramalakshmi, A. Rex Macedo Arokiaraj, S. D. Kanmani, C. Venkatesan, and C. S. G. Dhas, Analysis of DNA Sequence Classification Using CNN and Hybrid Models, Comput Math Methods Med, 2021, 1, p. 1835056, Jan. 2021. https://doi.org/10.1155/2021/1835056.

[8] A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski, Alignment-free sequence comparison: benefits, applications, and tools, Genome Biology 2017 18:1, 18, 1, 1–17, Oct. 2017. https://doi.org/10.1186/S13059-017-1319-7.

[9] A. Zielezinski et al., Benchmarking of alignment-free sequence comparison methods, Genome Biol, 20, 1, 1–18, Jul. 2019. https://doi.org/10.1186/S13059-019-1755-7/TABLES/1.

[10] R. Rizzo, A. Fiannaca, M. La Rosa, and A. Urso, A deep learning approach to DNA sequence classification, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9874 LNCS, 129–140, 2016. https://doi.org/10.1007/978-3-319-44332-4_10/FIGURES/7.

[11] O. Bonham-Carter, J. Steele, and D. Bastola, Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis, Brief Bioinform, 15, 6, 890–905, Nov. 2014. https://doi.org/10.1093/BIB/BBT052.

[12] M. Uddin, M. K. Islam, M. R. Hassan, F. Jahan, and J. H. Baek, A fast and efficient algorithm for DNA sequence similarity identification, Complex and Intelligent Systems, 9, 2, 1265–1280, Apr. 2023. https://doi.org/10.1007/S40747-022-00846-Y/TABLES/12.

[13] S. Zou, L. Wang, and J. Wang, A 2D graphical representation of the sequences of DNA based on triplets and its application, EURASIP J Bioinform Syst Biol, 2014, 1, 2014. https://doi.org/10.1186/1687-4153-2014-1.

[14] N. Jafarzadeh and A. Iranmanesh, C-curve: A novel 3D graphical representation of DNA sequence based on codons, Math Biosci, 241, 2, 217–224, Feb. 2013. https://doi.org/10.1016/J.MBS.2012.11.009.

[15] P. Waz and D. Bielińska-Waz, Non-standard similarity/dissimilarity analysis of DNA sequences, Genomics, 104, 6, 464–471, Dec. 2014. https://doi.org/10.1016/J.YGENO.2014.08.010.

[16] B. Liao, M. Tan, and K. Ding, A 4D representation of DNA sequences and its application, Chem Phys Lett, 402, 4–6, 380–383, Feb. 2005. https://doi.org/10.1016/J.CPLETT.2004.12.062.

[17] B. Liao, R. Li, W. Zhu, and X. Xiang, On the similarity of DNA primary sequences based on 5-D representation, J Math Chem, 42, 1, 47–57, Jul. 2007. https://doi.org/10.1007/S10910-006-9091-Z/METRICS.

[18] B. Liao and T. M. Wang, Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases, J Chem Inf Comput Sci, 44, 5, 1666–1670, Sep. 2004. https://doi.org/10.1021/CI034271F/ASSET/IMAGES/LARGE/CI034271FF3.JPEG.

[19] E. Delibaş and A. Arslan, DNA sequence similarity analysis using image texture analysis based on first-

order statistics, J Mol Graph Model, 99, p. 107603, Sep. 2020. https://doi.org/10.1016/j.jmgm.2020.107603.

[20] W. Chen, B. Liao, and W. Li, Use of image texture analysis to find DNA sequence similarities, J Theor Biol, 455, 1–6, Oct. 2018. https://doi.org/10.1016/J.JTBI.2018.07.001.

[21] M. Li and P. Vitányi, An introduction to Kolmogorov complexity and its applications, 3. Springer, 2008.

[22] H. H. Otu and K. Sayood, A new sequence distance measure for phylogenetic tree construction, Bioinformatics, 19, 16, 2122–2130, Nov. 2003. https://doi.org/10.1093/BIOINFORMATICS/BTG295.

[23] E. Delibaş and A. Arslan, A new feature vector model for alignment-free DNA sequence similarity analysis, Sigma Journal of Engineering and Natural Sciences, 40, 3, 610–619, Oct. 2022. https://doi.org/10.14744/sigma.2022.00065.

[24] J. P. Bao and R. Y. Yuan, A wavelet-based feature vector model for DNA clustering, Genet Mol Res, 14, 4, 19163–19172, Dec. 2015. https://doi.org/10.4238/2015.DECEMBER.29.26.

[25] G. Mendizabal-Ruiz, I. Román-Godínez, S. Torres-Ramos, R. A. Salido-Ruiz, H. Vélez-Pérez, and J. A. Morales, Genomic signal processing for DNA sequence clustering, PeerJ, 2018, 1, p. e4264, Jan. 2018. https://doi.org/10.7717/PEERJ.4264/SUPP-2.

[26] S. Dey, P. Ghosh, and S. Das, Positional difference and Frequency (PdF) based alignment-free technique for genome sequence comparison, J Biomol Struct Dyn, Oct. 2023. https://doi.org/10.1080/07391102.2023.2272748.

[27] S. Akbari Rokn Abadi, A. Mohammadi, and S. Koohi, A new profiling approach for DNA sequences based on the nucleotides' physicochemical features for accurate analysis of SARS-CoV-2 genomes, BMC Genomics, 24, 1, Dec. 2023. https://doi.org/10.1186/S12864-023-09373-7.

[28] M. K. Ganapathiraju, A. D. Mitchell, M. Thahir, K. Motwani, and S. Ananthasubramanian, Suite of tools for statistical N-gram language modeling for pattern mining in whole genome sequences, J Bioinform Comput Biol, 10, 6, Dec. 2012. https://doi.org/10.1142/S0219720012500163.

[29] H. U. Osmanbeyoglu and M. K. Ganapathiraju, N-gram analysis of 970 microbial organisms reveals presence of biological language models, BMC Bioinformatics, 12, p. 12, Jan. 2011. https://doi.org/10.1186/1471-2105-12-12.

[30] M. R. Kantorovitz, G. E. Robinson, and S. Sinha, A statistical method for alignment-free comparison of regulatory sequences, Bioinformatics, 23, 13, i249–i255, Jul. 2007. https://doi.org/10.1093/BIOINFORMATICS/BTM211.

[31] K. Song, J. Ren, G. Reinert, M. Deng, M. S. Waterman, and F. Sun, New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing, Brief Bioinform, 15, 3, 343–353, May 2014. https://doi.org/10.1093/BIB/BBT067.

[32] H.-H. Huang and C. Yu, Clustering DNA sequences using the out-of-place measure with reduced n-grams, J Theor Biol, 406, 61–72, 2016. https://doi.org/https://doi.org/10.1016/j.jtbi.2016.06.029.

[33] M. S. Nawaz, P. Fournier-Viger, M. Aslam, W. Li, Y. He, and X. Niu, Using alignment-free and pattern mining methods for SARS-CoV-2 genome analysis, Applied Intelligence, 53, 19, 21920–21943, Oct. 2023. https://doi.org/10.1007/S10489-023-04618-0/TABLES/13.

[34] T. Wang, Z. G. Yu, and J. Li, CGRWDL: alignment-free phylogeny reconstruction method for viruses based on chaos game representation weighted by dynamical language model, Front Microbiol, 15, p. 1339156, Mar. 2024. https://doi.org/10.3389/FMICB.2024.1339156/BIBTEX.

[35] B. Morgenstern, J. Söding, C. Bleidorn, A. Sturm, J. de Vries, and F. Manea, Alignment-free Phylogenetic Placement and its Applications, Feb. 2023. https://doi.org/10.53846/GOEDISS-9762.

[36] J. S. Almeida, J. A. Carriço, A. Maretzek, P. A. Noble, and M. Fletcher, Analysis of genomic sequences by Chaos Game Representation, Bioinformatics, 17, 5, 429–437, May 2001. https://doi.org/10.1093/BIOINFORMATICS/17.5.429.

[37] S. Safoury and W. Hussein, Enriched DNA strands classification using CGR images and convolutional neural network, ACM International Conference Proceeding Series, 87–92, Oct. 2019. https://doi.org/10.1145/3369166.3369176.

[38] K. Dick and J. R. Green, Chaos Game Representations Deep Learning for Proteome-Wide Protein Prediction, Proceedings - IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE 2020, 115–121, Oct. 2020. https://doi.org/10.1109/BIBE50027.2020.00027.

[39] R. Rizzo, A. Fiannaca, M. La Rosa, and A. Urso, Classification experiments of DNA sequences by using a deep neural network and chaos game representation, ACM International Conference Proceeding Series, 1164, 222–228, Jun. 2016. https://doi.org/10.1145/2983468.2983489.

[40] K. Zheng, Z. H. You, J. Q. Li, L. Wang, Z. H. Guo, and Y. A. Huang, ICDA-CGR: Identification of circRNA-disease associations based on Chaos Game Representation, PLoS Comput Biol, 16, 5, May 2020. https://doi.org/10.1371/JOURNAL.PCBI.1007872.

[41] C. Sravani, P. Pavani, G. Y. Vybhavi, G. Ramesh, A. Farman, and L. Venkareswara Reddy, Decoding the Human Genome: Machine Learning Techniques for DNA Sequencing Analysis, E3S Web of Conferences, 430, Oct. 2023. https://doi.org/10.1051/E3SCONF/202343001067.

[42] A. Yang, W. Zhang, J. Wang, K. Yang, Y. Han, and L. Zhang, Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA, Front Bioeng Biotechnol, 8, Sep. 2020. https://doi.org/10.3389/FBIOE.2020.01032.

[43] B. A. Bredesen and M. Rehmsmeier, DNA sequence models of genome-wide Drosophila melanogaster Polycomb binding sites improve generalization to independent Polycomb Response Elements, Nucleic Acids Res, 47, 15, 7781–7797, Sep. 2019. https://doi.org/10.1093/NAR/GKZ617.

[44] S. Das, A. Das, D. K. Bhattacharya, and D. N. Tibarewala, A new graph-theoretic approach to determine the similarity of genome sequences based on nucleotide triplets, Genomics, 112, 6, 4701–4714, Nov. 2020. https://doi.org/10.1016/J.YGENO.2020.08.023.

[45] 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, 3–4, 134–142, Oct. 2016. https://doi.org/10.1016/J.YGENO.2016.08.002.

[46] T. Hoang, C. Yin, H. Zheng, C. Yu, R. Lucy He, and S. S. T. Yau, A new method to cluster DNA sequences using Fourier power spectrum, J Theor Biol, 372, 135–145, May 2015. https://doi.org/10.1016/J.JTBI.2015.02.026.

[47] D. Quan, N. Nguyen, L. Xing, P. Dong, T. Le, and L. Lin, A graph-theoretical approach to DNA similarity analysis, bioRxiv, p. 2021.08.05.455342, Aug. 2021. https://doi.org/10.1101/2021.08.05.455342.

[48] X. Jin et al., A novel DNA sequence similarity calculation based on simplified pulse-coupled neural network and Huffman coding, Physica A: Statistical Mechanics and its Applications, 461, 325–338, Nov. 2016. https://doi.org/10.1016/J.PHYSA.2016.05.004.

[49] E. Delibaş, A. Arslan, A. Şeker, and B. Diri, A novel alignment-free DNA sequence similarity analysis approach based on top-k n-gram match-up, J Mol Graph Model, 100, p. 107693, Nov. 2020. https://doi.org/10.1016/j.jmgm.2020.107693.

[50] R. Dong, L. He, R. L. He, and S. S. T. Yau, A novel approach to clustering genome sequences using inter-nucleotide covariance, Front Pharmacol, 10, FEB, p. 423682, Apr. 2019. https://doi.org/10.3389/FGENE.2019.00234/BIBTEX.

[51] H. J. Jeffrey, Chaos game representation of gene structure., Nucleic Acids Res, 18, 8, p. 2163, Apr. 1990. https://doi.org/10.1093/NAR/18.8.2163.

[52] F. Zhuang et al., A Comprehensive Survey on Transfer Learning, Proceedings of the IEEE, 109, 1, 43–76, Jan. 2021. https://doi.org/10.1109/JPROC.2020.3004555.

[53] S. Eskandari, A. Eslamian, and Q. Cheng, Comparative Analysis of Transfer Learning Models for Breast Cancer Classification, Aug. 2024. https://doi.org/10.1109/AIC61668.2024.10731032.

[54] K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition, 2016. Accessed: Oct. 16, 2024. [Online]. Available: http://image-net.org/challenges/LSVRC/2015/

[55] F. Hong, D. W. L. Tay, and A. Ang, Intelligent Pick-and-Place System Using MobileNet, Electronics 2023, 12, Page 621, 12, 3, p. 621, Jan. 2023. https://doi.org/10.3390/ELECTRONICS12030621.

[56] J. Ren et al., Alignment-Free Sequence Analysis and Applications, Annu Rev Biomed Data Sci, 1, p. 93, Jul. 2018. https://doi.org/10.1146/ANNUREV-BIODATASCI-080917-013431.

[57] K. Tamura, G. Stecher, and S. Kumar, MEGA11: Molecular Evolutionary Genetics Analysis Version 11, Mol Biol Evol, 38, 7, 3022–3027, Jun. 2021. https://doi.org/10.1093/MOLBEV/MSAB120.