Positional frequency chaos game representation for machine learning-based classification of crop lncRNAs

Athanasios Papastathopoulos-Katsaros*†, Zhandong Liu*†
*Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA
{athanasios.papastathopoulos-katsaros, zhandonl}@bcm.edu
†Data Science Center, Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX, USA

***Abstract*—Alignment-based methods are fundamental for sequence comparison but are often computationally prohibitive for large-scale genomic analyses. This limitation has spurred the development of quicker, alignment-free alternatives, such as k-mer analysis, which are crucial for studying long non-coding ribonucleic acids (lncRNAs) in plants. These lncRNAs play critical roles in regulating gene expression at both the epigenetic and transcriptomic levels. However, existing alignment-free approaches typically lose positional information, which can be vital for achieving accurate classification. We propose positional frequency chaos game representation (PFCGR), a novel encoding that improves the traditional frequency chaos game representation (FCGR) by incorporating four statistical moments of k-mer positions: mean, standard deviation, skewness, and kurtosis. This creates a multi-channel image representation of genomic sequences, enabling machine learning models such as Logistic Regression, Random Forests, and Convolutional Neural Networks to classify plant lncRNAs directly from raw genomic sequences. Tested on seven major crop species, our PFCGR-based classifiers achieve classification accuracies comparable to or exceeding those of the computationally intensive DNABERT-based model [1], while requiring 80% to 95% less computational time. These results demonstrate PFCGR's potential as an efficient and accurate tool for plant lncRNA identification, as well as its ability to facilitate large-scale computational studies in genomics.**

***Index Terms*—long non-coding RNA, genomic sequence classification, k-mer analysis, Frequency chaos game representation**

I. Introduction

Alignment-based methods are essential for analyzing biological sequences but require significant computational resources. This need has accelerated the development of faster, alignment-free options, like k-mer analysis, which are vital for the study of different biological organisms. In plant biology, alignment-free methods show great promise for identifying long non-coding RNAs (lncRNAs) in crop species, offering significant advantages over traditional alignment-based approaches.

LncRNAs, defined as RNA molecules exceeding 200 nucleotides in length, are a critical component of the plant transcriptome with no coding potential but significant regulatory functions [2]–[4]. They are transcribed by RNA Polymerase II and are often polyadenylated, though significant exceptions exist. Plant genomes, unlike those of yeast and mammals, also utilize the plant-specific RNA polymerases Pol IV and Pol V to produce distinct lncRNAs, some of which are non-polyadenylated. These include upstream non-coding transcripts (UNTs) in *Arabidopsis* that are analogous to mammalian promoter upstream transcripts (PROMPTs), suggesting a nuanced regulatory landscape [5].

LncRNAs are involved in a myriad of regulatory roles across various stages of plant development and are differentially expressed across tissues, reflecting their specialized functions. Notably, lncRNAs influence key developmental processes, such as the timing of flowering [6], [7], and are crucial for plants' resilience to environmental stresses [8]–[12].

In terms of functionality, lncRNAs modulate gene expression through a variety of mechanisms, including epigenetic modifications [13]–[15]. For instance, in rice, the lncRNA LAIR notably increases grain yield by upregulating the LRK gene cluster, which leads to larger and more numerous panicles [16]. Another lncRNA, Ef-cd, sourced from the antisense strand of the OsSOC1 locus, significantly shortens the maturity period for late-maturing rice varieties without reducing yield, showcasing lncRNAs' ability to fine-tune plant growth under cultivation constraints [17]. These advancements in lncRNA research are paving the way for significant contributions to agricultural biotechnology, including the development of genetically engineered crops with increased yields, improved resistance to pests and diseases, and greater environmental stress tolerance. Thus, understanding and predicting lncRNAs remain crucial in modern plant science and biotechnology. For additional implications of lncRNAs in physiological processes, we refer the reader to [18].

To address the challenges of existing methods, we introduce a novel approach for classifying plant lncRNAs based on genomic sequences, using machine-learning techniques such as Logistic Regression (LR), Random Forest (RF), and Convolutional Neural Networks (CNN). Our approach uses Frequency Chaos Game Representation (FCGR) and our innovative improved version, "positional FCGR" (PFCGR), which incorporates the following statistical features of the positions of all the k-mers of the sequences: mean, standard deviation, skewness, and kurtosis. Therefore, the aims of this study are: (i) to address FCGR's limitation of losing positional information; (ii) to develop and evaluate PFCGR-

based machine-learning models for the identification of plant lncRNAs directly from genomic sequences; (iii) to compare the accuracy and computational efficiency of PFCGR-based models against the very accurate and much more complex DNABERT-based method across seven crop species.

II. Related work

Bioinformatics pipelines for lncRNA prediction often combine custom scripts and tools like CPC2 [19], LncMachine [20], CPAT [21], PlncPRO [22], PLEK [23] and PLEKv2 [24] to filter and estimate the coding potential of transcripts from transcriptome datasets. These methods typically involve manually defining features from transcripts, which are then used in machine-learning models to classify transcripts as mRNA or lncRNA [25], [26]. However, the reliance on handcrafted features can introduce bias and affect classification accuracy [27], [28], while the training of models depends heavily on transcriptome data from sources like CANTATAdb [29], PLncDB [2], GreeNC [30], [31] and AlnC [32]. For this reason, a shift towards models that utilize genomic references could improve the detection of lncRNAs, particularly given their tissue- or stage-specific low expression rates.

One such work [1] is based on DNABERT [33], a specialized adaptation of the language BERT model [34] tailored for DNA sequences. It addresses the binary classification task of distinguishing true lncRNA from synthetic lncRNA sequences, as explained in more detail in subsequent sections. DNABERT is capable of performing sequence classification by encoding DNA sequences into k-mer tokens, which allows it to capture intricate biological features. This is achieved through the transformer architecture and the attention mechanism [35], which unfortunately comes with a significant computational cost and the black-box nature of the model. While other recent transformer-based approaches like LncRNA-BERT [36] and LoRA-BERT [37] have demonstrated strong performance for lncRNA classification, they present similar computational challenges and are primarily trained on human data with limited evaluation on plant species. For these reasons, we only compare our method against [1] to ensure a fair and equal comparison on the exact same dataset. For a more extensive discussion on related lncRNA work, refer to [38] and [1].

Lastly, we acknowledge that the loss of positional information, the main limitation of FCGR, has been a subject of previous research. Methods have been developed to decode positional information from FCGR images using techniques such as wavelet transformation (Frequency Chaos Game Signal [39]) or by analyzing correlations among k-mers [40]. However, these approaches differ significantly from ours, as none have implemented a similar strategy of using statistical features of k-mer positions to capture and utilize this information. By introducing PFCGR, we propose a distinct and novel way to address this long-standing limitation, and we choose to benchmark it to the challenging task of crop lncRNA classification.

III. Methodology

A. FCGR encoding

The Chaos Game Representation (CGR), originally used to construct fractals from random inputs, was later adapted to encode DNA sequences [41]. For readers unfamiliar with CGR, a simple illustration is provided in Figure 1. In this work, we use the Frequency CGR (FCGR), a discrete matrix equivalent of CGR that counts k-mer frequencies at each position. Unlike traditional encoding methods such as one-hot encoding, FCGR encodes nucleotide sequences into images/numerical data, standardizing sequences of varying lengths into images of the same size. For further reading, see [42].

In a nutshell, the idea behind FCGR is to map k-mers onto a unit square by recursively dividing the square into smaller regions according to the sequence of bases in each k-mer. Each base (A, C, G, T) determines which quadrant of the current region to select, and this subdivision continues for all positions in the k-mer, resulting in a unique location within a $2^k \times 2^k$ grid corresponding to that k-mer. The FCGR image is constructed by plotting each k-mer's frequency on that grid, which is normalized to compute pixel intensities. In this study, k ranges from 3 to 6. Figure 2 shows FCGR images from the *Zea Mays* dataset, highlighting some key regions that could potentially aid in sequence classification.

FCGR has shown success in various applications, such as detecting COVID-19 [43], [44], distinguishing human coding and non-coding regions [45], and predicting antimicrobial resistance [46]. It has also been shown to outperform traditional one-hot encoding in bacterial 16S rRNA classification [47]. Recently, a new FCGR-based representation was developed based on ranked frequencies instead of raw, which was integrated with Transformers and showed improved performance, particularly in low-coverage genomic data [48]. Lastly, reference [49] applied FCGR in unsupervised clustering of DNA sequences using CNN combined with twin contrastive learning, demonstrating high clustering accuracy across viral genomes and other genomic datasets.

B. Positional FCGR encoding

One major drawback of FCGR is that it does not retain positional information. For this reason, we extend the FCGR method by adding positional information (PFCGR) through statistical measures of the k-mer positions, including mean, standard deviation, skewness, and kurtosis. These are computed using Welford's online algorithm [50], reducing memory requirements and computational overhead. Each statistical measure generates a separate $2^k \times 2^k$ image or is kept as tabular data, as explained below. The intensity of each pixel corresponds to the statistical value of the respective k-mer and is normalized by min-max normalization. This added positional context improves FCGR by capturing more complex patterns in genomic sequences, where k-mer placement matters.

There are two ways to encode these added statistical measures. On the one hand, we can organize the extracted features into tabular data. For FCGR, each column represents a k-mer

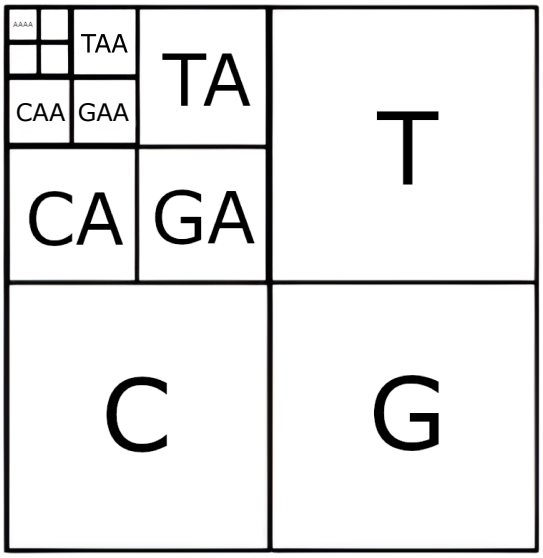

Fig. 1. **Illustration of the CGR algorithm**: The square is initially divided into four equal parts, each representing one of the four nucleotides found in DNA: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). Each quadrant is assigned to one nucleotide. Each of these initial quadrants is further subdivided into four smaller squares. This division process respects the nucleotide of the larger square from which it is derived. For example, the square for Adenine (A) is subdivided into AA, TA, CA, and GA. This method applies recursively to all quadrants. With each iteration, the squares become increasingly specific to particular sequences of nucleotides. For instance, the top left quadrant, after the second iteration, shows further subdivisions like AAA, TAA, CAA, and GAA, each representing a specific 3-mer, and the grid in the top-left corner of this figure represents the 4-mer AAAA. Finally, to map a DNA sequence onto this square, we begin at the center and move to each subsequent quadrant based on the next nucleotide in the sequence. The position within the square becomes more precise with each nucleotide added, tracing a path through increasingly smaller quadrants.

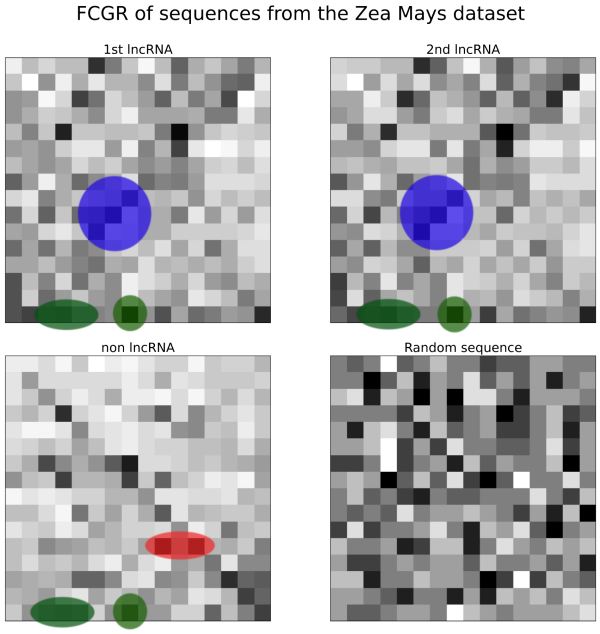

Fig. 2. FCGR images with k=4 are displayed for four sequences: two lncRNAs, one non-lncRNA, and one random sequence. Green circles highlight patterns common to both lncRNAs and the non-lncRNA, indicating shared genomic motifs. Blue circles indicate patterns unique to lncRNAs, suggesting motifs specific to lncRNA. Red circles highlight patterns exclusive to the non-lncRNA sequence. Not all patterns are marked to maintain visual clarity in the images.

(e.g., "AAT"), and each cell contains its frequency in each sequence of the dataset. For PFCGR, the columns are labeled with both the k-mer and the statistical descriptors (Frequency, Mean, Standard Deviation, Skewness, Kurtosis). The data in this format can then be inputted into LR and RF models. On the other hand, the data in image format (which can be thought of as either five images or one image with five channels) can be inputted into CNN models, which are effective in identifying patterns in the generated images, even across channels. For missing k-mers, we assign zero frequencies in the FCGR matrix. We experimented with different imputation strategies for handling the rest of the missing statistics in the PFCGR matrices, including filling with zeros, using -1 as a sentinel value, and adding an additional binary masking channel indicating presence or absence of a k-mer. Since all approaches yielded similar performance, we opted to use zeros for simplicity. This issue does not affect the LR and RF methods, as the tabular format does not contain those columns.

C. Model architectures and training

In selecting our classifiers, we prioritized simplicity and computational efficiency to establish a proof of concept for PFCGR's effectiveness, rather than pursuing complex architectures for absolute maximum performance. This study represents the first application of PFCGR, and specifically using PFCGR for lncRNA classification, making it essential to demonstrate the method's viability using straightforward models. Therefore, LR and RF were chosen for their low computational cost and ease of computing feature importance, while CNNs were included for their ability to capture complex patterns in image-based data. Our CNN architecture for FCGR includes, in this order, a convolutional layer, followed by a pooling layer, a fully-connected hidden layer, and a final sigmoid output neuron. All other layers use ReLU activation functions. The difference for PFCGR is that the five images are processed separately by one convolutional layer before merging their outputs to the rest of the neural network, which has the same architecture as the single-image FCGR setup. This is intended to capture unique features of each image before further processing.

Each species has a dedicated classifier, individually trained in species-specific datasets, to accurately distinguish true lncRNA from synthetic non-lncRNA sequences. For LR, we applied L2 regularization, which demonstrated more robust performance compared to LASSO. For RF, we optimized the number of trees and their maximum depths using grid search. Specifically, the number of trees was varied between 250 and 1250, while the maximum depth ranged from unrestricted to a limit ranging from 10 to 40 to reduce overfitting. Lastly, for the CNN, the number of filters and filter sizes in the convolutional layers were optimized via grid search. The max pooling

layer's window size and stride were also tuned, along with the number of neurons in the fully connected layer. Specifically, the number of filters was varied between 16 and 128, while kernel sizes ranged from 1×1 to 8×8. Pooling window sizes and strides were explored from 2×2 up to 6×6 to identify the optimal configuration. The model was trained using binary cross-entropy loss with the Adam optimizer (learning rate = 0.001), batch sizes ranging from 8 to 32, and for 7 to 10 epochs. To prevent overfitting and ensure reliable performance, stratified 5-fold cross-validation was applied to all models. The CNN models were trained on an NVIDIA A30 GPU to reduce training time.

TABLE I
MEDIAN SEQUENCE LENGTHS AND LABEL COUNTS FOR EACH CROP SPECIES.

Species	Median seq. length	lncRNA labels	Non-lncRNA labels
A. thaliana	1,514	3,467	3,464
B. napus	1,854	8,544	7,442
B. oleracea	1,832	5,350	4,997
B. rapa	2,070	7,075	6,882
G. max	2,143	2,320	2,288
O. sativa	1,943	2,335	2,334
Z. mays	1,978	8,889	8,747

D. Computational complexity of PFCGR compared to DNABERT

To compare the computational complexity of PFCGR to DNABERT's, we consider both feature extraction and classifier complexity, using RF for illustration. For a sequence of length N and a fixed k-mer size, feature extraction involves iterating over the sequence once (sliding a window of length k), with each k-mer requiring $\mathcal{O}(1)$ time because of hashing and Welford's online algorithm ability to compute the statistical measures on-the-fly. Thus, the total complexity for generating the PFCGR feature vector is $\mathcal{O}(N)$. For M training sequences, each with a feature vector of length $\approx 5 \cdot 4^k$, the RF complexity with T trees and tree depth $\mathrm{d_{tree}}$ is $\mathcal{O}\big(M \times (5 \cdot 4^k) \times T \times \mathrm{d_{tree}}\big)$. The term 4^k represents the maximum possible number of unique k-mers. For small k, the feature vector size is manageable. For large k, the number of unique k-mers present is typically much smaller than 4^k. In this case, the number of unique k-mers is bounded by the sequence length and therefore, while the theoretical upper bound is exponential in k, the practical complexity remains tractable for both small and large values of k.

In contrast, DNABERT, has a computational complexity of $\mathcal{O}(L \times (N^2 \times d + N \times d^2))$ for a sequence of length N, where d is the embedding size and L is the number of feed-forward layers. The self-attention mechanism introduces a quadratic complexity on N, which makes it significantly more computationally demanding than PFCGR. This is particularly problematic for the long sequences in our study (median length 1500-2200 nucleotides, see Table I), as DNABERT's performance and efficiency degrade with increasing sequence length due to the high computational cost and challenges in capturing long-range dependencies.

A direct comparison of training times between PFCGR-based RF and CNN models relative to DNABERT is provided in Table II. Our approaches are about 80% to 95% faster, with the exact speedup varying by classifier and dataset. It is important to acknowledge that this comparison is complex and our PFCGR implementation has room for further optimization. DNABERT is designed to exploit high-performance hardware, taking advantage of large-scale GPUs, mixed-precision training, and optimized matrix operations. In contrast, our current PFCGR implementation does not fully capitalize on these advanced GPU optimizations. These discrepancies in hardware and software optimizations can make a direct, apples-to-apples comparison of raw execution times challenging. Even with these factors making DNABERT's times appear closer to our current implementation than they are in reality, our results still convincingly demonstrate the superior computational efficiency of the PFCGR-based methods.

TABLE II
COMPARISON OF THE TRAINING TIMES (IN MINUTES) FOR DNABERT, RF WITH PFCGR, AND CNN WITH PFCGR. ALL MODELS WERE TRAINED ON THE COMPLETE DATASET USING THEIR RESPECTIVE OPTIMAL PARAMETERS (FOR DNABERT WE REFER THE READER TO [1]) USING $k = 6$. DNABERT AND CNN MODELS UTILIZED AN NVIDIA A30 GPU, WHILE THE RF MODEL WAS TRAINED ON AN INTEL XEON SILVER 4314 CPU. CROSS-VALIDATION AND INFERENCE TIMES ARE EXCLUDED.

Species	RF w/ PFCGR	CNN w/ PFCGR	DNABERT
A. thaliana	13	3	54
B. napus	34	9	180
B. oleracea	24	6	120
B. rapa	38	9	216
G. max	24	8	96
O. sativa	20	6	84
Z. mays	45	13	288

E. Datasets and experimental setup

We used the same data set as described by [1]. In brief, the lncRNA datasets for all crop species (*Arabidopsis thaliana*, *Brassica napus*, *Brassica oleracea*, *Brassica rapa*, *Glycine max*, *Oryza sativa*, and *Zea mays*) were sourced from CantataDB 2.0 [29] and PLncDB 2.0 [2]. To enrich the dataset, flanking regions were added both upstream and downstream of each lncRNA sequence to prevent the classifier from merely identifying broad genomic features. A control dataset was constructed by extracting random genomic stretches, explicitly excluding the lncRNA sequences and their flanking regions. This approach increases the complexity of the classification task compared to the simpler mRNA versus lncRNA, encouraging the model to learn discriminative features specific to lncRNA regions within their genomic context. For a comprehensive explanation of the dataset construction, we direct the reader to [1].

In our experiments, we performed binary classification between lncRNA and non-lncRNA sequences separately for each species. This choice was motivated by intrinsic differences in the datasets, such as variations in sequence length distributions, nucleotide composition, and class imbalances. Training separate classifiers per species allows them to better

capture features unique to each genome. Table I summarizes the median length of the sequences for each species, showing the scale of the datasets, as well as the class distribution for each dataset. The class distribution varies between species, reflecting differences in available annotated sequences and the size of the control sets. For this reason, to robustly evaluate model performance, we employed stratified 5-fold cross-validation. Each fold was used once as the test set, while the remaining four folds served as training data. This approach ensures that the evaluation is reliable and that each fold contains representative samples of both classes.

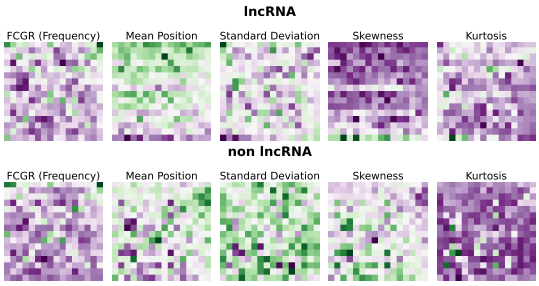

Fig. 3. One lncRNA and one non-lncRNA sequence are shown using the standard FCGR and the positional FCGR. The top row displays the lncRNA, while the bottom row depicts the non-lncRNA. Despite very similar FCGRs for both sequences, PFCGRs reveal differences, emphasizing their discriminative power. The color scale ranges from dark purple (indicating low values) to dark green (indicating high values), with all values normalized across the statistics using min-max normalization to ensure consistent comparison.

Metrics: As in [1], model effectiveness was measured using accuracy, precision, recall, and F1 score. True positives, false negatives, and false positives represent correctly identified lncRNAs, misclassified lncRNAs, and falsely identified lncRNAs, respectively.

IV. Results and discussion

We conducted two sets of experiments to evaluate our method. In the first set, we used LR, RF, and CNN for binary classification between non-lncRNA and lncRNA sequences in the seven species. We applied FCGR and PFCGR and optimized the k-mer size for each species, ranging from 3-mers to 6-mers, to evaluate the impact of the additional statistical features and their importance. This set also included a comparison with the DNABERT model [1]. In the second set of experiments, we aimed to improve classification accuracy by applying RF and CNN across all k-mer sizes combined (k=3 to k=6), to further improve the model's performance and provide a better comparison with DNABERT. This set of experiments served to support the findings of the first.

A. Performance evaluation using the optimal k-mers

To evaluate whether PFCGR boosts performance over the traditional FCGR method, we apply LR, RF, and CNN, using various k-mer sizes to determine the best accuracy. The range of k-mer lengths considered (k ranging from 3 to 6) reflects a common practice in k-mer-based approaches: smaller k values tend to be overly frequent and less informative, while k values greater than 6 become increasingly rare, resulting in sparse and potentially noisy features that can skew model performance. Thus, the selected range is more balanced.

TABLE III
Evaluation metrics for species-specific LR, RF, and CNN models using FCGR and PFCGR, and a comparison with DNABERT, for the binary classification of lncRNAs across the seven crop species. The table reports the mean Accuracy, Precision, Recall, and F1 score from stratified 5-fold cross-validation, using the optimal k-mer size for each method.

Species	Method	Acc.	Prec.	Rec.	F1
A. thaliana	LR - FCGR	0.56	0.56	0.55	0.56
	LR - PFCGR	0.59	0.59	0.60	0.60
	RF- FCGR	0.61	0.61	0.64	0.62
	RF - PFCGR	0.65	**0.67**	0.60	0.63
	CNN - FCGR	0.62	0.61	**0.68**	0.64
	CNN - PFCGR	**0.66**	0.66	0.65	**0.65**
	DNABERT	0.65	0.65	0.65	**0.65**
B. napus	LR - FCGR	0.55	0.54	1.00	0.70
	LR - PFCGR	0.64	0.66	0.68	0.67
	RF- FCGR	0.66	0.64	0.81	0.72
	RF - PFCGR	0.70	0.69	0.80	0.74
	CNN - FCGR	0.66	0.64	**0.81**	0.72
	CNN - PFCGR	0.70	0.69	0.80	**0.74**
	DNABERT	**0.75**	**0.75**	0.74	**0.74**
B. oleracea	LR - FCGR	0.59	0.56	0.94	0.70
	LR - PFCGR	0.63	0.63	0.66	0.65
	RF- FCGR	0.67	0.64	0.78	0.71
	RF - PFCGR	0.71	0.71	0.74	0.73
	CNN - FCGR	0.69	0.68	0.76	0.72
	CNN - PFCGR	0.68	0.65	**0.85**	**0.74**
	DNABERT	**0.74**	**0.75**	0.74	**0.74**
B. rapa	LR - FCGR	0.55	0.55	0.64	0.59
	LR - PFCGR	0.55	0.55	0.55	0.55
	RF- FCGR	0.58	0.58	0.65	0.61
	RF - PFCGR	**0.60**	**0.60**	0.67	0.63
	CNN - FCGR	0.59	0.59	0.66	0.62
	CNN - PFCGR	0.56	0.55	**0.80**	**0.65**
	DNABERT	0.58	0.58	0.58	0.58
G. max	LR - FCGR	0.50	0.50	1.00	0.67
	LR - PFCGR	0.62	0.63	0.63	0.63
	RF- FCGR	0.70	0.68	0.77	0.72
	RF - PFCGR	0.72	0.70	0.78	**0.74**
	CNN - FCGR	0.68	0.64	0.83	0.72
	CNN - PFCGR	0.70	0.66	**0.84**	**0.74**
	DNABERT	**0.73**	**0.73**	0.73	0.73
O. sativa	LR - FCGR	0.54	0.55	0.48	0.51
	LR - PFCGR	0.57	0.57	0.57	0.57
	RF- FCGR	0.61	0.61	0.61	0.61
	RF - PFCGR	**0.63**	**0.64**	0.60	0.62
	CNN - FCGR	0.62	0.62	0.65	**0.63**
	CNN - PFCGR	0.57	0.56	**0.67**	0.61
	DNABERT	0.62	0.62	0.62	0.62
Z. mays	LR - FCGR	0.72	0.69	0.80	0.74
	LR - PFCGR	0.72	0.72	0.74	0.73
	RF- FCGR	0.79	0.75	**0.88**	0.81
	RF - PFCGR	0.79	0.75	**0.88**	0.81
	CNN - FCGR	0.80	0.77	0.86	0.81
	CNN - PFCGR	0.81	0.78	**0.88**	**0.83**
	DNABERT	**0.83**	**0.84**	0.84	**0.83**

Before presenting our results, we first want to address any skepticism regarding the utility of PFCGR over regular FCGR. To do that, we present an example from the *Zea Mays* dataset in Figure 3. As it can be seen, while lncRNA and non-lncRNA sequences exhibit similar FCGR patterns,

TABLE IV
A COMPARISON OF THE ROBUSTNESS OF FCGR AND PFCGR FOR DIFFERENT K-MER LENGTHS BASED ON ACCURACY AND F1 SCORE METRICS. ALL RESULTS ARE COMPUTED USING RF AND THE *A. thaliana* DATASET AND REPRESENT THE MEAN PERFORMANCE OBTAINED FROM STRATIFIED 5-FOLD CROSS-VALIDATION.

k	Metric	FCGR	PFCGR
3	Accuracy	0.61	0.65
4	Accuracy	0.61	0.64
5	Accuracy	0.61	0.64
6	Accuracy	0.59	0.65
3	F1 - score	0.62	0.63
4	F1 - score	0.62	0.63
5	F1 - score	0.60	0.62
6	F1 - score	0.57	0.63

TABLE V
MEAN ABSOLUTE CORRELATIONS (MEAN ABS COR.) AND THEIR STANDARD DEVIATIONS (ST.D. OF THE COR.) AMONG ALL FEATURES FOR DIFFERENT K. ALL RESULTS COMPUTED USING RF AND THE *A. thaliana* DATASET.

k	Statistic	FCGR	PFCGR
3	Mean abs cor.	0.20	0.06
3	St.d. of the cor.	0.15	0.08
4	Mean abs cor.	0.10	0.03
4	St.d. of the cor.	0.10	0.05
5	Mean abs cor.	0.05	0.04
5	St.d. of the cor.	0.06	0.05
6	Mean abs cor.	0.03	0.06
6	St.d. of the cor.	0.04	0.06

incorporating additional statistical measures reveals significant differences. The four new metrics clearly distinguish between the two sequence types, suggesting that the PFCGR method offers more information, potentially improving lncRNA classification accuracy.

In Table III, we summarize our results obtained with the optimal k-mer lengths for each species, where "optimal" refers to the k-mer size within the range 3 to 6 that yields the highest classification accuracy for that species. Specifically, a k-mer of length 3 works best for *A. thaliana*, *B. napus*, *B. oleracea*, *B. rapa*, and *O. sativa*, while lengths of 4 and 6 are optimal for *G. max* and *Z. mays* respectively. All results reported are the averages obtained from stratified 5-fold cross-validation to ensure robustness of the models. Our results demonstrate that PFCGR improves the classification performance for nearly all species compared to FCGR. For certain ones, RF and CNN models achieve results comparable to or better than DNABERT, whose results were fine-tuned separately for each species to optimize performance. In particular, the higher recall across the board suggests that our models are better at identifying lncRNAs, which is critical in most applications.

Another important observation is that despite logistic regression's tendency to predict an excessive number of lncRNA cases, which inflates recall, PFCGR moderates these predictions leading to more balanced performance metrics. We also notice some disparities between the CNN and RF models when it comes to the accuracy and F1 score metrics, particularly evident in the PFCGR scenario. In slightly imbalanced datasets like ours, there is no clear answer as to which metric we should use, therefore both the CNN and RF models remain accurate, indicating that for most of the cases PFCGR improves the standard FCGR by providing a more robust classification method for lncRNAs without significantly increasing the computational cost. However, we do realize that in a few cases, PFCGR may actually degrade the performance of the CNN model compared to FCGR, which we attribute to potential overfitting.

Before we move on to the next section, we want to mention a few more observations. The difference in performance between different k-mer lengths is smaller for PFCGR, as shown in Table IV for the example of *A. thaliana*. This suggests that PFCGR adds useful features that contribute to a more robust model. These features refer to the coordinates of FCGR/PFCGR in tabular format. Moreover, feature importance analysis revealed that, for both LR and RF models using PFCGR, the majority of the top 20 significant features were standard deviations and frequencies. This emphasizes the value of incorporating measures of k-mer dispersion along the sequence. To calculate feature importance, we used model-specific criteria: for the LR model, the absolute values of the regression coefficients served as a proxy, while for the RF model, we relied on the average decrease in Gini impurity to measure each individual contribution. Although a few skewness and kurtosis-related features were identified, they were fewer in number, aligning with expectations since they capture more specific distributional aspects and ought to be related to secondary effects. While feature importance analysis is also feasible for CNNs, it tends to be more elaborate and we did not see the need to pursue it here, as we expect the results to be similar to those obtained with LR and RF models.

Lastly, the features showed low correlations (Table V), indicating that they provide complementary information. To compute feature correlations, we calculated the absolute mean correlation and its standard deviation, using Pearson's correlation coefficient as a measure. This low correlation among features is critical, as it reduces the likelihood of multicollinearity, which can skew the model performance.

TABLE VI
COMPARISON OF CLASSIFICATION ACCURACY FOR SINGLE-K-MER MODELS, MULTI-K-MER MODELS (WITH k RANGING FROM 3 TO 6), AND THE BEST-PERFORMING DNABERT MODEL ACROSS SEVEN PLANT SPECIES. THE RESULTS REPRESENT THE MEAN PERFORMANCE OBTAINED FROM STRATIFIED 5-FOLD CROSS-VALIDATION.

	Single k		Multi k		Single k
Species	RF	CNN	RF	CNN	DNABERT
A. thaliana	0.66	0.66	0.68	**0.69**	0.65
B. napus	0.70	0.70	0.72	0.73	**0.75**
B. oleracea	0.71	0.68	0.72	0.70	**0.74**
B. rapa	0.60	0.56	**0.61**	0.60	0.58
G. max	0.70	0.70	0.73	**0.74**	0.73
O. sativa	0.63	0.57	**0.64**	0.62	0.62
Z. mays	0.81	0.81	0.80	0.82	**0.83**

B. Performance evaluation using all k-mers combined

In our earlier experiments, we observed that the performance did not improve consistently with increasing k-mer size, highlighting the complex relationship between k and model accuracy. To address this, we tested multiple k values and selected the best one, as performed in the DNABERT study. Building on this, we now integrate multi-scale k-mer statistics, spanning from k=3 to k=6. We use both RF and CNN models in this approach.

For the RF model, we maintain the same tabular format as before, but this time we incorporate multiple k-mers into a unified feature set. In the case of the CNN model, an ensemble approach was used. Each individual CNN processes a separate k-mer size (k=3, k=4, k=5, and k=6) with the same architecture as previously used, but with the last layer removed. Instead of the original output layer, the outputs from each model are flattened and passed through a fully connected layer (meta-model), and then through a final sigmoid neuron to generate the final prediction. The dimension of the fully-connected layer was optimized separately for each species using grid search, using values ranging from 32 to 512. This multi-scale integration approach, while non-trivial for DNABERT, is straightforward in our framework and effectively combines the complementary strengths of different k-mer sizes.

However, feature importance analysis reveals that when specific k-mers are significant, their related variants (e.g., subsets or extensions) may also appear important, indicating potential multicollinearity. While this could affect some classifiers, it is not a major issue for RF and CNN in practice.

Our results, summarized in Table VI, represent the average performance from stratified 5-fold cross-validation. The multi-k RF and CNN models consistently match or exceed DNABERT's performance, while being far less computationally demanding. In addition, our approaches demonstrate a notable advantage on smaller datasets, such as those for *A. thaliana*, *G. max* and *O. sativa*, as well as on datasets with longer sequences, including *B. rapa* and *G. max*. These findings are particularly encouraging given the challenges associated with large-scale genomic studies. Access to extensive datasets is often limited, and lncRNAs are known for their considerable length, which poses a well-documented challenge for LLM-based models like DNABERT, due to their context length limitations.

V. Conclusion and future work

This study presents a novel approach for extending FCGR with positional information (PFCGR) by incorporating statistical measures of k-mer positions, such as mean, standard deviation, skewness, and kurtosis. In this way, our method improves the standard FCGR by adding layers of positional information, providing a richer set of features for the classifiers. We combine this approach with machine-learning techniques, such as Logistic Regression (LR), Random Forest (RF) and Convolutional Neural Networks (CNN), to classify lncRNAs of seven crop species. Our method reduces computational time by 80% to 95% compared to the state-of-the-art DNABERT for this task, while maintaining competitive accuracy, and even in some instances surpassing it.

In addition, our experiments across multiple crop species illustrate the robustness and adaptability of our approach. Notably, the RF and the CNN have shown significant promise, leading to high accuracy and improved classification performance, particularly when using the multi-k feature sets. However, we note that applying our approach to other types of sequences and integrating more diverse datasets are necessary to validate and potentially improve the generalizability of our findings. Importantly, the PFCGR methodology is sequence-agnostic and can be readily extended to any DNA, RNA or protein, assuming the protein-adapted FCGR is used, sequence classification task, making it a versatile tool for diverse applications.

Lastly, we want to underline that this study not only advances the computational tools available for genomic research, but also contributes to our understanding of lncRNA functions and their regulatory roles in plants. By providing a more efficient and accurate tool for lncRNA identification, we pave the way for discoveries in plant biology, improve prospects for agricultural innovation, and accelerate plant breeding programs.

In future work, we can focus on exploring alternative neural network architectures, such as transformers, which could improve accuracy, especially with larger datasets. Additionally, incorporating gapped k-mers could improve generalizability and robustness by accommodating sequencing errors and variations. These strategies will further improve the accuracy and versatility of our method.

VI. Acknowledgments

This work was supported by the National Institutes of Health (NIH) under grant number 5R01HG011795 and by the Cancer Prevention and Research Institute of Texas (CPRIT) under grant number RP240131.

References

[1] M. F. Danilevicz, M. Gill, C. G. T. Fernandez, J. Petereit, S. R. Upadhyaya, J. Batley, M. Bennamoun, D. Edwards, and P. E. Bayer, "DNABERT-based explainable lncRNA identification in plant genome assemblies," *Comput. Struct. Biotechnol. J.*, vol. 21, 2023.

[2] J. Jin, P. Lu, Y. Xu, Z. Li, S. Yu, J. Liu, H. Wang, N.-H. Chua, and P. Cao, "PLncDB V2.0: a comprehensive encyclopedia of plant long noncoding RNAs," *Nucleic Acids Res.*, vol. 49, no. D1, 2021.

[3] H. Wang, P. J. Chung, J. Liu, I.-C. Jang, M. J. Kean, J. Xu, and N.-H. Chua, "Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in *Arabidopsis*," *Genome Res.*, vol. 24, no. 3, 2014.

[4] Y.-C. Zhang, J.-Y. Liao, Z.-Y. Li, Y. Yu, J.-P. Zhang, Q.-F. Li, L.-H. Qu, W.-S. Shu, and Y.-Q. Chen, "Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice," *Genome Biol*, vol. 15, no. 12, 2014.

[5] J. A. Chekanova, B. D. Gregory, S. V. Reverdatto, H. Chen, R. Kumar, T. Hooker, J. Yazaki, P. Li, N. Skiba, Q. Peng, J. Alonso, V. Brukhin, U. Grossniklaus, J. R. Ecker, and D. A. Belostotsky, "Genome-Wide High-Resolution Mapping of Exosome Substrates Reveals Hidden Features in the Arabidopsis Transcriptome," *Cell*, vol. 131, no. 7, 2007.

[6] H. Lee, S. J. Yoo, J. H. Lee, W. Kim, S. K. Yoo, H. Fitzgerald, J. C. Carrington, and J. H. Ahn, "Genetic framework for flowering-time regulation by ambient temperature-responsive miRNAs in Arabidopsis," *Nucleic Acids Res.*, vol. 38, no. 9, 2010.

[7] S. Waheed and L. Zeng, "The Critical Role of miRNAs in Regulation of Flowering Time and Flower Development," *Genes*, vol. 11, no. 3, 2020.

[8] F. Thiebaut, C. A. Rojas, C. Grativol, M. R. Motta, T. Vieira, M. Regulski, R. A. Martienssen, L. Farinelli, A. S. Hemerly, and P. C. Ferreira, "Genome-wide identification of microRNA and siRNA responsive to endophytic beneficial diazotrophic bacteria in maize," *BMC Genomics*, vol. 15, no. 1, 2014.

[9] F. Thiebaut, C. A. Rojas, K. L. Almeida, C. Grativol, G. C. Domiciano, C. R. C. Lamb, J. De Almeida Engler, A. S. Hemerly, and P. C. G. Ferreira, "Regulation of miR319 during cold stress in sugarcane," *Plant Cell & Environment*, vol. 35, no. 3, 2012.

[10] R. Salvador-Guirao, P. Baldrich, D. Weigel, I. Rubio-Somoza, and B. San Segundo, "The MicroRNA miR773 Is Involved in the *Arabidopsis* Immune Response to Fungal Pathogens," *MPMI*, vol. 31, no. 2, 2018.

[11] Q. Liu, T. Yang, T. Yu, S. Zhang, X. Mao, J. Zhao, X. Wang, J. Dong, and B. Liu, "Integrating Small RNA Sequencing with QTL Mapping for Identification of miRNAs and Their Target Genes Associated with Heat Tolerance at the Flowering Stage in Rice," *Front. Plant Sci.*, vol. 8, 2017.

[12] G. Hu, M. Hao, L. Wang, J. Liu, Z. Zhang, Y. Tang, Q. Peng, Z. Yang, and J. Wu, "The Cotton miR477- *CBP60A* Module Participates in Plant Defense Against *Verticillium dahlia*," *MPMI*, vol. 33, no. 4, 2020.

[13] J. B. Heo and S. Sung, "Vernalization-Mediated Epigenetic Silencing by a Long Intronic Noncoding RNA," *Science*, vol. 331, no. 6013, 2011.

[14] S. Guil and M. Esteller, "Cis-acting noncoding RNAs: friends and foes," *Nat Struct Mol Biol*, vol. 19, no. 11, 2012.

[15] M. C. D. O. Urquiaga, F. Thiebaut, A. S. Hemerly, and P. C. G. Ferreira, "From Trash to Luxury: The Potential Role of Plant LncRNA in DNA Methylation During Abiotic Stress," *Front. Plant Sci.*, vol. 11, 2021.

[16] Y. Wang, X. Luo, F. Sun, J. Hu, X. Zha, W. Su, and J. Yang, "Overexpressing lncRNA LAIR increases grain yield and regulates neighbouring gene cluster expression in rice," *Nat Commun*, vol. 9, no. 1, 2018.

[17] J. Fang, F. Zhang, H. Wang, W. Wang, F. Zhao, Z. Li, C. Sun, F. Chen, F. Xu, S. Chang, L. Wu, Q. Bu, P. Wang, J. Xie, F. Chen, X. Huang, Y. Zhang, X. Zhu, B. Han, X. Deng, and C. Chu, "*Ef-cd* locus shortens rice maturity duration without yield penalty," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 116, no. 37, 2019.

[18] H.-L. V. Wang and J. A. Chekanova, "Long Noncoding RNAs in Plants," in *Long Non Coding RNA Biology*, M. Rao, Ed. Singapore: Springer Singapore, 2017, vol. 1008, pp. 133–154, series Title: Advances in Experimental Medicine and Biology.

[19] Y.-J. Kang, D.-C. Yang, L. Kong, M. Hou, Y.-Q. Meng, L. Wei, and G. Gao, "CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features," *Nucleic Acids Res.*, vol. 45, no. W1, 2017.

[20] H. B. Cagirici, S. Galvez, T. Z. Sen, and H. Budak, "LncMachine: a machine learning algorithm for long noncoding RNA annotation in plants," *Funct Integr Genomics*, vol. 21, no. 2, 2021.

[21] L. Wang, H. J. Park, S. Dasari, S. Wang, J.-P. Kocher, and W. Li, "CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model," *Nucleic Acids Res.*, vol. 41, no. 6, 2013.

[22] U. Singh, N. Khemka, M. S. Rajkumar, R. Garg, and M. Jain, "PLncPRO for prediction of long non-coding RNAs (lncRNAs) in plants and its application for discovery of abiotic stress-responsive lncRNAs in rice and chickpea," *Nucleic Acids Res.*, vol. 45, no. 22, 2017.

[23] A. Li, J. Zhang, and Z. Zhou, "PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme," *BMC Bioinformatics*, vol. 15, no. 1, 2014.

[24] A. Li, H. Zhou, S. Xiong, J. Li, S. Mallik, R. Fei, Y. Liu, H. Zhou, X. Wang, X. Hei, and L. Wang, "PLEKv2: predicting lncRNAs and mRNAs based on intrinsic sequence features and the coding-net model," *BMC Genomics*, vol. 25, no. 1, 2024.

[25] C. Pian, G. Zhang, Z. Chen, Y. Chen, J. Zhang, T. Yang, and L. Zhang, "LncRNApred: Classification of Long Non-Coding RNAs and Protein-Coding Transcripts by the Ensemble Algorithm with a New Hybrid Feature," *PLoS ONE*, vol. 11, no. 5, 2016.

[26] X.-N. Fan and S.-W. Zhang, "lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning," *Mol. BioSyst.*, vol. 11, no. 3, 2015.

[27] G. Eraslan, Z. Avsec, J. Gagneur, and F. J. Theis, "Deep learning: new computational modelling techniques for genomics," *Nat Rev Genet*, vol. 20, no. 7, 2019.

[28] A. Shrestha and A. Mahmood, "Review of Deep Learning Algorithms and Architectures," *IEEE Access*, vol. 7, 2019.

[29] M. W. Szcześniak, O. Bryzghalov, J. Ciomborowska-Basheer, and I. Makałowska, "CANTATAdb 2.0: Expanding the Collection of Plant Long Noncoding RNAs," in *Plant Long Non-Coding RNAs*. New York, NY: Springer New York, 2019, vol. 1933, pp. 415–429.

[30] A. Paytuví Gallart, A. Hermoso Pulido, I. Anzar Martínez de Lagrán, W. Sanseverino, and R. Aiese Cigliano, "GREENC: a Wiki-based database of plant lncRNAs," *Nucleic Acids Res*, vol. 44, no. D1, 2016.

[31] M. Di Marsico, A. Paytuvi Gallart, W. Sanseverino, and R. Aiese Cigliano, "GreeNC 2.0: a comprehensive database of plant long non-coding RNAs," *Nucleic Acids Res.*, vol. 50, no. D1, 2022.

[32] A. Singh, A. T. Vivek, and S. Kumar, "AlnC: An extensive database of long non-coding RNAs in angiosperms," *PLoS ONE*, vol. 16, no. 4, 2021.

[33] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome," *Bioinformatics*, vol. 37, no. 15, 2021.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv*, 2018.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv*, 2017.

[36] L. Romeijn, D. Cats, K. Wolstencroft, and H. Mei, "LncRNA-BERT: An RNA Language Model for Classifying Coding and Long Non-Coding RNA," *bioRxiv*, 2025.

[37] N. Jeon, X. Qian, L. SaidyKhan, P. de Figueiredo, and B.-J. Yoon, "LoRA-BERT: a Natural Language Processing Model for Robust and Accurate Prediction of long non-coding RNAs," *arXiv*, 2024.

[38] J. Li, X. Zhang, and C. Liu, "The computational approaches of lncRNA identification based on coding potential: Status quo and challenges," *Comput. Struct. Biotechnol. J.*, vol. 18, 2020.

[39] I. Messaoudi, A. E. Oueslati, and Z. Lachiri, "Wavelet analysis of frequency chaos game signal: a time-frequency signature of the C. elegans DNA," *J Bioinform Sys Biology*, vol. 2014, no. 1, 2014.

[40] A. Sievers, K. Bosiek, M. Bisch, C. Dreessen, J. Riedel, P. Froß, M. Hausmann, and G. Hildenbrand, "K-mer Content, Correlation, and Position Analysis of Genome DNA Sequences for the Identification of Function and Evolutionary Features," *Genes*, vol. 8, no. 4, 2017.

[41] H. Jeffrey, "Chaos game representation of gene structure," *Nucleic Acids Res.*, vol. 18, no. 8, 1990.

[42] H. F. Löchel and D. Heider, "Chaos game representation and its applications in bioinformatics," *Comput. Struct. Biotechnol. J.*, vol. 19, 2021.

[43] M. S. Hammad, V. F. Ghoneim, M. S. Mabrouk, and W. I. Al-atabany, "A hybrid deep learning approach for COVID-19 detection based on genomic image processing techniques," *Sci Rep*, vol. 13, no. 1, 2023.

[44] J. Avila Cartes, S. Anand, S. Ciccolella, P. Bonizzoni, and G. Della Vedova, "Accurate and fast clade assignment via deep learning and frequency chaos game representation," *GigaScience*, vol. 12, 2022.

[45] F. B. Nasr and A. E. Oueslati, "A new automatic method for human coding and non-coding zones characterization and classification based on FCGR coding and CNN classifier," in *2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. Sfax, Tunisia: IEEE, 2022, pp. 1–6.

[46] Y. Ren, T. Chakraborty, S. Doijad, L. Falgenhauer, J. Falgenhauer, A. Goesmann, A.-C. Hauschild, O. Schwengers, and D. Heider, "Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning," *Bioinformatics*, vol. 38, no. 2, 2022.

[47] 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," in *Proceedings of the 17th International Conference on Computer Systems and Technologies 2016*. Palermo Italy: ACM, 2016, pp. 222–228.

[48] B. De Medeiros, L. Cai, P. Flynn, Y. Yan, X. Duan, L. Marinho, C. Anderson, and C. Davis, "A universal DNA barcode for the Tree of Life," *EcoEvoRxiv*, 2024.

[49] F. Alipour, K. A. Hill, and L. Kari, "CGRclust: Chaos Game Representation for twin contrastive clustering of unlabelled DNA sequences," *BMC Genomics*, vol. 25, no. 1, 2024.

[50] B. P. Welford, "Note on a Method for Calculating Corrected Sums of Squares and Products," *Technometrics*, vol. 4, no. 3, 1962.