Biochimie 199 (2022) 112–122

Contents lists available at ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

A novel binary *k*-mer approach for classification of coding and non-coding RNAs across diverse species

Neha Periwal [a, 1], Priya Sharma [a, 1], Pooja Arora [b], Saurabh Pandey [a], Baljeet Kaur [c, **], Vikas Sood [a, *]

[a] Department of Biochemistry, School of Chemical and Life Sciences, Jamia Hamdard, Delhi, 110062, India
[b] Department of Zoology, Hansraj College, University of Delhi, Delhi, 110007, India
[c] Department of Computer Science, Hansraj College, University of Delhi, Delhi, 110007, India

ARTICLE INFO

Article history:
Received 20 October 2021
Received in revised form
12 March 2022
Accepted 21 April 2022
Available online 25 April 2022

Keywords:
Binary k-mer
ncRNA
CDS
Machine learning
k-nearest neighbour

ABSTRACT

Classification among coding sequences (CDS) and non-coding RNA (ncRNA) sequences is a challenge and several machine learning models have been developed for the same. Since the frequency of curated CDS is many-folds as compared to that of the ncRNAs, we devised a novel approach to work with the complete datasets from fifteen diverse species. In our proposed binary approach, we replaced all the 'A's and 'T's with '0's and 'G's and 'C's with '1's to obtain a binary form of CDS and ncRNAs. The *k*-mer analysis of these binary sequences revealed that the frequency of binary patterns among the CDS and ncRNAs can be used as features to distinguish among them. Using insights from these distinguishing frequencies, we used *k*-nearest neighbor classifier to classify among them. Our strategy is not only time-efficient but leads to significantly increased performance metrics in terms of Matthews Correlation Coefficient (MCC), Accuracy, F1 score, Precision, Recall and AUC-ROC, for species like *P. paniscus, M. mulatta, M. lucifugus, G. gallus, C. japonica, C. abingdonii, A. carolinensis, D. melanogaster* and *C. elegans* when compared with the conventional ATGC approach. Additionally, we also show that the performance obtained for diverse species tested on the model based on *H. sapiens,* correlated with the geological evolutionary timeline, thereby further strengthening our approach. Therefore, we propose that CDS and ncRNAs can be efficiently classified using "2-character" binary frequency as compared to "4-character" frequency of ATGC approach. Thus, our highly efficient binary approach can replace the more complex ATGC approach successfully.

1. Introduction

The central dogma of molecular biology suggests that the genetic information is typically processed from DNA to RNA and from RNA to proteins. Less than 5% of genomic DNA is capable to encode proteins and can be deciphered in the form of coding sequences (CDS) [1]. These protein coding sequences have been studied extensively. The remaining 95% of genomic DNA sequences that do not encode for proteins, were assumed to be junk in nature with no function in biological processes. The portions from this non-coding genomic DNA which gets transcribed are termed as non-coding RNAs (ncRNAs). However, with advances in the biological sciences, these ncRNAs have been shown to regulate gene expression and play a vital role in various biological processes [2]. The importance of ncRNAs can be conjectured by the fact that the repertoire of ncRNA has been directly related to the biological complexity [3]. Non-coding RNAs play a critical role in the biological processes including reproduction [4], embryonic development [5], immune responses [6], cancer [7] and viral infections [8–10]. Several ncRNAs have also been shown to act as biomarkers in various diseases including cancer, tuberculosis, dengue and autoimmune diseases [11–17].

Attesting to this complexity of ncRNAs, recently, another class of ncRNA known as circular RNA were identified in several eukaryotes [18–20] and viruses [21–23]. Circular RNAs have been shown to

* Corresponding author.
** Corresponding author.
E-mail addresses: baljeetkaur26@hotmail.com (B. Kaur), v.sood@jamiahamdard.ac.in, vikas1101@gmail.com (V. Sood).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.biochi.2022.04.012
0300-9084/

regulate gene expression via sequestering the micro-RNAs inside the cells [24] as well as play a critical role in regulating host innate immune responses [25]. Apart from their role in gene regulation, circular RNAs can act as biomarkers for several diseases [26] further adding towards the importance of ncRNAs in biological functions.

Owing to their role and propensity towards disease diagnosis and pathogenesis, categorization and characterization of ncRNAs is essential to understand the fundamentals of biological processes and disease progression. Additionally, next generation sequencing is generating huge amounts of data consisting of thousands of novel transcripts thereby making ncRNA identification and classification a mammoth task. Therefore, there is a need to devise approaches based on novel strategies and methods that can classify CDS and ncRNAs with high accuracy and reliability in resource and time efficient manner.

Past research to distinguish CDS and ncRNAs has been based on features which are derived from transcript nucleotide sequences, ORF length, ORF integrity, isoelectric point (pI), gene structure, potential codon sequence, conservation and frequencies of nucleotide pattern [27–35]. However, these techniques have not yielded high performance and are complex in terms of computation and time. Moreover, most of these studies have used balanced positive and negative datasets which tend to lose the critical features that may help to differentiate the CDS and ncRNAs. To overcome these limitations, this article proposes a machine learning based binary *k*-mer approach to classify all curated CDS from ncRNAs without the loss of crucial features.

Machine learning (ML) based approaches have been successfully used in the classification problems. Among several ML models, *k*-Nearest Neighbor (*k*NN) is a simple yet powerful method that is non parametric in nature [36]. It can resolve complex boundaries between classes and discriminate data effectively. In ML based decision systems, appropriate features are a prerequisite to build a robust model. This paper proposes a binary *k*-mer approach to build a robust feature set and to overcome the pertinent challenge of complexity and time. In this approach the 'A's and 'T's nucleotides are replaced with a zero ('0') whereas 'G's and 'C's with a one ('1'). This approach was formulated keeping in mind that 'A' and 'T' both pair with each other with double bonds (low energy) whereas 'G' and 'C' pair with each other using triple bonds (high energy).

The ML model using *k*NN was then built to classify CDS and ncRNAs using the features extracted from the binary *k*-mer approach. To validate our approach, exhaustive experiments were carried out on fifteen phylogenetically diverse species including *Homo sapiens, Pan paniscus, Macaca mulatta, Mus musculus, Myotis lucifugus, Gallus gallus, Coturnix japonica, Chelonoidis abingdonii, Anolis carolinensis, Xenopus tropicalis, Leptobrachium leishanense, Salmo trutta, Danio rerio, Drosophila melanogaster* and *Caenorhabditis elegans.* The *k*-mer analysis of binary sequences revealed the enrichment of distinguishing patterns among CDS and ncRNAs from all the species. Using the frequency of various patterns of *k*-mer in each sequence in binary files as the features, we show that our ML model based on *k*NN is highly efficient in classifying CDS and ncRNAs as compared to the conventional ATGC *k*-mer approach both in terms of computational time as well as Matthews Correlation Coefficient (MCC) which is the preferred metric to evaluate the performance of the imbalanced data. Thus, we have identified a unique binary strategy which can be explored to differentiate among CDS and ncRNAs in a highly efficient manner.

2. Related work

With an ever increasing role of ncRNAs in biological functions, their discovery and characterization has become the need of the hour. Experimental validation for ncRNA characterization is time consuming and resource intensive step. Therefore, several computational approaches have been used towards the classification of CDS and ncRNAs. In this direction, several groups have successfully used ML and deep learning based approaches to classify CDS from ncRNAs. For example, the support vector machine (SVM) classifier based Coding Potential Calculator (CPC) utilized six features of transcripts to classify 5610 protein coding cDNA and 2670 ncRNAs with good accuracy [27]. Another SVM model by Sun et al. utilized ten features from the broad categories including sequence conservation, open reading frame (ORF) features and frequencies of seven di-nucleotide or tri-nucleotide sequences to identify long intergenic non-coding RNA (lincRNA) sequences. The study included 5079 and 889 lincRNA sequences from human and mouse respectively and demonstrated higher prediction accuracy [28]. However, one of the drawbacks of the model was that it was restricted to sequences obtained from human and mouse only. Another tool named Coding Non-Coding Index (CNCI) was created that was based on intrinsic sequence information i.e., frequency of adjoining nucleotide triplets (ANT) to predict Most-Like CDS regions (MLCDS). Although the model was able to predict ncRNAs in a variety of species especially with poor genomic annotation, it had the tendency to misclassify transcripts with sequencing errors such as insertions and deletions [29]. The shortcomings of the CNCI tool were taken care of by another binary classifier named PLEK (predictors of long-non coding RNAs and messenger RNA) which was based on an improved *k*-mer scheme and SVM algorithm [30]. The classifier was trained on 34,691 human CDS and 22,389 ncRNA sequences and provided better specificity and sensitivity as compared to CNCI on real sequencing data. Another tool named long non-coding RNA identification tool (LncRNA-ID) was built on random forest (RF) classification method and explored the coding potential of a transcript based on multiple features broadly covering three categories i.e., ORF, ribosome interaction and protein conservation [31]. Other SVM based method utilized various features derived from gene structure, transcript sequence, potential codon sequence, conservation, ORF length, ORF relative length and frequencies of nucleotide pattern to improve the sensitivity and accuracy of CDS and ncRNAs from humans and several other species [32].

In order to attain more specificity and sensitivity, several groups have successfully utilized *k*-mer based deep learning approaches for classification and characterization of CDS and ncRNAs. In one of the studies, it was observed that ncRNAs performing related functions had a similar *k*-mer profile. This observation enabled them to cluster ncRNAs based on their functional profile [33]. Another study used *k*-mer embedding vectors to construct a deep-learning model to distinguish among CDS and ncRNA sequences. This five layered model included Bidirectional Long Short Term Memory (BLSTM) layer, a convolutional neuronal network (CNN) layer and three hidden layers to achieve 96.4% accuracy [34]. Some studies have used *k*-mer approach to differentiate among CDS and ncRNAs as well as classify ncRNAs into relevant families [35].

These approaches are primarily dependent on features which are derived from transcript nucleotide sequences, ORF length, ORF integrity, isoelectric point (pI), gene structure, potential codon sequence, conservation and frequencies of nucleotide pattern to classify CDS from ncRNAs. However, features based on ORFs or protein conservation might be inept because certain ncRNAs especially long non-coding RNAs may have long putative ORFs or evolutionary conserved sequences [37] which result in false classifying of ncRNAs as CDS. Moreover, certain coding transcripts don't have conserved ORFs resulting in false negative findings [38].

Majorly, most of the earlier methods that distinguish CDS and ncRNAs have used balanced positive and negative datasets that have same number of the positive and negative class samples. Since

the number of CDS is many folds as compared to the number of ncRNA sequences, hence generating balanced datasets might lead to loss of highly important features that may have contributed towards their classification. Although *k*-mer based approaches have been explored in several models including deep-learning, these approaches are generally computationally intensive. The complexity of the searches based on the four nucleotides is high. Since DNA consists of four nucleotides 'A', 'T', 'G' and 'C', there are 16 unique permutations (4^2) for a *2*-mer, 64 permutations (4^3) for *3*-mer and 256 permutations (4^4) for *4*-mer that need to be investigated. As the size of *k* grows, it can be observed that the number of permutations grow exponentially, hence the volume of search becomes complex. In this article, we propose a novel binary *k*-mer based approach where the 'A' and 'T' are replaced by '0' whereas 'C' and 'G' are replaced by '1' in both CDS and ncRNA sequences. As a result there are only 4 unique permutations (2^2) for a *2*-mer, 8 permutations (2^3) for *3*-mer and 16 permutations (2^4) for *4*-mer that need to be investigated in our proposed approach. Utilizing the frequency of binary patterns in CDS and ncRNA sequences as features, *k*NN was successfully implemented for the classification of CDS and ncRNAs. Comparison of our proposed binary approach with the conventional ATGC approach, we show that our approach is highly efficient in time and computational resources and results in better performance values in several diverse species. Additionally, we have retained all sequences to capture all possible patterns in the data without loss of crucial information resulting in better MCC as compared to the traditional 'ATGC' approach.

3. Material and methods

3.1. Coding and ncRNA datasets

The CDS as well as ncRNA sequences for all the fifteen species included in this study were manually downloaded from Ensembl [39] in FASTA format on October 30, 2020. In order to capture the genetic diversity, we included the species on the following criteria: (a) both coding and ncRNA data should be available for the species and (b) the species should have been extensively characterized. Based on the above-mentioned criteria's, representative sequences from various diverse species from important classes including mammals, aves, reptiles, amphibians, pisces, insects and nematodes were included. In this study, we mainly focused on the model and highly characterized organisms including *Homo sapiens, Pan paniscus, Macaca mulatta, Mus musculus, Myotis lucifugus, Gallus gallus, Coturnix japonica, Chelonoidis abingdonii, Anolis carolinensis, Xenopus tropicalis, Leptobrachium leishanense, Salmo trutta, Danio rerio, Drosophila melanogaster and Caenorhabditis elegans*.

3.2. Preprocessing of data

3.2.1. Removal of characters other than 'A', 'T', 'G' and 'C'

In order to make sure that the input files did not contain any character other than 'A', 'T', 'G', and 'C', the input file (in FASTA format) was processed using Python script that removes all the characters other than 'A', 'T', 'G', and 'C'. The script was run on both the coding as well as ncRNA input files of all the species.

3.2.2. Removal of short sequences

Once the characters other than 'A', 'T', 'G', and 'C' were removed from the input file, we removed short sequences from both the files. We retained all the sequences greater than 50 nucleotides in our coding and ncRNA files so as to include them in our study. The short sequences were removed because it is very unlikely that such short sequences might have any coding potential [40].

3.3. Conversion of DNA sequences into the proposed binary form

Several groups have investigated CDS and ncRNA sequences using conventional approaches taking all the four nucleotides into account which thereby leads to enhanced complexity. However, we speculated that we could replicate their key findings using only two characters ('0's and '1's) representing all the four nucleotides. Therefore, all the 'A's and 'T's were replaced by '0's whereas 'G's and 'C's were replaced by '1's in both the coding and ncRNA processed files of all the fifteen species. This leads to the generation of binary files where each sequence is represented by a string of '0's and '1's. This replacement decreased the genome complexity as well as requirement of computer hardware to process the full genomic sequences.

3.4. Generation of possible k-mer patterns

The next step towards the analysis of coding and ncRNA binary files involved generating all the possible patterns for a given *k*-mer. For each *k*-mer (k = 3 to k = 18), 2^k permutations of patterns were extracted for both CDS and ncRNAs of all the species. Supplementary Table S1 illustrates the technique for generating all possible permutations (patterns) for a given 3-mer.

3.5. Finding frequency of binary k-mer in CDS and ncRNAs

For every 2^k pattern of a given *k*-mer ($k = 3$ to $k = 18$), sequences were read in an overlapping manner to calculate the corresponding frequency of that pattern in a binary CDS and ncRNAs file. Supplementary Table S1 and Table S2 exhibit the process of formulation of the feature vector in the proposed approach. The approach used in this study can be summarised as follows:

i. Remove characters other than 'A', 'T', 'G' and 'C'
ii. Remove short sequences
iii. Convert sequences into the proposed binary form
iv. Choose a *k* for *k*-mer
v. Formulate the set of all *k*-mer patterns for a chosen k
vi. Enumerate the frequency of each *k*-mer pattern
vii. Find the ratio of each frequency of each pattern over the sum
viii. For a given observation (sequence), each feature vector is the vector of frequency ratios of the particular *k*-mer chosen

All the steps of data pre-processing, binary conversion and calculating frequencies were performed by in-house Python script.

3.6. Machine learning model

We used the *k*-Nearest Neighbor (*k*NN), a well-known classifier to build the decision model. The *k*NN classifier is useful in the CDS and ncRNAs classification as it helps differentiate non-linear classes effectively. The frequency ratios of the *k*-mer of the two classes are discriminative. The *k*NN uses the similarity measure between the feature vectors of CDS and ncRNAs to classify the test samples. Based on the choice of *k*, the number of neighbors, a test data point is classified. The class labels of *k* nearest data points in the training data set are noted and the majority class is assigned to the test data.

Each sequence of the binary file of CDS and ncRNAs was read in an overlapping manner to calculate the frequency of each pattern for a given *k*-mer. The input file contains the sequences along the rows and the frequency ratio of each *k*-mer pattern corresponding to the sequence along the columns. For all the species, the input file was prepared using the inhouse Python script. Our next important task was to find the most suitable *k* (for the *k*-mer) to distinguish CDS from ncRNAs. To determine the optimum *k*-mer, we tested the *k*NN (with 5 neighbors) using the input file at different *k*-mer ($k = 3$

to $k = 10$). The value of k which gave the best performance metrics was chosen as the optimum k for further experiments. In the present study, the positive class (CDS) was labelled as "1" whereas the negative class (ncRNA) was labelled as "0".

3.6.1. Performance measures

For each experiment, 10 independent runs of 10-fold cross validation were performed for the *k*NN classifier, with 5 nearest neighbors. By averaging over the ten runs of the 10-fold cross validation, a realistic performance score was reported for all our experiments. For each experiment, True Positives (TP) is the number of correctly classified CDS, True Negatives (TN) is the number of correctly classified ncRNAs, False Negatives (FN) is the number of CDS predicted as ncRNAs and False Positives (FP) is the number of ncRNAs predicted as CDS. The performance of the ML algorithm was quantified in terms of classification accuracy, precision, recall, F1 score, area under the curve-receiver operating characteristics (AUC-ROC). The classification accuracy, precision, recall, F1 score and MCC are described below.

(i) Classification accuracy: Classification accuracy is the ratio of the correctly classified sequences.

$$\text{Accuracy} = (TP + TN)/(TP + FP + TN + FN)$$

(ii) Precision is the ratio of correctly classified CDS in the set of True and False Positives.

$$\text{Precision} = TP/(TP + FP)$$

(iii) Recall (sensitivity): Recall is the proportion of the correctly classified CDS in the set of all available coding sequences.

$$\text{Recall} = TP/(TP + FN)$$

(iv) F1 score: F1 score is the harmonic mean of precision and recall

$$\text{F1 score} = 2 * ((\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall}))$$

(v) AUC-ROC: The AUC-ROC is commonly used to compare classification models for imbalanced problems.

(vi) Matthews Correlation Coefficient (MCC): MCC is a reliable statistical metric for unbalanced data which gives a high score when most of the positive and the negative predictions are correct.

$$\text{MCC} = (TP*TN - FP*FN)/\text{sqrt}((TP + FP)(TP + FN)(TN + FP)(TN + FN))$$

3.6.2. Optimum k-mer for high classification performance

Our experiments used all the available data of CDS and ncRNAs of each species to build our model using 10 independent runs of 10-fold cross validation. The average results of each performance measure are reported. Through these experiments we found the optimum *k*-mer for classifying the CDS and ncRNAs across the species.

3.7. Efficacy of the binary approach in comparison to the ATGC k-mer approach

Further, to establish that we do not lose any relevant information by converting the 'A', 'T' 'G' and 'C' sequences into binary sequences, we compared the performance of the ATGC as well the binary *5*-mer approach for all the species included in this study. The reason why we have used binary *5*-mer is due to its best average performance across all species. The results of various performance measures (Accuracy, F1 score, Precision, Recall, AUC-ROC and MCC) as well as the time taken for building the feature frequency datasets and the average model building time for all the species for $k = 5$ are reported. These experiments are conducted with the *k*NN classifier with 10 independent runs of 10-fold cross validations.

3.8. Evaluation of the model on novel species

There is a need to classify the CDS and ncRNAs for all possible species, but in the case of some species, due to lack of sufficient and relevant data, ML models cannot be built. The evolutionary relationship among species motivated us for our next set of experiments. The *k*NN model was built with the binary *5*-mer patterns as the features on the human dataset. The model was evaluated across several new species some of which were not part of this study. The MCC scores are also depicted on the time tree which was built using Time Tree web server [41].

4. Results and discussion

4.1. Identification of highly occurring patterns among human CDS

As the first line of study, we used CDS of the human genome to investigate whether any permutation of *3*-mer binary patterns exists across the sequences. We used computational approaches to identify the frequency of all the permutations of a *3*-mer among the human CDS. The process of generation of permutations for a *k*-mer is illustrated in Supplementary Table S1. We downloaded the human CDS (in FASTA format) from Ensembl and converted the DNA sequences into binary form as described above. The resulting human CDS file now consisted of '0's and '1's only. We started our investigation with a set of *3*-mer patterns. In order to find the most commonly occurring patterns, we wrote a Python script to calculate the frequency of all the possible *3*-mer patterns (of '0' and '1') in the human CDS. Our analysis revealed that out of all the eight permutations that arise from permutations of a *3*-mer, the pattern '101' has the highest occurrence followed by the pattern '010' and '011'. The pattern '000' has the lowest occurrence (Fig. 1A, Supplementary Table S1). This data suggested that binary human CDS had a distinct frequency of *3*-mer patterns. It was also observed that the frequency of the highest occurring pattern '101' was 1.74 times higher as compared to the lowest occurring one '000'. This difference in the frequency of various patterns among the human binary CDS pointed towards a possible 'binary signature' and encouraged us to take the approach further. We then investigated human CDS by calculating occurrences of all the possible permutations of *4*-mer, *5*-mer, and so on, up to *18*-mer. We observed that the ratio of the highest occurring pattern to the lowest occurring pattern kept on increasing with the increase in the *k*-mer length (Fig. 1B). A closer inspection of top occurring sequences revealed the repetitive occurrence of the seed sequence i.e., '101' that was obtained from the *3*-mer. We thus identified a stretch of '$(101)_n$' that was being enriched in human CDS (Fig. 1C). This seed sequence is observed to be enriched up to *16*-mer in human CDS. However, in *17*-mer and *18*-mer we can observe that this seed sequence is no longer highly enriched but occupies a second position. Therefore

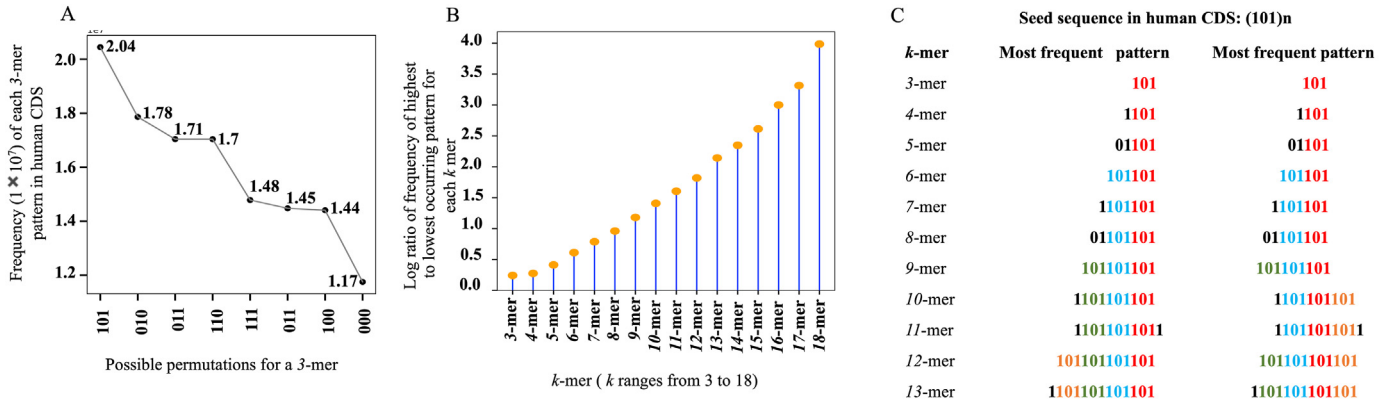

Fig. 1. Identification of highly occurring patterns among human CDS. (A) Plot showing frequencies (occurrences) of all possible permutations (patterns) for a 3-mer in human CDS. The pattern '101' has the highest frequency while the pattern '000' has the lowest frequency amongst all eight permutations. Here, X-axis represents all the possible permutations of a 3-mer whereas the Y-axis represents the frequency of that pattern. (B) Lollipop plot showing the log ratio of frequencies of the highest occurring to the lowest occurring patterns for different *k*-mer. X-axis represents the *k*-mer where value of *k* ranges from 3 to 18. Y-axis represents the log ratio of the highest occurring pattern to the lowest occurring pattern. It can be observed that as the value of *k* increases, the log ratio of the highest occurring pattern to the lowest occurring pattern keeps on increasing. (C) Pictorial representation of two types of possible arrangements of the highest occurring pattern for various *k*-mer ($k = 3$ to 18). It can be observed that the seed sequence '101' is being enriched as the size of *k*-mer increases in the human CDS. Note: *denotes the second most frequent pattern.
Abbreviations: CDS, coding sequences.

our data suggest that there might be some binary patterns among the human CDS.

4.2. Frequent occurring pattern is the characteristic of CDS in several other model organisms

Since analysis of the human CDS led us to the identification of the frequently occurring pattern, we sought to investigate whether we could replicate our findings among other model organisms as well. Therefore, we used similar approach to investigate CDS of several other model organisms including *Pan paniscus, Macaca mulatta, Mus musculus, Myotis lucifugus, Gallus gallus, Coturnix japonica, Chelonoidis abingdonii, Anolis carolinensis, Xenopus tropicalis, Leptobrachium leishanense, Salmo trutta, Danio rerio, Drosophila melanogaster and Caenorhabditis elegans.* The analysis of 3-mer for these organisms, revealed that the pattern '101' was the highest occurring in CDS for most of them (Fig. 2A). Similar results were obtained when the analysis was performed on 4-mer and 5-mer respectively (Fig. 2B and C). As observed in humans, we found that the same patterns were enriched in most of the model organisms. Moreover, the ratio of the highest occurring to the lowest occurring pattern also increased with increase in the *k*-mer among all the species (Fig. 2D). Therefore, we observed that a binary pattern ('101') was highly enriched among CDS of humans and other diverse species. Our results from CDS of various species suggested that there could be some conserved binary signatures among the CDS of species under the investigation. The fact that similar sequences were enriched in multiple species pointed towards the role of these sequences in biological processes.

4.3. Coding and non-coding sequences are characterized by unique occurring patterns

Since we have identified the highest occurring binary pattern among CDS of various species, we sought to investigate whether a similar pattern was enriched among the ncRNAs. Therefore, we planned to investigate ncRNAs of humans and all other species. Similar to the CDS, all the ncRNA sequences were obtained from Ensembl in a FASTA format and pre-processed in exactly the same way to obtain ncRNA binary files. We then performed the *k*-mer analysis on these ncRNA binary sequences to investigate whether there is an enrichment of similar pattern or not. We started our analysis with 3-mer on human ncRNA binary sequences and observed that a particular pattern '000' was highly enriched among them (Fig. 3A). We found that the highly enriched pattern '000' in ncRNA was in fact the least occurring pattern in the human CDS (Fig. 3B). This observation suggested that there was a huge difference among the architecture of human CDS and ncRNA sequences. Apart from the highly occurring sequences, it was also observed that frequency of various other binary patterns was different among CDS and ncRNA sequences. Moreover, similar to the CDS data, the ratio of the highest occurring to the lowest occurring pattern kept on increasing with an increase in the *k*-mer among human ncRNA binary sequences (Fig. 3C). There was an enrichment of repeating pattern '(000)n' among various *k*-mer in human ncRNA sequences (Fig. 3D). The enrichment of seed sequence '000' in case of ncRNA was different as compared to that of CDS where the enrichment of seed sequence '101' was observed. We then investigated ncRNA binary data from the same set of species that was used for CDS analysis and the same pattern '000' was found to be enriched in most of them (Fig. 3E). The frequency of binary patterns for various *k*-mer among CDS and ncRNAs was different and this observation helped us to formulate the strategy in distinguishing among coding and non-coding transcripts with high efficiency.

4.4. Determining the optimum k-mer for high classification performance

Extensive experiments have been performed to study the efficacy of the *k*NN classifier. The *k*NN classifier is suitable when there is no prior information available about the distribution of data and is a useful technique for non-linear data. Being a non-parametric

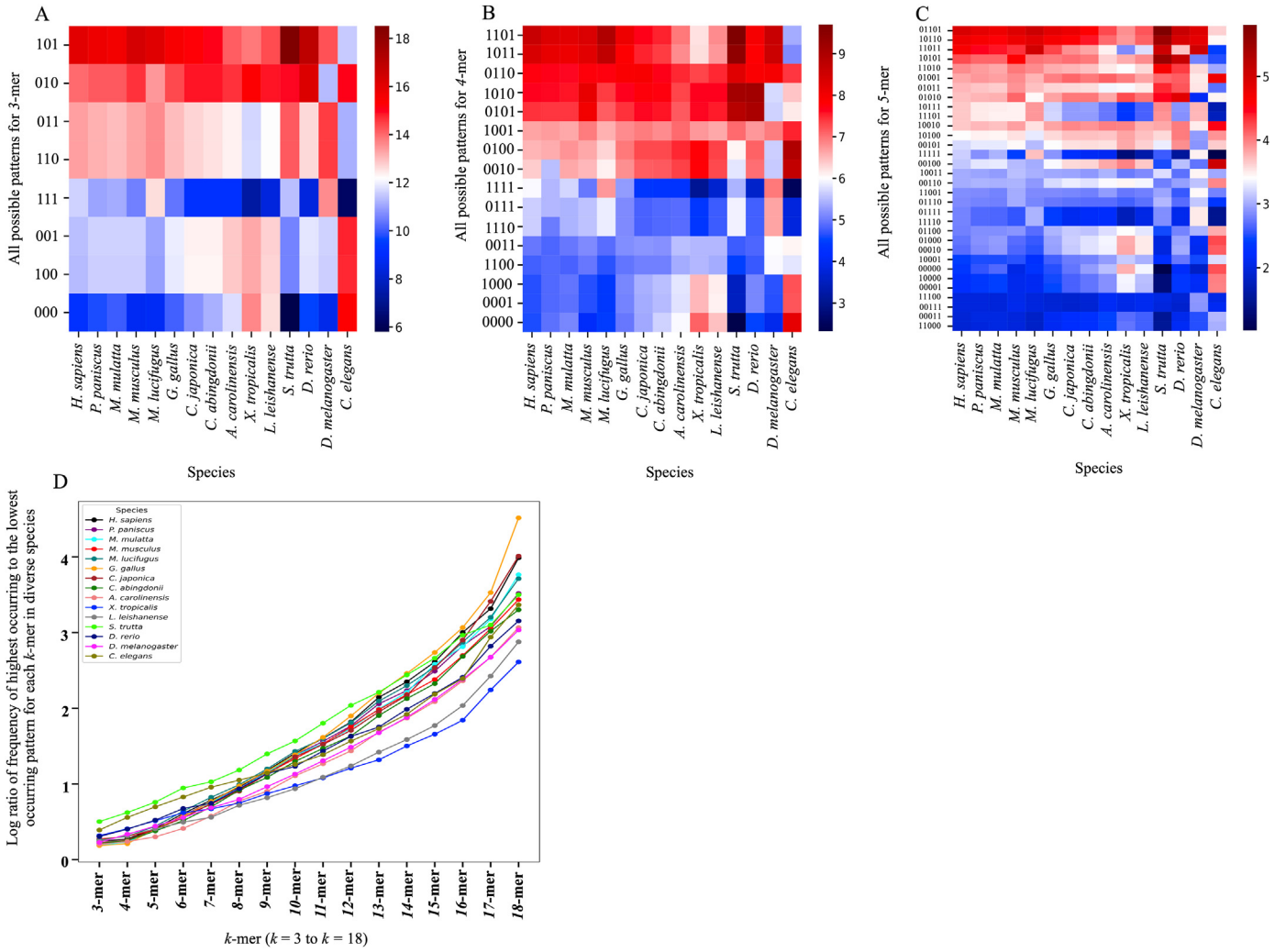

Fig. 2. Frequent occurring binary pattern among human CDS is also enriched in the CDS of 15 diverse species. Heatmap plots representing the frequency (occurrence) of 3-mer (A), 4-mer (B) and 5-mer (C) respectively in CDS of diverse species. The X-axis represents the species name whose CDS sequence is analyzed and the Y-axis represents the frequency of respective binary *k*-mer pattern in the corresponding CDS sequences. Red and blue boxes indicate, respectively, different extents of frequencies in a given CDS (highest to lowest occurring). (D) Plot representing the log ratio of frequency of the highest occurring to the lowest occurring pattern for different *k*-mer in the CDS of 15 diverse species. The X-axis represents the *k*-mer that ranges from *3 to* 18. The Y-axis represents the log ratio of the highest occurring to the lowest occurring representative *k*-mer pattern among the CDS of various species. The ratio of the highest to the lowest occurring pattern among the CDS increases with increase in the *k*-mer.
Abbreviations: CDS, coding sequences.

classifier, it is used widely in ML problems. A thorough study of the optimum *k*-mer for various species was undertaken and the most suitable *k* value for the classifier was determined. Since the data of CDS and ncRNAs of all the species is highly skewed (Table 1), it becomes imperative to measure the performance in terms of metrics that take into consideration the imbalanced data problem. To adjudge the performance of classifiers in such biological data, measures like accuracy tend to give a high value even when the minority class gets heavily mis-classified. Other metrics that are reported are the TP, TN, precision and recall, which alone do not give the true efficacy of the classifier in an unbalanced setup. The metrics like AUC, F1 and MCC are used for classification and evaluation of imbalanced data. AUC is sensitive to class imbalance and does not have a closed form [42]. The F1 score is commonly used for imbalanced data but it is independent of TN and is not symmetric for class swapping [43]. In contrast, MCC generates a high score when the majority of positive data instances and the majority of negative data instances are correctly classified. MCC hence is effective as a relevant and correct measure of the efficiency of a decision algorithm. In our opinion, as illustrated mathematically and empirically in literature, MCC is the most appropriate measure to evaluate an unbalanced dataset [42,43]. We have used all the available data so as not to lose any significant information from the datasets. Some work in literature has randomly selected data to obtain balanced data using very less number of samples, and built machine learning models. This approach may be restrictive in nature as it does not include all possible data and hence the model is not trained as robustly. The high-performance measures reported might be unreliable in these studies. Owing to the preponderance of MCC towards the unbalanced data, we report the performance of the *k*NN classifier in terms of MCC in addition to other performance metrics including Accuracy, Precision, F1 score, Recall and AUC-ROC (Supplementary Table S3).

The performance measures for the fifteen species are reported in Table 2 for various *k*-mer ($k = 3$ to $k = 10$). It can be observed that, for *Homo sapiens*, the highest accuracy and MCC were obtained for *9*-mer. We can observe that the values of accuracy and MCC for all the other species were very encouraging. In case of *P. paniscus, M. mulatta, G. gallus, C. japonica, X. tropicalis*, and *D. melanogaster*, the highest MCC is observed for *5*-mer. For *M. lucifugus, C. abingdonii, A.*

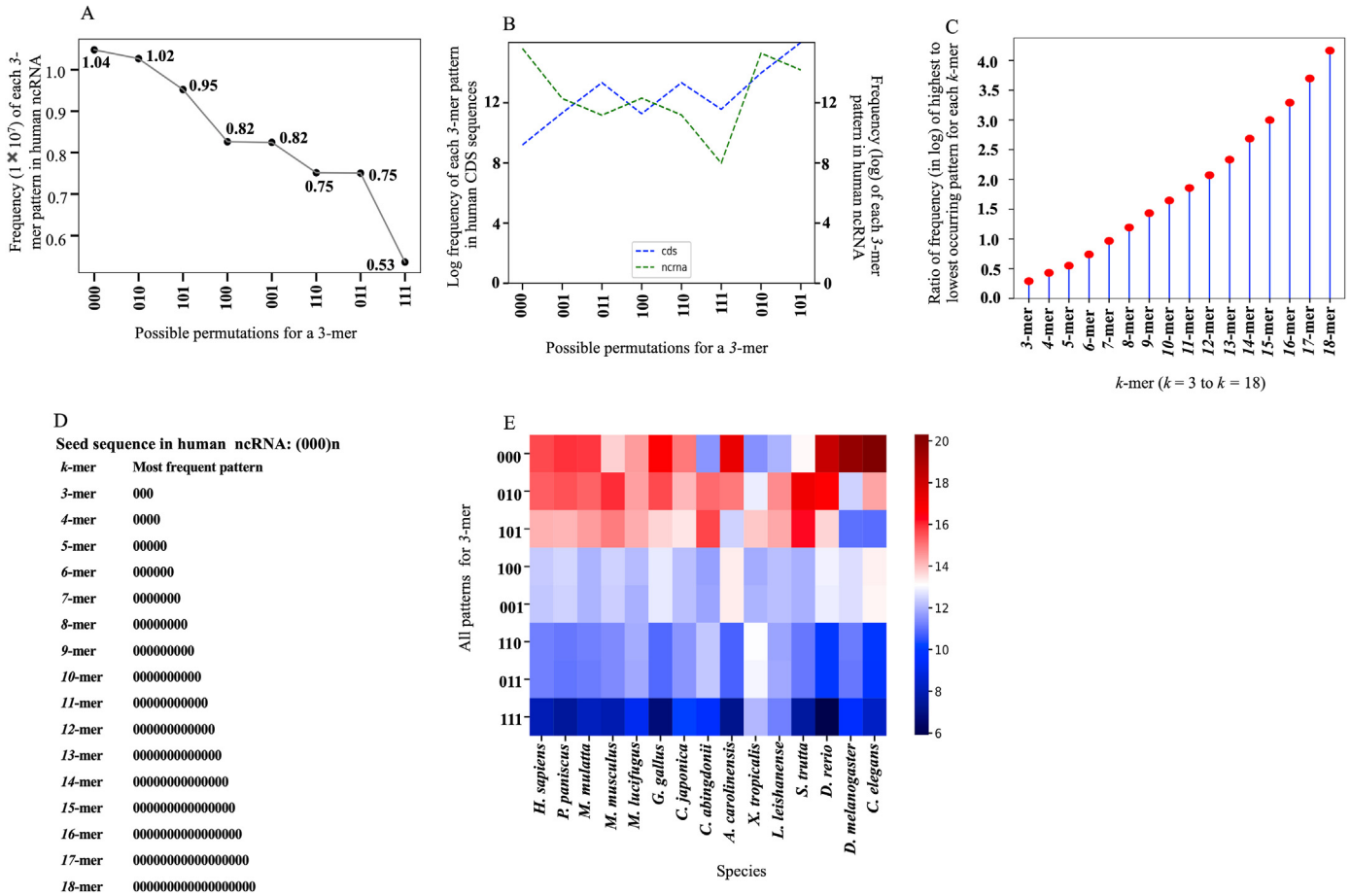

Fig. 3. Coding and non-coding sequences are characterized by unique occurring patterns. (A) Plot showing frequencies (occurrences) of all possible permutations (patterns) for a *3*-mer in human ncRNAs. The pattern '000' has the highest frequency while the pattern '111' has the lowest frequency amongst all eight permutations in human ncRNAs. Here, X-axis represents all the possible patterns of a *3*-mer whereas the Y-axis represents the frequency of that pattern in human ncRNAs. (B) Plot showing frequencies (occurrences) of all possible permutations (patterns) for a *3*-mer in human CDS and ncRNA sequences. Pattern '000' has the highest frequency in human ncRNAs whereas it has the least frequency in the human CDS. Here, X-axis represents all the possible patterns of *3*-mer whereas the LHS Y-axis represents the frequency of the patterns in human CDS and RHS Y-axis represents the frequency of the patterns in human ncRNA sequences. (C) Lollipop plot showing the log ratio of frequencies of the highest occurring to the lowest occurring patterns for different *k*-mer in human ncRNAs. The X-axis represents the *k*-mer that ranges from *3 to* 18. The Y-axis represents the log ratio of the highest occurring to the lowest occurring representative *k*-mer pattern among the ncRNAs of various species. It can be observed that as the value of *k* increases, the log ratio of the highest occurring pattern to the lowest occurring pattern keeps on increasing. (D) Pictorial representation of the possible arrangements of highest occurring pattern for various *k*-mer ($k = 3$ to $k = 18$) in human ncRNAs. It can be observed that the seed sequence '000' is being enriched as the size of *k*-mer increases in the human ncRNAs. (E) Heatmap plots representing the frequency (occurrence) of various *3*-mer in ncRNAs of several diverse species. The X-axis represents the species name whose ncRNA sequence is analyzed and the Y-axis represents all the permutations of *3*-mer and their frequency in the respective ncRNA sequences. Red and blue boxes indicate, respectively, different extents of frequencies in a given ncRNA sequences (highest to lowest occurring). *Abbreviations:* CDS, coding sequences; ncRNAs, non-coding sequences.

Table 1
The final number of coding sequences (CDS) and non-coding RNAs (ncRNAs) after preprocessing.

S.No	Species	Number of Sequences	
		CDS	ncRNAs
1	*Homo sapiens*	111199	58247
2	*Pan paniscus*	43225	9563
3	*Macaca mulatta*	48770	14665
4	*Mus musculus*	68026	22025
5	*Myotis lucifugus*	20714	4402
6	*Gallus gallus*	28442	10429
7	*Coturnix japonica*	27883	9577
8	*Chelonoidis abingdonii*	31797	2954
9	*Anolis carolinensis*	19175	7835
10	*Xenopus tropicalis*	54849	1384
11	*Leptobrachium leishanense*	49452	912
12	*Salmo trutta*	116557	5241
13	*Danio rerio*	52028	8113
14	*Drosophila melanogaster*	30580	3943
15	*Caenorhabditis elegans*	33484	9305

carolinensis, and C. elegans, 4-mer were the most distinguishing patterns. The species *M. musculus, L. leishanense, and D. rerio* performed best for *7*-mer.

Using binary *k*-mer, we have substantially reduced the complexity of the determination of frequency as well as the building of the decision algorithm. To decide on the optimal value of *k*, the average rank over the 15 species was considered. On the basis of the average MCC among all the species that we studied, the efficacy of each *k*-mer was ranked in Fig. 4A. It can be observed that on an average, the features formed using *5*-mer are the most distinguishing feature set for classifying CDS and ncRNAs. We can thus infer that ML models with binary *5*-mer are the most efficient for distinguishing CDS from the ncRNAs.

4.5. Comparison between ATGC and binary 5-mer approach

In literature, CDS and ncRNAs have been identified using the *k*-mer approach based on the 'A', 'T', 'G' and 'C' nucleotides. To gather

Table 2
The accuracy and Matthews Correlation Coefficient (MCC) from $k = 3$ to $k = 10$ among various species. The values in the bold represent the highest values.

Species	3-mer		4-mer		5-mer		6-mer		7-mer		8-mer		9-mer		10-mer	
	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC
H. sapiens	0.795	0.535	0.826	0.607	0.858	0.681	0.6874	0.718	0.883	0.752	0.889	0.752	**0.891**	**0.757**	0.889	0.755
P. paniscus	0.881	0.554	0.895	0.607	**0.897**	**0.620**	0.889	0.582	0.880	0.542	0.872	0.497	0.864	0.455	0.857	0.418
M. mulatta	0.875	0.623	0.899	0.702	**0.906**	**0.73**	0.900	0.704	0.887	0.662	0.873	0.615	0.860	0.573	0.849	0.531
M. musculus	0.810	0.439	0.831	0.508	0.854	0.577	0.863	0.602	**0.867**	**0.611**	0.865	0.609	0.864	0.604	0.859	0.588
M. lucifugus	0.911	0.666	**0.935**	**0.763**	0.931	0.737	0.918	0.694	0.907	0.649	0.899	0.617	0.894	0.588	0.890	0.575
G. gallus	0.862	0.633	0.879	0.680	**0.885**	**0.700**	0.885	0.693	0.876	0.668	0.859	0.624	0.834	0.550	0.808	0.470
C. japonica	0.863	0.623	0.880	0.668	**0.881**	**0.669**	0.872	0.644	0.862	0.615	0.845	0.561	0.828	0.508	0.815	0.459
C. abingdonii	0.929	0.415	**0.936**	**0.475**	0.932	0.424	0.924	0.314	0.921	0.242	0.920	0.221	0.918	0.220	0.919	0.199
A. carolinensis	0.809	0.510	**0.836**	**0.578**	0.835	0.571	0.816	0.523	0.799	0.474	0.786	0.433	0.778	0.407	0.770	0.384
X. tropicalis	0.987	0.699	0.989	0.753	**0.991**	**0.775**	0.990	0.765	0.989	0.762	0.989	0.762	0.989	0.763	0.989	0.761
L. leishanense	0.987	0.560	0.989	0.668	0.992	0.728	0.992	0.742	**0.993**	**0.752**	0.992	0.743	0.992	0.747	0.992	0.745
S. trutta	0.973	0.630	0.976	0.675	0.978	0.711	0.979	0.718	**0.979**	**0.717**	**0.979**	**0.717**	**0.979**	**0.717**	0.979	0.713
D. rerio	0.905	0.542	0.925	0.647	0.934	0.692	0.937	0.704	**0.938**	**0.711**	0.935	0.697	0.934	0.687	0.931	0.670
D. melanogaster	0.943	0.699	0.949	0.725	**0.949**	**0.728**	0.949	0.728	0.948	0.711	0.944	0.692	0.941	0.676	0.937	0.646
C. elegans	0.905	0.706	**0.914**	**0.732**	0.903	0.700	0.879	0.618	0.867	0.535	0.841	0.469	0.834	0.429	0.829	0.407
Average	0.895	0.589	0.910	0.652	**0.915**	**0.669**	0.898	0.649	0.906	0.626	0.899	0.600	0.893	0.579	0.887	0.554

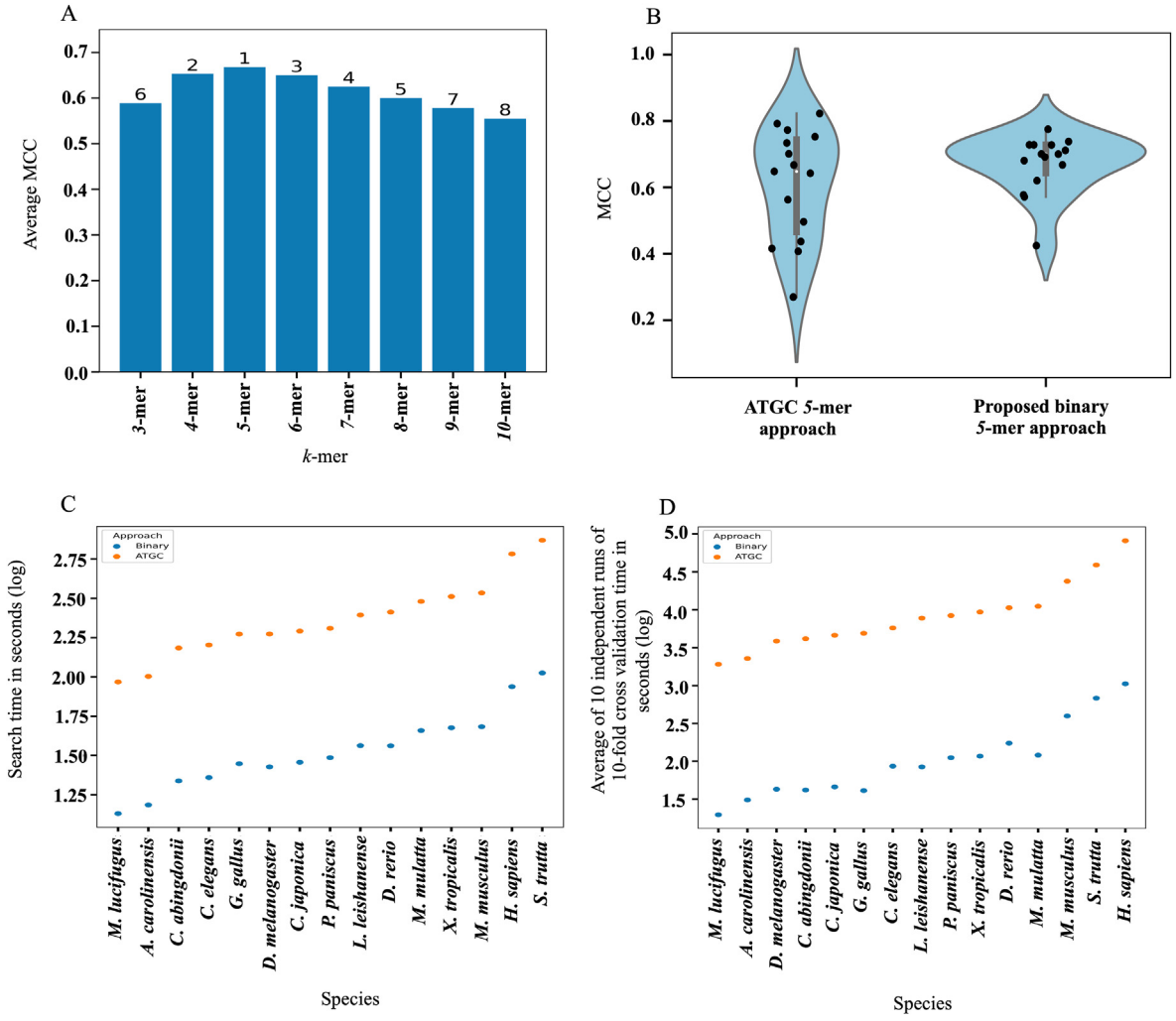

Fig. 4. Determining the most efficient binary *k*-mer for *k*NN classifier and comparison between ATGC and proposed binary approach. (A) Bar plot representing the ranking of MCC score for different *k*-mer ($k = 3$ to $k = 10$) average over the fifteen species using *k*NN classifier (with 5 neighbors). The X-axis represents the number of various *k*-mer that were used in the study and the Y-axis represents the average MCC score. The MCC score that was obtained by including all the species is presented. The models based on the binary 5-mer were found to be the most efficient for distinguishing CDS from the ncRNAs (B) Violin plot showing the average MCC scores (*k*NN with 5 neighbors) of fifteen species using ATGC and binary 5-mer approach. The X-axis represents ATGC and proposed binary approach and the Y-axis represents the MCC score obtained on the 5-mer. (C) A scatter plot for the comparison of time taken by ATGC 5-mer and proposed binary 5-mer approach to search the occurrence of patterns in each species. Blue dots represent the time taken to search the 32 patterns (2^5) generated by the binary 5-mer approach whereas orange dots represent the time taken to search the 1024 patterns (4^5) generated by ATGC 5-mer approach in the CDS and ncRNAs of each species. (D) Scatter plot showing the log average time (in seconds) that was taken for 10 independent runs of 10-fold cross validation. The comparison of 5-mer model building time taken by ATGC and binary approaches for each species is depicted.
Abbreviations MCC, Matthew's Correlation Coefficient; *k*NN, *k*-Nearest Neighbor; CDS, Coding Sequences; ncRNAs, non-coding RNAs.

Table 3
Comparison of 5-mer search time, model building time and performance in terms of Matthews Correlation Coefficient (MCC) among traditional ATGC and proposed binary approach. Results shown below are for 5-mer patterns in diverse species. The values in the bold represents the high MCC values.

Species	5-mer Search (seconds)		Model building (seconds) Average 10-fold cross validation time		MCC	
	ATGC	Proposed Binary	ATGC	Proposed Binary	ATGC	Proposed Binary
H. sapiens	606.170	86.715	81399.000	1053.404	**0.822**	0.681
P. paniscus	203.729	30.635	8379.376	111.655	0.563	**0.620**
M. mulatta	302.192	45.588	11122.520	120.713	0.667	**0.730**
M. musculus	342.657	48.221	23795.240	396.350	**0.648**	0.577
M. lucifugus	92.846	13.493	1908.876	19.726	0.701	**0.737**
G. gallus	187.289	28.011	4895.296	41.076	0.437	**0.700**
C. japonica	195.757	28.618	4610.225	45.925	0.415	**0.669**
C. abingdonii	152.769	21.777	4150.723	41.831	0.269	**0.424**
A. carolinensis	100.748	15.319	2273.838	30.915	0.407	**0.571**
X. tropicalis	324.764	47.434	9344.900	116.880	**0.792**	0.775
L. leishanense	247.689	36.537	7755.518	84.439	**0.773**	0.728
S. trutta	740.901	105.996	38956.190	682.069	**0.752**	0.711
D. rerio	258.719	36.458	10622.830	173.530	**0.734**	0.692
D. melanogaster	187.502	26.765	3861.512	42.826	0.642	**0.728**
C. elegans	159.688	22.879	5753.597	86.269	0.496	**0.700**

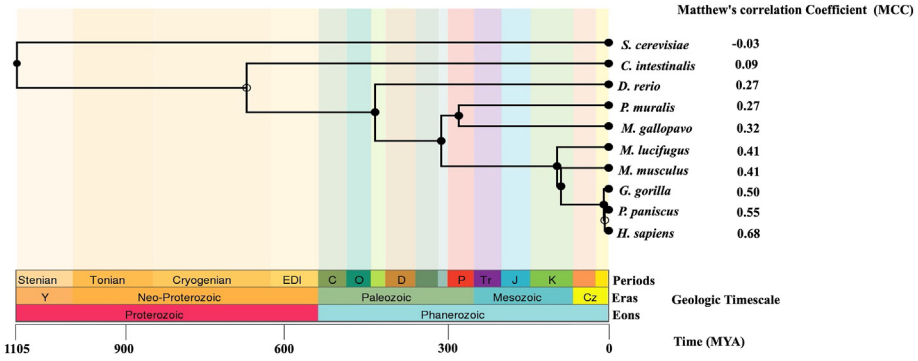

Fig. 5. MCC scores correlate with evolutionary history among several species. Dendrogram illustrating the evolutionary geological time scale of 10 species. The timeline was created using the TimeTree webserver. It can be observed that the MCC scores correlate well with evolutionary history among these species. The MCC values of *k*NN (with 5 neighbors) model trained on *H. sapiens* and tested on nine diverse species correlate with their evolution. It can be observed that the species that are closer to *H. sapiens* in evolution have higher MCC values as compared to those that are farther from *H. sapiens.*
Abbreviations MCC, Matthews Correlation Coefficient; *k*NN, *k*-Nearest Neighbor.

the confidence in our binary *k*-mer approach, we illustrate the difference in the conventional ATGC and the proposed binary approach on the basis of performance metrics. We conducted experiments with the 1024 possible patterns for the conventional 5-mer (ATGC approach) to classify the CDS and ncRNAs of fifteen species, using the *k*NN classifier. This experiment was compared to the results of the binary 5-mer approach with the 32 patterns only. This was to understand which one of the two approach is superior. Table 3 illustrates the MCC reported in both the cases. The average MCC of all the species in conventional ATGC 5-mer was 0.608 whereas it was 0.669 in case of binary 5-mer (Fig. 4B). However, it can be seen that MCC values increased by nearly one and a half times of conventional ATGC approach in *G. gallus* and *C. elegans* thereby supporting the robustness of our proposed binary approach. The time complexity is also tabulated in Table 3. The traditional approach took an average of 273.56 s to search for the 5-mer patterns over fifteen species whereas it took 39.62 s to process the data among the same set of species using our proposed binary approach. Similarly, the traditional approach took an average of 14588.64 s to build the *k*NN model for the 5-mer whereas the same model was built in 203.17 s using our binary approach (Fig. 4C and D, Supplementary Table S4). This clearly suggests that binary 5-mer with only 32 permutations out-performed the ATGC approach. In fact, the time complexity of the proposed approach is extremely less and the performance is highly encouraging.

4.6. MCC scores correlates with evolutionary history among some species

Having established the success of the binary *k*-mer approach as an effective way to identify coding and non-coding sequences, we enhance this study by testing the ability of the ML model built by us. The ML model trained on human dataset was evaluated across several species such as *Pan paniscus*, *Gorilla gorilla*, *Mus musculus*, *Myotis lucifugus*, *Meleagris gallopavo*, *Podarcis muralis*, *Danio rerio*, *Ciona intestinalis* and *Saccharomyces cerevisiae*. As can be seen in Fig. 5 and Supplementary Table S5, an ML model trained on humans gives high MCC for species closer to humans in the geological time scale. As humans diverge from other species in geological time scales, MCC score gradually decreases. For instance - MCC scores of reptiles (*P. muralis*: 0.27), fish (*D. rerio*: 0.27) and tunicates (*C. intestinalis*: 0.09) are lower than birds (*M. gallopavo*: 0.32) and mammals (*P. paniscus*: 0.55) which directly correlates the phylogenetic nearness of these species. It further emphasizes the fact that the model built on one species can be used to evaluate coding and non-coding transcripts of other species which are phylogenetically close to it. This data further attest to our binary approach where the MCC values correlated with the phylogenetic nearness of the species.

5. Conclusion

The major contributions of this paper are (i) distinguishing binary *k*-mer patterns were discovered, that were highly enriched in CDS and ncRNAs, (ii) a binary *k*-mer approach was devised to identify frequency of 'binary signatures' among CDS and ncRNAs thereby enabling us to use them as the features that effectively discriminate CDS and ncRNAs. This approach decreased the genome complexity as well as requirement of computer hardware to process the full genomic sequences consisting of four characters. Here, 2^k permutations were searched against the 4^k permutations of conventional ATGC *k*-mer approach, (iii) utilizing the complete data sets to build a robust ML model for classification. This approach utilizes all the crucial information in the CDS as well as ncRNAs as compared to the models built on a balanced dataset with only a few randomly chosen data samples and (iv) since next generation sequencing is generating huge amounts of data consisting of thousands of novel transcripts, our proposed method facilitates classification of novel CDS and ncRNAs. This can be achieved by building models of related known species and testing species that are phylogenetically closer to them.

Author contributions

VS conceived the idea. NP and PS performed the experiments. BK contributed to the ML setup and experiments. PA, SP, BK and VS supervised the study. VS wrote the first draft. PA, BK and VS edited and finalized the draft.

Declaration of competing interest

Authors declare no conflict of interest.

Acknowledgements

NP and PS are recipients of PhD fellowship from UGC and CSIR respectively. VS is thankful to UGC for the Faculty Recharge award and start-up grant. We are thankful to Ishpreet Singh and Bhuvnesh Tanwar for their assistance in optimizing the Python scripts.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biochi.2022.04.012.

References

[1] National Research Council, Mapping and Sequencing the Human Genome, National Academies Press, 1988.
[2] L. Statello, C.-J. Guo, L.-L. Chen, M. Huarte, Gene regulation by long non-coding RNAs and its biological functions, Nat. Rev. Mol. Cell Biol. 22 (2021) 96–118.
[3] G. Liu, J. Mattick, R.J. Taft, A meta-analysis of the genomic and transcriptomic composition of complex life, Cell Cycle 12 (2013) 2061–2072.
[4] B. Kamalidehghan, M. Habibi, S.S. Afjeh, M. Shoai, S. Alidoost, R.A. Ghale, N. Eshghifar, F. Pouresmaeili, The importance of small non-coding RNAs in human reproduction: a review article, Appl. Clin. Genet. 13 (2020) 1.
[5] A. Pauli, J.L. Rinn, A.F. Schier, Non-coding RNAs as regulators of embryogenesis, Nat. Rev. Genet. 12 (2011) 136–149.
[6] A. Mehta, D. Baltimore, MicroRNAs as regulatory elements in immune system logic, Nat. Rev. Immunol. 16 (2016) 279–294.
[7] C.P. Bracken, H.S. Scott, G.J. Goodall, A network-biology perspective of microRNA function and dysfunction in cancer, Nat. Rev. Genet. 17 (2016) 719–732.
[8] R. Mishra, A. Kumar, H. Ingle, H. Kumar, The interplay between viral-derived miRNAs and host immunity during infection, Front. Immunol. 10 (2020) 3079.
[9] P. Wang, The opening of pandora's box: an emerging role of long noncoding RNA in viral infections, Front. Immunol. 9 (2019) 3138.
[10] N. Sharma, S.K. Singh, Implications of non-coding RNAs in viral infections, Rev. Med. Virol. 26 (2016) 356–368.
[11] P. Waller, A. Blann, Non-coding RNAs—A primer for the laboratory scientist, Br. J. Biomed. Sci. 76 (2019) 157–165.
[12] H. Long, X. Wang, Y. Chen, L. Wang, M. Zhao, Q. Lu, Dysregulation of microRNAs in autoimmune diseases: pathogenesis, biomarkers and potential therapeutic targets, Cancer Lett. 428 (2018) 90–103.
[13] R. Ojha, R. Nandani, R.K. Pandey, A. Mishra, V.K. Prajapati, Emerging role of circulating microRNA in the diagnosis of human infectious diseases, J. Cell. Physiol. 234 (2019) 1030–1043.
[14] L. Tribolet, E. Kerr, C. Cowled, A.G. Bean, C.R. Stewart, M. Dearnley, R.J. Farr, MicroRNA biomarkers for infectious diseases: from basic research to biosensing, Front. Microbiol. 11 (2020) 1197.
[15] Y. Xiao, T. Xiao, W. Ou, Z. Wu, J. Wu, J. Tang, B. Tian, Y. Zhou, M. Su, W. Wang, LncRNA SNHG16 as a potential biomarker and therapeutic target in human cancers, Biomarker Research 8 (2020) 1–11.
[16] Z. Chen, L.-L. Wei, L.-Y. Shi, M. Li, T.-T. Jiang, J. Chen, C.-M. Liu, S. Yang, H. Tu, Y. Hu, Screening and identification of lncRNAs as potential biomarkers for pulmonary tuberculosis, Sci. Rep. 7 (2017) 1–10.
[17] A.D. Pandey, S. Goswami, S. Shukla, S. Das, S. Ghosal, M. Pal, B. Bandyopadhyay, V. Ramachandran, N. Basu, V. Sood, Correlation of altered expression of a long non-coding RNA, NEAT1, in peripheral blood mononuclear cells with dengue disease progression, J. Infect. 75 (2017) 541–554.
[18] P.L. Wang, Y. Bao, M.-C. Yee, S.P. Barrett, G.J. Hogan, M.N. Olsen, J.R. Dinneny, P.O. Brown, J. Salzman, Circular RNA is expressed across the eukaryotic tree of life, PLoS One 9 (2014), e90859.
[19] A. Ivanov, S. Memczak, E. Wyler, F. Torti, H.T. Porath, M.R. Orejuela, M. Piechotta, E.Y. Levanon, M. Landthaler, C. Dieterich, Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals, Cell Rep. 10 (2015) 170–177.
[20] W.R. Jeck, J.A. Sorrentino, K. Wang, M.K. Slevin, C.E. Burd, J. Liu, W.F. Marzluff, N.E. Sharpless, Circular RNAs are abundant, conserved, and associated with ALU repeats, RNA 19 (2013) 141–157.
[21] T. Tagawa, S. Gao, V.N. Koparde, M. Gonzalez, J.L. Spouge, A.P. Serquiña, K. Lurain, R. Ramaswami, T.S. Uldrick, R. Yarchoan, Discovery of Kaposi's sarcoma herpesvirus-encoded circular RNAs and a human antiviral circular RNA, Proc. Natl. Acad. Sci. Unit. States Am. 115 (2018) 12805–12810.
[22] J. Huang, J. Chen, L. Gong, Y. Bi, J. Liang, L. Zhou, D. He, C. Shao, Identification of virus-encoded circular RNA, Virology 529 (2019) 144–151.
[23] Y. Li, U. Ashraf, Z. Chen, D. Zhou, M. Imran, J. Ye, H. Chen, S. Cao, Genome-wide profiling of host-encoded circular RNAs highlights their potential role during the Japanese encephalitis virus-induced neuroinflammatory response, BMC Genom. 21 (2020) 1–11.
[24] L.S. Kristensen, M.S. Andersen, L.V. Stagsted, K.K. Ebbesen, T.B. Hansen, J. Kjems, The biogenesis, biology and characterization of circular RNAs, Nat. Rev. Genet. 20 (2019) 675–691.
[25] I. Li, Y.G. Chen, Emerging roles of circular RNAs in innate immunity, Curr. Opin. Immunol. 68 (2021) 107–115.
[26] L. Verduci, E. Tarcitano, S. Strano, Y. Yarden, G. Blandino, CircRNAs: role in human diseases and potential use as biomarkers, Cell Death Dis. 12 (2021) 1–12.
[27] L. Kong, Y. Zhang, Z.-Q. Ye, X.-Q. Liu, S.-Q. Zhao, L. Wei, G. Gao, CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine, Nucleic Acids Res. 35 (2007) W345–W349.
[28] K. Sun, X. Chen, P. Jiang, X. Song, H. Wang, H. Sun, iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data, BMC Genom. 14 (2013) 1–10.
[29] L. Sun, H. Luo, D. Bu, G. Zhao, K. Yu, C. Zhang, Y. Liu, R. Chen, Y. Zhao, Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts, Nucleic Acids Res. 41 (2013) e166–e166.
[30] A. Li, J. Zhang, Z. Zhou, PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme, BMC Bioinf. 15 (2014) 1–10.
[31] R. Achawanantakun, J. Chen, Y. Sun, Y. Zhang, LncRNA-ID: long non-coding RNA IDentification using balanced random forests, Bioinformatics 31 (2015) 3897–3905.
[32] H.W. Schneider, T. Raiol, M.M. Brigido, M.E.M. Walter, P.F. Stadler, A support vector machine based method to distinguish long non-coding RNAs from protein coding transcripts, BMC Genom. 18 (2017) 1–14.
[33] J.M. Kirk, S.O. Kim, K. Inoue, M.J. Smola, D.M. Lee, M.D. Schertzer, J.S. Wooten, A.R. Baker, D. Sprague, D.W. Collins, Functional classification of long non-coding RNAs by k-mer content, Nat. Genet. 50 (2018) 1474–1482.
[34] X.-Q. Liu, B.-X. Li, G.-R. Zeng, Q.-Y. Liu, D.-M. Ai, Prediction of long non-coding RNAs based on deep learning, Genes 10 (2019) 273.
[35] J. Wen, Y. Liu, Y. Shi, H. Huang, B. Deng, X. Xiao, A classification model for lncRNA and mRNA based on k-mers and a convolutional neural network, BMC Bioinf. 20 (2019) 1–14.
[36] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theor. 13 (1967) 21–27.
[37] J. Ruiz-Orera, M.M. Albà, Conserved regions in long non-coding RNAs contain abundant translation and protein–RNA interaction signatures, NAR Genomics and Bioinformatics 1 (2019) e2–e2.
[38] M.W. Orr, Y. Mao, G. Storz, S.-B. Qian, Alternative ORFs and small ORFs: shedding light on the dark proteome, Nucleic Acids Res. 48 (2020) 1029–1042.
[39] K.L. Howe, P. Achuthan, J. Allen, J. Allen, J. Alvarez-Jarreta, M.R. Amode, I.M. Armean, A.G. Azov, R. Bennett, J. Bhai, Ensembl, Nucleic Acids Res. 49 (2021) D884–D891, 2021.

[40] F. Mignone, A. Anselmo, G. Donvito, G.P. Maggi, G. Grillo, G. Pesole, Genome-wide identification of coding and non-coding conserved sequence tags in human and mouse genomes, BMC Genom. 9 (2008) 1–12.
[41] S. Kumar, G. Stecher, M. Suleski, S.B. Hedges, TimeTree: a resource for time-lines, timetrees, and divergence times, Mol. Biol. Evol. 34 (2017) 1812–1819.
[42] S. Boughorbel, F. Jarray, M. El-Anbari, Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric, PLoS One 12 (2017), e0177678.
[43] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genom. 21 (2020) 1–13.