Jurnal AL-AZHAR INDONESIA SERI SAINS DAN TEKNOLOGI
Vol. 11, Nomor 01, January 2026, hal. 85-93
DOI http://dx.doi.org/10.36722/sst.v11i1.5258 https://jurnal.uai.ac.id/index.php/SST

Transforming DNA Sequences Into Musical Patterns Via A 3-mer Classification

Abel Prayoga[1], Elis Khatizah[1*]

[1]Mathematics Study Program, School of Data Science, Mathematics, and Informatics, IPB University, Kampus IPB Dramaga, Bogor, 16680.

Corresponding author /E-mail: elis_khatizah@apps.ipb.ac.id

Abstract – **DNA can be viewed as a symbolic sequence with patterns that vary across species. This study explores DNA sequences through two complementary approaches: species classification using simple machine learning methods and transformation of DNA into musical note representations. In the first task, DNA sequences from five organisms with different evolutionary distances are represented using 3-mer and 6-mer features. These k-mers form a vocabulary whose frequency counts are converted into feature vectors. Random Forest (RF) and Support Vector Machine (SVM) models are then applied for five-class classification. Using an 80:20 train-test split and 10-fold cross-validation, the SVM model achieved average accuracies above 0.90 for 3-mer features, with low standard deviation, indicating stable performance. In the second approach, 3-mer motifs are mapped to musical notes to generate species-based musical patterns. The resulting musical representations exhibit distinct structural differences across species, reflecting variations in underlying sequence composition. Overall, the results demonstrate that 3-mer features are effective for species discrimination and that musical transformation provides an alternative and intuitive way to visualize DNA sequence patterns.**

Keywords *– DNA Classification, DNA-to-Music, Random Forest, SVM.*

INTRODUCTION

DNA serves as the fundamental blueprint of all living organisms. Its sequence acts as an instruction set that governs cellular development and biological function. Although composed of only four nucleotides, each DNA sequence is highly specific, and even small variations can lead to distinct structural or functional outcomes in cells and tissues[1]. For example, a gene expressed in eye cells encodes instructions that guide the formation of ocular tissue rather than any unrelated structure. This functional specificity arises from the ordered arrangement of nucleotides, making DNA both structurally constrained and mathematically interpretable.

Because of its inherent uniqueness, DNA enables species identification through computational and statistical analysis. Particularly, machine learning methods have become central tools for modelling the high-dimensional patterns embedded in nucleotide sequences. For instance, a Random Forest (RF) classifier was used to model precursor micro RNA data from 16 species [2]. The results demonstrated that prediction accuracy is influenced by evolutionary distance, with performance ranging from approximately 80% to 93% when distinguishing Homo sapiens from members of the Brassicaceae family. Their study relied on k-mer representations, specifically 1-mers, 2-mers, and 3-mers, as numerical features, noting that larger k-mers rapidly expand the feature space without providing proportional improvements in accuracy. Moreover, k-mers are well suited for analyzing large sequencing datasets because they are computationally efficient, require less memory and still capture important biological information [3].

Another approach can be seen in [4], who used a multiclass Support Vector Machine (SVM) combined with the N-best algorithm to classify microbial marker clades. It evaluated k-mer sizes of 10, 20, 30, 40 and 50 and used the N-best algorithm to address overlap and redundancy between feature sets. Across 17 species, the method achieved over

Received: 17 December 2025, Accepted: 19 January 2026, Published: January 2026

28% accuracy for the top-1 prediction and over 91% accuracy for the top-10 predictions in both training and testing phases. These findings reinforce that k-mers remain a widely used feature representation for machine learning-based genomic classification. In addition, as sequencing technologies continue to advance, research efforts need to improve k-mer counting methods to handle the growing size of sequencing datasets more effectively [5].

Research [4] also noted that many other machine learning and statistical methods are commonly used in genomic analysis, such as BLAST, Hidden Markov Model (HMM) and others. Beyond these traditional approaches, deep learning methods such as CNNs and LSTMs have also been used to classify DNA sequences directly without relying on the k-mer representation, achieving acceptable performance even for the sequences originate from the same species[6].

Interestingly, music also contains structured patterns that can be described through pitch relationships, interval sequences and rhythmic repetition[7]. These elements resemble feature extraction in data analysis, where patterns are represented in numerical or symbolic form and examined systematically. To the best of our knowledge, although informal online sources have experimented with mapping DNA sequences to musical notes, there is limited peer-reviewed work that examines this idea within a quantitative or computational framework. Therefore, this study proposes a dual consecutive approach. First, machine learning techniques are applied to classify DNA sequences from different species using k-mer-based representations. This step serves as a validation of the selected data and feature representations. Second, the DNA sequences are transformed into melodic patterns based on amino acid polarity, with the aim of exploring whether the resulting musical structures reflect the underlying characteristics of the DNA sequences. Through this approach, the study provides an alternative way to visualize DNA using musical representations.

METHODOLOGY

This study uses DNA sequence data obtained from 5 organisms representing different biological domains: human (*Homo sapiens*), a vertebrate animal (*Mus musculus*), a model plant (*Arabidopsis thaliana*), bacteria (*Escherichia coli*) and virus (Human adenovirus). These organisms were intentionally selected to represent clearly distinct biological groups. Two types of exploration were conducted on the data are a classification task and a simple musical pattern analysis as shown in figure 1.

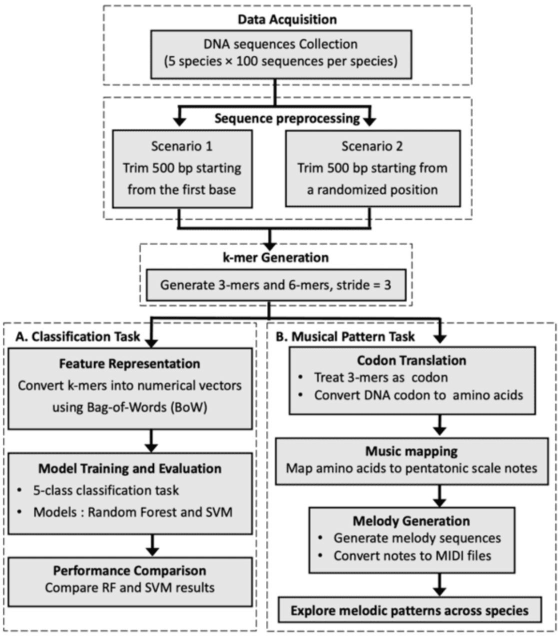

Data Acquisition
DNA sequences Collection
(5 species × 100 sequences per species)
Sequence preprocessing
Scenario 1
Trim 500 bp starting from the first base
Scenario 2
Trim 500 bp starting from a randomized position
k-mer Generation
Generate 3-mers and 6-mers, stride = 3
A. Classification Task
Feature Representation
Convert k-mers into numerical vectors using Bag-of-Words (BoW)
Model Training and Evaluation
• 5-class classification task
• Models : Random Forest and SVM
Performance Comparison
Compare RF and SVM results
B. Musical Pattern Task
Codon Translation
• Treat 3-mers as codon
• Convert DNA codon to amino acids
Music mapping
Map amino acids to pentatonic scale notes
Melody Generation
• Generate melody sequences
• Convert notes to MIDI files
Explore melodic patterns across species

Figure 1. Research Method Flowchart

Data

The DNA sequences were obtained from the National Center for Biotechnology Information (NCBI). We used Coding Sequences (CDS) and the list of accession IDs is provided in Supplementary Material. For each of the five species, 100 sequences were collected in FASTA format. All sequences were trimmed to a uniform length of 500 base pairs to standardize the input data. In addition, one representative 500-base-pair sequence from each species was later translated into its corresponding amino acid chain and subsequently mapped to musical notes using a pentatonic scale to generate melody patterns for comparison.

Method

All DNA sequences were downloaded using Python with the Biopython library to ensure reproducibility. Two trimming strategies were applied to obtain 500-base-pair input sequences. In the first scenario, the sequence was taken starting from the first base pair. In the second scenario, the starting position was randomized to introduce variation. This randomized trimming was intended to introduce positional variation and to evaluate the robustness of the classification approach with respect to sequence starting positions.

To prepare the data for classification task, the trimmed DNA sequences were transformed into numerical features using the k-mer method. A k-mer represents a subsequence of length k extracted from a longer nucleotide sequence [8]. This study used 3-mers and 6-mers with a sliding window of size 3, allowing the capture of local and slightly longer-range sequence patterns. The resulting k-mers were converted into feature vectors using the Bag-of-Words technique. Bag-of-Words is a text vectorization method that ignores order or grammar and counts the frequency of unique "words" (in this case, k-mers) in the sequence [9]. These frequency vectors served as inputs for the classification models in the classification investigation. Since DNA consist of four nucleotides (A, T, G, and C), the feature vector size is $4^3 = 64$ for 3-mer s and 4^6 for 6-mer.

With the feature vectors constructed, two mathematical classification models, Random Forest (RF) and Support Vector Machine (SVM), were implemented. Random Forest can be viewed as an ensemble method that constructs a collection of decision functions, where each function corresponds to a decision tree trained on a randomly sampled subset of the observations and feature space. As described in [10], each tree is generated using an independently sampled random vector with an identical distribution, ensuring statistical diversity across the ensemble. The final classifier is obtained by aggregating the individual tree outputs, typically through majority voting, which approximates an ensemble decision function. This aggregation reduces variance and improves generalization, forming the basis of the RF model shown conceptually in figure 2.

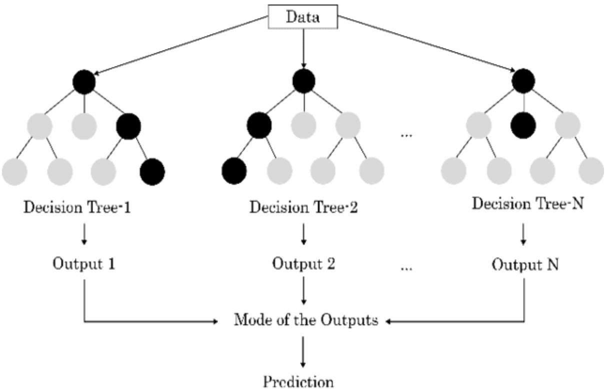


Figure 2. Random Forest Model Constructed from Multiple Decision Trees

On the other hand, SVM model constructs a separating hyperplane that divides the data into two classes[11]. The optimal hyperplane is obtained by solving an optimization problem that determines the parameters w and b. A simple two-dimensional linear example is shown in figure 3.

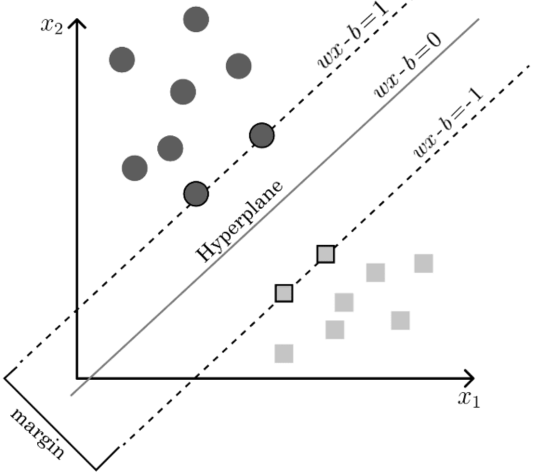


Figure 3. Support Vector Machine Algorithm Illustration

The 3-mer and 6-mer of DNA features were then used in both RF and SVM models and their results were compared. The experiments were designed by splitting the dataset into 80% for training and 20% for testing. Model performance was evaluated using accuracy, precision, recall and F1-score for each species. To ensure that the model is more stable and accurate, it is further evaluated using 10-fold cross-validation. The standard deviation and confidence interval are also calculated to measure how consistent the accuracy is across the different folds.

All experiments were conducted using a fixed random seed to ensure reproducibility. The random state parameter was set to 42 for both Support Vector Machine (SVM) and Random Forest (RF) models. Hyperparameter tuning was performed using GridSearchCV with cross validation on the training data.

For the Random Forest classifier, the evaluated parameters included the number of trees (n_estimators = {100, 200}), maximum tree depth (max_depth = {10, 20, None}), minimum samples required to split an internal node (min samples split = {2, 5, 10}) and class weight ({balanced, None}). For the Support Vector Machine classifier, the regularization parameter C ({0.1, 1, 10, 100}), kernel coefficient gamma ({scale, auto, 0.1, 0.01}) and kernel type ({linear, rbf}) were optimized. The optimal hyperparameters were selected based on classification accuracy obtained during cross validation and were used to train the final models.

As the second type of exploration, a simple musical pattern analysis was conducted to visualize structural pattern of DNA sequences. For this purpose, the extracted 3-mers were treated as codons and converted into amino acids according to the rules shown in figure 4 (with thymine replaced by uracil). The resulting amino acid sequences were then mapped to musical notes in the pentatonic scale, as listed in table 1.

It is important to note that 3-mers were treated as codon-like units and mapped to amino acids for the purpose of symbolic representation rather than biological translation. Although coding sequences (CDS) were used in this study, the translation step does not aim to identify functional proteins or biologically valid Open Reading Frames (ORFs). Instead, a fixed reading frame starting from the first nucleotide of each selected sequence was applied consistently to ensure a uniform and reproducible mapping across all species. This approach allows the 3-mer structure of DNA sequences to be systematically transformed into amino acid symbols, which are subsequently used for musical mapping and pattern visualization.

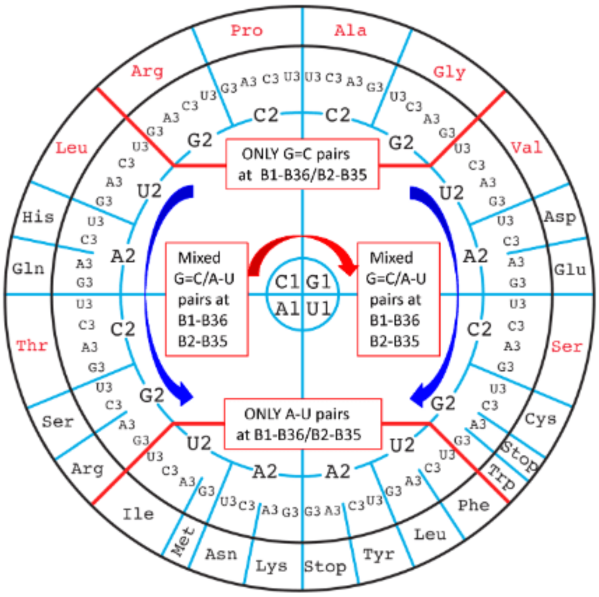

Pro
Ala
Gly
Val
Asp
Glu
Ser
Cys
Stop
Trp
Phe
Leu
Tyr
Stop
Lys
Asn
Met
Ile
Arg
Ser
Thr
Gln
His
Leu
Arg
ONLY G=C pairs at B1-B36/B2-B35
ONLY A-U pairs at B1-B36/B2-B35
Mixed G=C/A-U pairs at B1-B36 B2-B35
Mixed G=C/A-U pairs at B1-B36 B2-B35
C1 G1
A1 U1
C2
G2
U2
A2

Figure 4. Amino Acids Table [12]

Table 1. Pentatonic Scale Mapping of Amino Acids

Property Group	Amino Acids	Musical Note Range	Ratio-nale
Hydro-phobic and nonpolar	Glycine, Alanine, Valine, Leucine, Isoleucine	C3 – A3	Lower pitch
	Methionine, Phenylalanine, Tryptophan, Proline	C4 – G4	
Polar (un-charged)	Serine, Threonine, Cysteine, Tyrosine, Asparagine, Glutamine	A4 – A5	Middle range
Polar (charged)	Aspartic Acid, Glutamic acid, Lysine, Arginine, Histidine	C6 – A6	Higher pitch
Start codon	Methionine	C4	Middle C note
Stop codon	UAG, UAA, UGA	C7	Highest note

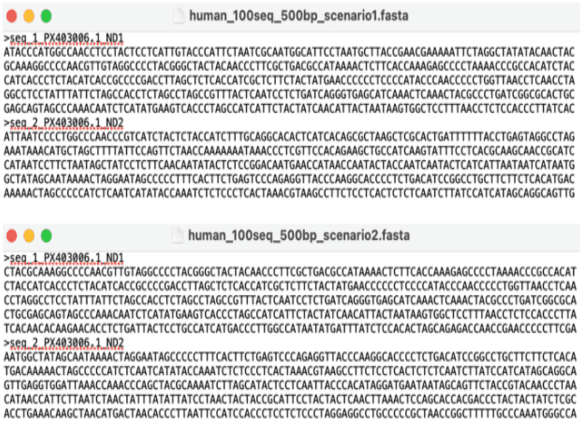
A musical scale is not simple to define. It is essentially a sequence of notes arranged by specific intervals, which represent the frequency differences between consecutive tones[13]. One familiar example is the major scale: do, re, mi, fa, sol, la, ti, do. The pentatonic scale, as the name suggests, consists of five notes: do, re, mi, sol, and la, and is widely used due to its simplicity and ease of improvisation[14]. This five-note structure provides a clearer way to identify recurring patterns in DNA sequences, thereby facilitating qualitative comparison of sequence characteristics across different species.

To show the pattern formed by these musical notes, the notes were then converted into a melody using the MIDIFile library in python by playing these notes next to each other at a certain speed and then turning that into a midi. A melody in essence is a sequence of musical notes that exhibit complex dependencies in different time scales [15]. The generated midi file was then put in FL Studio, a digital audio workstation (DAW), where the sound characteristics were refined and a backing track was added to facilitate clearer auditory comparison of the resulting melodies.

RESULT AND DISCUSSION

DNA Sequence Classification Task

The performance of two classification scenarios is evaluated, Scenario I, in which each 500-base-pair DNA segment is taken directly from the first 500 bp of the corresponding NCBI sequence and Scenario II, in which the starting position for each 500 bp segment is randomized. figure 5 displays the first two human DNA sequences obtained using both scenarios. As shown in the figure, the resulting sequences from the two scenarios begin with different bases or nucleotides. For example, in the first sequence, Scenario I starts with ATACCC…, whereas Scenario II starts with CTACGC…. These conditions therefore reflect that the extracted sequences accommodate the introduced positional variation arising from different sequence starting positions.


human_100seq_500bp_scenario1.fasta

```
>seq_1_PX403006.1_ND1
ATACCCATGGCCAACCTCCTACTCCTCATTGTACCCATTCTAATCGCAATGGCATTCCTAATGCTTACCGAACGAAAAATTCTAGGCTATATACAACTAC
GCAAAGGCCCCAACGTTGTAGGCCCCTACGGGCTACTACAACCCTTCGCTGACGCCATAAAACTCTTCACCAAAGAGCCCCTAAAACCCGCCACATCTAC
CATCACCCTCTACATCACCGCCCCGACCTTAGCTCTCACCATCGCTCTTCTACTATGAACCCCCCTCCCCATACCCAACCCCCTGGTTAACCTCAACCTA
GGCCTCCTATTTATTCTAGCCACCTCTAGCCTAGCCGTTTACTCAATCCTCTGATCAGGGTGAGCATCAAACTCAAACTACGCCCTGATCGGCGCACTGC
GAGCAGTAGCCCAAACAATCTCATATGAAGTCACCCTAGCCATCATTCTACTATCAACATTACTAATAAGTGGCTCCTTTAACCTCTCCACCCTTATCAC
>seq_2_PX403006.1_ND2
ATTAATCCCCTGGCCCAACCCGTCATCTACTCTACCATCTTTGCAGGCACACTCATCACAGCGCTAAGCTCGCACTGATTTTTTACCTGAGTAGGCCTAG
AAATAAACATGCTAGCTTTTATTCCAGTTCTAACCAAAAAAATAAACCCTCGTTCCACAGAAGCTGCCATCAAGTATTTCCTCACGCAAGCAACCGCATC
CATAATCCTTCTAATAGCTATCCTCTTCAACAATATACTCTCCGGACAATGAACCATAACCAATACTACCAATCAATACTCATCATTAATAATCATAATG
GCTATAGCAATAAAACTAGGAATAGCCCCCTTTCACTTCTGAGTCCCAGAGGTTACCCAAGGCACCCCTCTGACATCCGGCCTGCTTCTTCTCACATGAC
AAAAACTAGCCCCCATCTCAATCATATACCAAATCTCTCCCTCACTAAACGTAAGCCTTCTCCTCACTCTCTCAATCTTATCCATCATAGCAGGCAGTTG
```

human_100seq_500bp_scenario2.fasta

```
>seq_1_PX403006.1_ND1
CTACGCAAAGGCCCCAACGTTGTAGGCCCCTACGGGCTACTACAACCCTTCGCTGACGCCATAAAACTCTTCACCAAAGAGCCCCTAAAACCCGCCACAT
CTACCATCACCCTCTACATCACCGCCCCGACCTTAGCTCTCACCATCGCTCTTCTACTATGAACCCCCCTCCCCATACCCAACCCCCTGGTTAACCTCAA
CCTAGGCCTCCTATTTATTCTAGCCACCTCTAGCCTAGCCGTTTACTCAATCCTCTGATCAGGGTGAGCATCAAACTCAAACTACGCCCTGATCGGCGCA
CTGCGAGCAGTAGCCCAAACAATCTCATATGAAGTCACCCTAGCCATCATTCTACTATCAACATTACTAATAAGTGGCTCCTTTAACCTCTCCACCCTTA
TCACAACACAAGAACACCTCTGATTACTCCTGCCATCATGACCCTTGGCCATAATATGATTTATCTCCACACTAGCAGAGACCAACCGAACCCCCTTCGA
>seq_2_PX403006.1_ND2
AATGGCTATAGCAATAAAACTAGGAATAGCCCCCTTTCACTTCTGAGTCCCAGAGGTTACCCAAGGCACCCCTCTGACATCCGGCCTGCTTCTTCTCACA
TGACAAAAACTAGCCCCCATCTCAATCATATACCAAATCTCTCCCTCACTAAACGTAAGCCTTCTCCTCACTCTCTCAATCTTATCCATCATAGCAGGCA
GTTGAGGTGGATTAAACCAAACCCAGCTACGCAAAATCTTAGCATACTCCTCAATTACCCACATAGGATGAATAATAGCAGTTCTACCGTACAACCCTAA
CATAACCATTCTTAATCTAACTATTTATATTATCCTAACTACTACCGCATTCCTACTACTCAACTTAAACTCCAGCACCACGACCCTACTACTATCTCGC
ACCTGAAACAAGCTAACATGACTAACACCCTTAATTCCATCCACCCTCCTCTCCCTAGGAGGCCTGCCCCCGCTAACCGGCTTTTTGCCCAAATGGGCCA
```

Figure 5. First Two Human DNA Sequences Extracted Using Fixed and Randomized Trimming

Using k-mer sizes of $k = 3$ and $k = 6$ with a stride of 3 resulted in 166 3-mers and 83 6-mers, respectively, for each 500 bp sequence. For example, for Sequence 1 under Scenario I in Figure 5, the 3-mer representation yields the list {ATA, CCC, ATG, …, ATC}, whereas the 6-mer representation yields the list {ATACCC, CCCATG, ATGGCC, …, CTTATC}. The remaining two nucleotides at the end of the sequence were discarded because they do not form a complete k-mer under the selected stride. After converting the k-mers from all five species into feature vectors using the Bag-of-Words technique, Random Forest (RF) and Support Vector Machine (SVM) model were trained for a five-class classification task using an 80:20 train-test data split.

Performing hyperparameter tuning using GridSearchCV on the cleaned data from Scenario I, the optimal parameters for each model were obtained as follows. For k = 3, the Random Forest model achieved the best performance with 100 trees, a maximum depth of 10 and minimum samples split of 10, without class weighting. Meanwhile, the SVM model performed best using an RBF kernel with C = 10 and gamma set to scale. For k = 6, the optimal Random Forest configuration consisted of 200 trees, no depth limitation, minimum samples split of 2 and balanced class weights. In contrast, the SVM model achieved its best performance with a linear kernel, C = 0.1, and gamma set to scale. These tuned parameters were subsequently used in all classification experiments to ensure fair and consistent model evaluation.

Table 2 and 3 present the classification performance metrics for RF and SVM models obtained using Scenario I.

Table 2. Performance of RF Model Using Scenario I

***k*-mer Feature**	**Class**	**Metrics performance**			
		Acc.	**Prec.**	**Rec.**	**F1-Score**
$k = 3$	Animal	0.96	1.00	1.00	1.00
	Bacteria	0.96	0.95	1.00	0.98
	Human	0.96	1.00	1.00	1.00
	Plant	0.96	0.90	1.00	0.95
	Virus	0.96	1.00	0.80	0.89
$k = 6$	Animal	0.94	0.95	0.95	0.95
	Bacteria	0.94	0.95	0.95	0.95
	Human	0.94	0.95	0.95	0.95
	Plant	0.94	0.91	1.00	0.95
	Virus	0.94	0.94	0.85	0.89

Table 3. Performance of SVM Model Using Scenario I

k-mer Feature	**Class**	**Metrics performance**			
		Acc.	**Prec.**	**Rec.**	**F1-Score**
$k = 3$	Animal	0.98	1.00	1.00	1.00
	Bacteria	0.98	1.00	1.00	1.00
	Human	0.98	1.00	1.00	1.00
	Plant	0.98	1.00	0.95	0.97
	Virus	0.98	0.94	1.00	0.97
$k = 6$	Animal	0.95	0.94	0.85	0.89
	Bacteria	0.95	0.95	1.00	0.98
	Human	0.95	0.90	0.95	0.93
	Plant	0.95	0.95	1.00	0.98
	Virus	0.95	1.00	0.95	0.97

Based on tables 2 and 3, SVM model achieves overall slightly better performance for both 3-mers and 6-mers classification, outperforming RF model. Furthermore, the 3-mer features yield a higher accuracy than the 6-mer features.

For Scenario II, table 4 and 5 show that SVM model consistently achieves higher accuracy than RF model. Similar to the previous scenario, the 3-mer feature again provide better metric performance. Additionally, the overall classification performance in Scenario II slightly decreases, which may be due to the broader positional variation introduced by random sequence starting positions. Nevertheless, the performance metrics across all species remain acceptable (all above 0.70), suggesting that the model maintains reliable classification performance.

Table 4. Performance of RF Model Using Scenario II

k-mer feature	Class	Metrics performance			
		Acc.	Prec.	Rec.	F1-Score
$k = 3$	Animal	0.95	0.95	0.95	0.95
	Bacteria	0.95	0.91	1.00	0.95
	Human	0.95	0.95	0.95	0.95
	Plant	0.95	1.00	0.95	0.97
	Virus	0.95	0.95	0.90	0.92
$k = 6$	Animal	0.79	0.73	0.80	0.76
	Bacteria	0.79	0.74	0.85	0.79
	Human	0.79	0.85	0.85	0.85
	Plant	0.79	0.83	0.75	0.79
	Virus	0.79	0.82	0.70	0.76

Table 5. Performance of SVM Model Using Scenario II

k-mer feature	Class	Metrics performance			
		Acc.	Prec.	Rec.	F1-Score
$k = 3$	Animal	0.97	1.00	0.95	0.97
	Bacteria	0.97	0.95	0.95	0.95
	Human	0.97	0.95	1.00	0.98
	Plant	0.97	1.00	1.00	1.00
	Virus	0.97	0.95	0.95	0.95
$k = 6$	Animal	0.93	0.94	0.75	0.83
	Bacteria	0.93	0.95	1,00	0.98
	Human	0.93	0.95	0.95	0.95
	Plant	0.93	0.83	1.00	0.91
	Virus	0.93	1.00	0.95	0.97

Although an initial 80:20 train-test split was used, this approach may be sensitive to data partitioning. Therefore, 10-fold cross-validation was employed to obtain a more reliable assessment of model performance, as presented in table 6 and 7.

Table 6. 10-Fold Cross-Validation Results for Scenario I

No	Fold	Accuracy			
		k = 3		k = 6	
		RF	SVM	RF	SVM
1	Fold 1	0.92	0.97	0.94	0.97
2	Fold 2	0.97	1.00	0.92	1.00
3	Fold 3	0.97	1.00	0.94	1.00
4	Fold 4	0.94	0.97	0.94	1.00
5	Fold 5	0.94	1.00	0.97	1.00
6	Fold 6	0.92	1.00	0.89	0.89
7	Fold 7	0.92	0.97	0.92	0.97
8	Fold 8	0.94	0.97	0.92	0.94
9	Fold 9	0.97	1.00	0.97	1.00
10	Fold 10	1.00	1.00	0.97	0.97
Mean		0.95	0.99	0.94	0.98
Standard deviation		0.03	0.01	0.03	0.03
Confidence Interval		(0.93, 0.97)	(0.98, 1.00)	(0.92, 0.96)	(0.95, 1.00)

Table 7. 10-Fold Cross-Validation Results for Scenario II

No	Fold	Accuracy			
		k = 3		k = 6	
		RF	SVM	RF	SVM
1	Fold 1	0.94	0.98	0.92	0.96
2	Fold 2	0.92	1.00	0.82	0.94
3	Fold 3	0.98	1.00	0.92	0.98
4	Fold 4	0.94	0.98	0.88	0.94
5	Fold 5	0.96	0.98	0.98	0.98
6	Fold 6	0.92	0.98	0.98	1.00
7	Fold 7	0.90	0.92	0.84	0.88
8	Fold 8	0.86	0.92	0.90	0.96
9	Fold 9	0.94	0.96	0.90	0.96
10	Fold 10	0.86	0.98	0.94	0.98
Mean		0.92	0.97	0.91	0.96
Standard deviation		0.04	0.03	0.05	0.03
Confidence Interval		(0.90, 0.95)	(0.95, 0.99)	(0.87, 0.95)	(0.94, 0.98)

Based on table 6, in scenario I, the classification models achieved an average accuracy above 0.90 with relatively small standard deviations for the 3-mer features, 0.03 for RF model and 0.01 for SVM model. The confidence interval for RF model, (0.93, 0.97), indicates that the model accuracy consistently falls between 93% and 97%, suggesting stable performance. From the mean accuracy, standard deviation values and confidence intervals, it can be observed that SVM model performs better than RF model and that the 3-mer features provide better classification performance than the 6-mer features.

Furthermore, as shown in table 7, the performance of both models slightly decreases in Scenario II. Again, this decrease may be due to the broader positional variation of nucleotides within sequences of the same species. These results are consistent with the evaluation obtained from the single train-test split, indicating that the models remain stable and maintain consistent performance across different data subsets.

DNA to Music and a Simple Musical Pattern Analysis

As an alternative perspective for observing differences in DNA sequence patterns across species, the sequences are transformed into musical notes. Based on the classification results, where the 3-mer features provide better classification performance, the DNA sequences are transformed into musical notes using the 3-mer strategy. Following the mapping in table 1, each 3-mer (with stride 3) is treated as a codon that determines a corresponding amino acid, which is then converted into a musical note.

Since each sequence contains 500 base pairs, a single sequence generates 166 notes. To maintain clarity and avoid excessive length, the melodies are limited to the first 500 bp of one representative sequence from each species. Specifically, we selected one sequence per species using the first accession ID available (PV177089.1_COX1 for animal, CP195866.1_dnaN for bacteria, PX403006.1_ND2 for human, BK010421.1_cox2 for plant, and MW306919.1_E1A for virus).

For example, Sequence 2 under Scenario I in figure 5 produces the 3-mer list {ATT, AAT, CCC, …, AGT}, which is converted into amino acids as {Isoleucine, Asparagine, Proline, …, Serine}. These amino acids are then mapped to musical notes {‘A3’,’G5’,’G4’, …, ‘A4’}. Complete examples of the resulting musical representations for human and animal sequences are provided in table 8.

Table 8. Notes generated from 500 bp of human and animal DNA sequences

Species	Notes
Human	['A3', 'G5', 'G4', 'G3', 'D3', 'A5', 'G4', 'E3', 'A3', 'E5', 'A4', 'C5', 'A3', 'D4', 'D3', 'C3', 'C5', 'G3', 'A3', 'C5', 'D3', 'G3', 'A4', 'A4', 'A6', 'C7', 'D4', 'D4', 'C5', 'C7', 'E3', 'C3', 'G3', 'D6', 'A3', 'G5', 'C4', 'G3', 'D3', 'D4', 'A3', 'G4', 'E3', 'G3', 'C5', 'E6', 'E6', 'A3', 'G5', 'G4', 'G6', 'A4', 'C5', 'D6', 'D3', 'D3', 'A3', 'E6', 'E5', 'D4', 'G3', 'C5', 'A5', 'D3', 'C5', 'D3', 'A4', 'A3', 'A3', 'G3', 'G3', 'A3', 'D3', 'A3', 'G3', 'D4', 'G5', 'G5', 'A3', 'G3', 'A4', 'C3', 'A5', 'C7', 'C5', 'A3', 'C5', 'G5', 'C5', 'C5', 'G5', 'A5', 'E5', 'A4', 'A4', 'G3', 'A3', 'A3', 'A3', 'C4', 'D3', 'A3', 'D3', 'A3', 'E6', 'G3', 'C3', 'A3', 'D3', 'G4', 'D4', 'A6', 'D4', 'C7', 'E3', 'G4', 'D6', 'E3', 'C5', 'A5', 'C3', 'C5', 'G4', 'G3', 'C5', 'A4', 'C3', 'G3', 'G3', 'G3', 'G3', 'C5', 'C7', 'A5', 'E6', 'G3', 'D3', 'G4', 'A3', 'A4', 'A3', 'A3', 'E5', 'A5', 'A3', 'A4', 'G4', 'A4', 'G3', 'G5', 'E3', 'A4', 'G3', 'G3', 'G3', 'C5', 'G3', 'A4', 'A3', 'G3', 'A4', 'A3', 'A3', 'D3', 'C3', 'A4']
Animal	['C4', 'D4', 'A3', 'G5', 'G6', 'C7', 'G3', 'D4', 'A4', 'C5', 'G5', 'A6', 'E6', 'C6', 'A3', 'C3', 'C5', 'G3', 'E5', 'G3', 'G3', 'D4', 'C3', 'D3', 'C7', 'D3', 'C3', 'A3', 'E3', 'C3', 'C5', 'D3', 'G3', 'A4', 'A3', 'G3', 'A3', 'G6', 'D3', 'D6', 'G3', 'C3', 'A5', 'G4', 'C3', 'D3', 'G3', 'G3', 'C3', 'C6', 'C6', 'A5', 'A3', 'E5', 'G5', 'E3', 'A3', 'E3', 'C5', 'D3', 'A6', 'D3', 'D4', 'E3', 'A3', 'A3', 'D4', 'D4', 'A3', 'E3', 'A3', 'G4', 'A3', 'A3', 'A3', 'C3', 'C3', 'D4', 'C3', 'G5', 'C7', 'G3', 'E3', 'G4', 'G3', 'A3', 'A3', 'C3', 'D3', 'G4', 'C6', 'A3', 'D3', 'D4', 'G4', 'G6', 'A3', 'G5', 'G5', 'A3', 'A4', 'D4', 'C7', 'G3', 'G3', 'G4', 'G4', 'A4', 'D4', 'G3', 'G3', 'G3', 'G3', 'D3', 'A4', 'A4', 'A3', 'E3', 'D6', 'D3', 'C3', 'D3', 'C3', 'C5', 'C3', 'C7', 'C5', 'E3', 'E5', 'G4', 'G4', 'G3', 'D3', 'C3', 'G5', 'G3', 'D3', 'A6', 'D3', 'C3', 'D3', 'A4', 'E3', 'C6', 'G3', 'C5', 'A3', 'D4', 'A4', 'G3', 'A6', 'G3', 'D3', 'C3', 'E3', 'A4', 'A4', 'A3', 'G3', 'C3', 'D3', 'A3', 'G5', 'D4', 'A3', 'C5']

The MIDI files for all species, as well as the musical arrangement, can be accessed at the following link: https://drive.google.com/drive/folders/1XMJ82GCnT-Y52-0Zocqaq0bohXKka2kG?usp=sharing.

The resulting melodies are also visualized using line plots to provide a clearer depiction of their structural patterns. The patterns generated from the human and animal sequences are shown in figure 6 and 7. Both species exhibit similar overall trends, with hydrophobic amino acids (9 of the 20 amino acids) appearing most frequently. In the human sequence, however the frequency of polar uncharged amino acids (6 of the 20 amino acids) is more balanced relative to the hydrophobic group, while the remaining categories appear only rarely. In contrast, the animal sequence shows more frequent fluctuations interval change in notes associated with non-hydrophobic amino acids, indicating greater variation in its amino acid composition. A direct comparison of the note line plots for human and animal sequences is provided in figure 8, which highlights these differences.

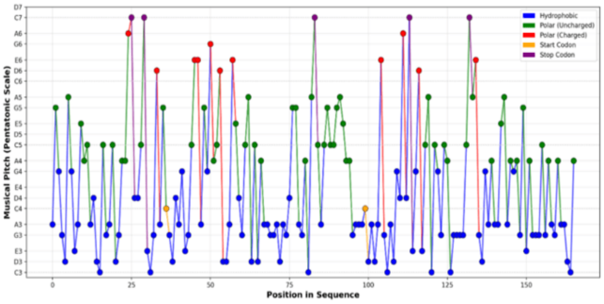


Figure 6. Human gene musical pattern

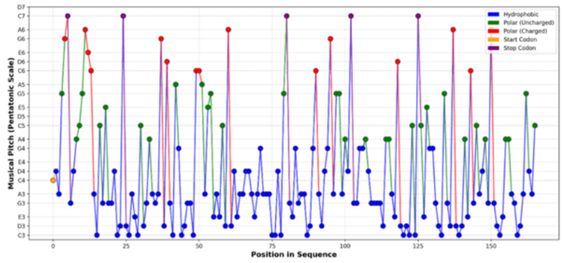


Figure 7. Animal Gene Musical Pattern

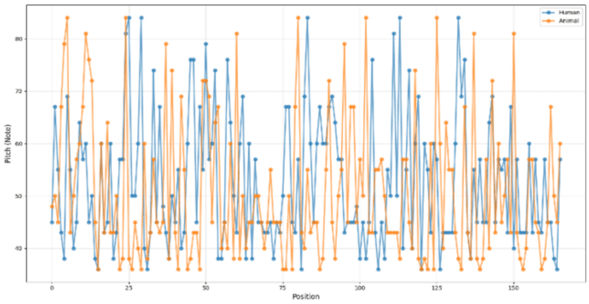


Figure 8. Human and Animal Gene Musical Pattern

As for the patterns generated from the plant, bacteria and virus sequences, the visualizations are shown in Figure 9, 10 and 11. The frequency of charged amino acids, which is relatively rare in human and animal sequences, appears more balanced with hydrophobic amino acids in these species. Although charged amino acids make up only about 25% of all amino acids (5 out of 20), their presence is notably more frequent in bacteria and virus. This can be clearly observed in Figure 10 and 11, where charged amino acids emerge much more often compared to the patterns seen in human and animal sequences. In addition, extreme interval changes, reaching up to 3 octaves are more commonly found in these sequences, reflected greater variability.

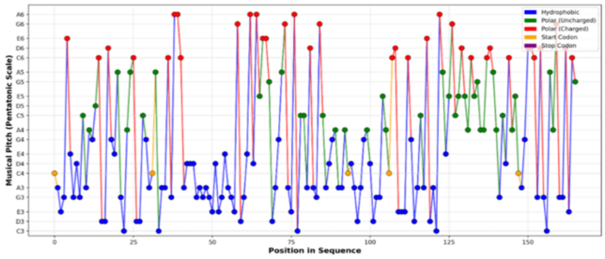


Figure 9. Plant Gene Musical Pattern

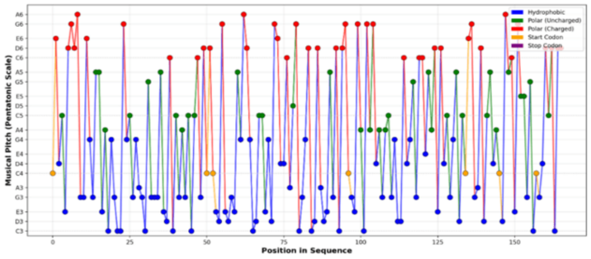


Figure 10. Bacteria Gene Musical Pattern

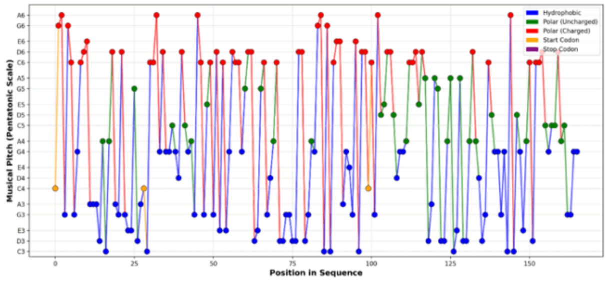


Figure 11. Virus Gene Musical Pattern

Overall, the musical transformation provides an alternative and intuitive way to observe differences in DNA sequence patterns across species. By converting nucleotide sequences into melodic structures, variations between species can be perceived more easily through differences in note distribution and interval patterns. However, it is important to emphasize that this transformation is not intended to represent biological protein synthesis. The 3-mer-to-amino-acid mapping was applied using a fixed reading frame solely for symbolic pattern transformation, not for biological interpretation of coding regions. In addition, this approach ignores codon degeneracy, as multiple codons that encode the same amino acid are treated in the same way. From the 64 possible codons, only 20 amino acids and one stop symbol are mapped to musical notes. This simplification makes the representation easier to apply, but it also reduces the biological detail of the method. Therefore, this approach should be considered a pattern visualization technique rather than a biologically precise model.

CONCLUSION

This study presents two related explorations of DNA sequences. First, species classification using 3-mer features with Random Forest (RF) and Support Vector Machine (SVM) models shows satisfactory performance, indicating that short k-mer representations are effective for distinguishing species with large evolutionary differences. Second, transforming 3-mer DNA sequences into musical notes provides an alternative and intuitive way to observe differences in DNA sequence patterns across species, as reflected by note distribution and interval patterns.

However, this study has some limitations. Only short DNA segments of 500 base pairs were analyzed, whereas real genomic sequences are much longer. In addition, the DNA-to-music transformation simplifies biological information by ignoring codon degeneracy, where multiple codons that encode the same amino acid are treated in the same way. This simplification makes the method easier to apply and suitable for exploratory analysis, but it reduces biological detail.

Despite these limitations, the simplicity of this approach supports exploratory data analysis and provides an intuitive symbolic view of DNA sequences. Future work may address these limitations to improve biological relevance while preserving the simplicity and interpretability of the musical visualization.

REFERENCES

[1] C. R. . Calladine, *Understanding DNA : the molecule & how it works*. Elsevier Academic Press, 2004.

[2] M. Yousef and J. Allmer, "Classification of Precursor MicroRNAs from Different Species Based on K-mer Distance Features," *Algorithms*, vol. 14, no. 5, p. 132, Apr. 2021, doi: 10.3390/a14050132.

[3] J. Liu, "Random Fragments Classification of Microbial Marker Clades with Multi-class SVM and N-Best Algorithm," Apr. 2019. DOI: 10.48550/arXiv.1904.09061.

[4] E. Khatizah and H. S. Park, "Country-Based COVID-19 DNA Sequence Classification in Relation with International Travel Policy," *Applied Sciences (Switzerland)*, vol. 14, no. 5, 2024, doi: 10.3390/app14051916.

[5] D. Temperley, "Melodic Pattern Repetition and Efficient Encoding: A Corpus Study," *Empirical Musicology Review*, vol. 18, no. 2, pp. 97–116, Jun. 2024, doi: 10.18061/emr.v18i2.9289.

[6] K. M. Jenike *et al.*, "*k* -mer approaches for biodiversity genomics," *Genome Res*, Jan. 2025, doi: 10.1101/gr.279452.124.

[7] S. H. Huspi, H. D. Abubakar, and M. Umar, "A Scheme of Pairwise Feature Combinations to Improve Sentiment Classification Using Book Review Dataset," *International Journal of Innovative Computing*, vol. 12, no. 1, pp. 25–33, Nov. 2021, doi: 10.11113/ijic.v12n1.344.

[8] F. Maleki, K. Ovens, K. Najafian, B. Forghani, C. Reinhold, and R. Forghani, "Overview of Machine Learning Part 1," *Neuroimaging Clin N Am*, vol. 30, no. 4, pp. e17–e32, Nov. 2020, doi: 10.1016/j.nic.2020.08.007.

[9] C. Avci, M. Budak, N. Yağmur, and F. Balçik, "Comparison between random forest and support vector machine algorithms for LULC classification," *International Journal of Engineering and Geosciences*, vol. 8, no. 1, pp. 1–10, Feb. 2023, doi: 10.26833/ijeg.987605.

[10] C. Moeckel et al., "A Survey of K-mer Methods and Applications in Bioinformatics," *Computational and Structural Biotechnology Journal*, vol. 23, pp. 2289–2303, Dec. 2024, doi: https://doi.org/10.1016/j.csbj.2024.05.025.

[11] S. C. Manekar and S. R. Sathe, "A benchmark study of k-mer counting methods for high-throughput sequencing," *GigaScience*, Oct. 2018, doi: https://doi.org/10.1093/gigascience/giy125.

[12] H. Grosjean and E. Westhof, "An integrated, structure- and energy-based view of the genetic code," *Nucleic Acids Res*, vol. 44, no. 17, pp. 8020–8040, Sep. 2016, doi: 10.1093/nar/gkw608.

[13] G. Hardegree, "Scales in Music," *Academia.edu*, 2001. https://www.academia.edu/65537416/Scales_in_Music (accessed Nov. 2, 2025).

[14] A. Nakamura, O. J. College, K. Kinoshita, and Y. Nanjo, "The Pentatonic Scale Gives Everyone a Chance to Create Music: Creating, Sharing, and Developing Music with Participants," *International Journal of Creativity in Music Education*, vol.09, pp.42-55, 2022. DOI https://doi.org/10.50825/icme.09.0_42.

[15] J. Wu, X. Liu, X. Hu, and J. Zhu, "PopMNet: Generating structured pop music melodies using neural networks," *Artificial Intelligence*, vol. 286, p. 103303, Sep. 2020, doi: 10.1016/j.artint.2020.103303.