Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 29 April 2021
Project Report

OXFORD

Motif Analysis

Identifying RNA Motifs of Specific RNA Motif Families Using Supervised and Unsupervised Machine Learning Algorithms

Md Mahfuzur Rahaman and Nabila Shahnaz Khan

Department of Computer Science, University of Central Florida, Orlando, Florida, 32816, USA.

Abstract

Motivation: In order to understand the functionality of genes and cell processes, identifying motif is very important. Though there are some existing tools and research works for identifying motifs, but no tool has been developed yet for identifying motif families based on their existing family patterns. Here, in this research work, we have tried to develop a model which will be able to identify motifs belonging to known motif families such as Kink-turn, Sarcin-ricin, Tandem-shear. For that, we have considered the sequential and structural features of the motifs and explored both supervised and unsupervised machine learning algorithms.
Results: Using supervised learning, we were able to achieve a high accuracy in predicting the family of a motif while unsupervised learning have shown moderate performance. In conclusion, supervised learning performs better in predicting motif families with an accuracy higher than 92%.
Code Availability: The codes for supervised learning and unsupervised learning are available at https://github.com/NabilaKhan/CAP_6545_Project
Data Availability: The Dataset is available at https://github.com/NabilaKhan/CAP_6545_Project

1 Introduction

Every living organism, from bacteria to human, has RNA that plays a wide variety of biological roles in different functions. According to the definition given by Brosiusa et al., transcripts or its processing products having specific functionality can be regarded as RNA [1]. Normally, it forms in the nucleolus and moves to specific regions based on its function. It is also important for drug design and RNA-based therapeutics. RNA is normally single-stranded in structure but due to different types of functions, it can fold and form a more stable 3D structure containing base pairs and stacks. In this structure, it might contain different kinds of loops like Hairpin, Internal or Multi loops.

Motifs are the recurring patterns in the genome and are also known to be the binding sites of transcription factors. They are known to contain hierarchically organized modular building blocks of RNA structure [2]. However, the definition of motif might vary depending on the research context. Here, we're defining Hairpin, Internal or Multi loops of RNA as the 3D structure of RNA motifs [3]. Based on the structures of these loops, the motifs are divided into several families. Some well-known motif families are Kink-turn, reverse Kink-turn, Sarcin-ricin, Tandem-shear, etc [4].

Till-date, motifs are considered to be critical components of RNA structure-function relationships and they might be capable of providing some important insight into the functionality of uncharacterized RNAs [5]. In order to understand the functionality of genes and cell processes, identifying motif is very important. Discovering motif is an open challenge to date. Generally, motif discovery tools like MEME [6], MEME-Chip [7], DREME [8] use approaches based on comparative genomics, motif profile search, and statistical analysis. But they might lack in both efficiency and accuracy. Also, tertiary structural information of motifs is not taken into consideration in most of the cases in spite of the well-known fact that, 3D structural information plays a vital role in RNA-protein binding. According to research, loop regions which are considered to be 3D structural motifs, are functional sites of an RNA that provide specific binding locations for proteins or other molecules [9]. Hence tertiary structure can be an important factor in motif identification. Furthermore, no such tool exists till now that can predict if a given RNA chain loop belongs to a certain renowned motif family or not.

Here, for our project, we have developed a model that will be able to identify motifs of specific motif families using the machine learning technique. For now, we are only focusing on some well-known motif families such as Kink-turn, Sarcin-ricin, Tandem-shear, etc. For developing the model, we have explored different supervised and unsupervised machine learning algorithms and compared their

performances. Finally, we selected the model with high accuracy and more flexibility for developing the tool that can predict specific RNA motif families. In our case, the proposed Ensemble Learner Base Classifier (ELBC) model seemed to have overall good performance with more flexibility for future improvements.

2 Methodology

For developing a predictive model of motifs of specific RNA families, first we had to collect motifs belonging to different well-known RNA motif families from different data sources as there was no such preexisting dataset. After that, we generated features based on their sequential and tertiary structural information. Then we performed normalization and scaling to improve data quality. Finally, we implemented different supervised and unsupervised learning algorithms and analyzed their results to see which algorithm might work well for our proposed model. The overall procedure has been described below step by step:

2.1 Data Collection

We have collected the information about motif instances belonging to Kink-turn, Reverse Kink-turn and Sarcin-ricin families from the research works as mentioned above [10, 11, 12]. Further motif instances of these families are collected from RNAMotifScan [13], RNAMotifScanX [14], RNA 3D Motif Atlas database [15], and RNAMotifContrast [16]. Finally, we collected the sequential and 3d structural information of these motif instances from the well-known PDB database [17] and extracted features as per requirement.

2.1.1 Motif Location Collection

To get the locations of the motif instances of the Kink-turn family, we followed the research work of Professor David M.J. Lilley [10]. They have provided several locations of Kink-turn motif instances in different RNAs. We have also found such locations for reverse Kink-turn and Sarcin-ricin from [11] and [12] respectively. The motif instances described in these papers is not enough to train our machine learning model. However, we have found some other research works that contain the locations of lots of such motif instances. Zhong et. al. [13] have done wonderful work on comparing and searching using RNA tertiary motifs and found a significant number of known motif instances. They have listed those locations on their website which we have used as our data source. Again, Petrov et. al. [15] have also worked on these motifs and provided a bunch of locations for such motif instances. Finally, we have got a significant number of known motif locations from the work of Islam et. al. [16].

2.1.2 Motif Sequence and 3D Structure Collection

We collected the specific PDBx file from the PDB database [17], annotated them using DSSR [18] and FR3D [19] annotation tools, and extracted the sequential and structural information (e.g. co-ordinates, base-pairings, base-stackings, etc.) of the motifs from specific RNA chains. We have also collected the FASTA files of the used PDBs to get the mapped indices between PDB and FASTA, and managed to generate the sequence information for the missing residues.

2.1.3 Collection of Testing Dataset

To test our machine learning model with unseen data, we needed to have motifs from other motif families or such loops which do not belong to any known motif family yet. And so, after getting the annotation information, we cut loops (internal or hairpin loops) from the randomly selected RNA chains using a single canonical W/W cis or more than one non-canonical W/W interaction as loop boundary. The procedure is the same to get motifs of the known families as well as unknown loops.

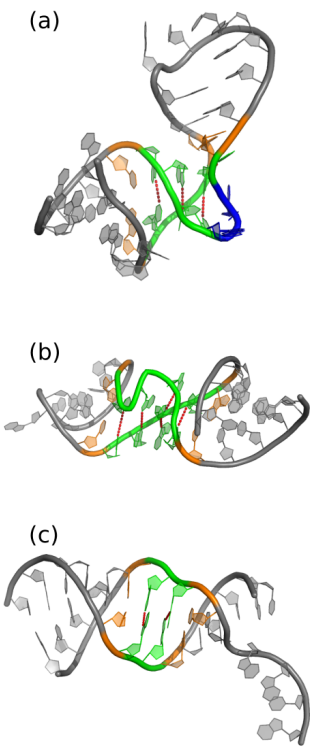

Fig. 1. Representative motifs for (a) Kink-turn, (b) Sarcin-ricin, and (c) Tandem-shear motif family.

2.2 Feature Extraction

To train our machine learning models (both supervised and unsupervised), we have extracted a couple of secondary and tertiary structural features. The basic features include Motif length and GC-percentage. For the 3D structural information, we have generated a representative motif of our selected families by using RNAMotifContrast [16] tool. In the process of generating representatives, we have used the motifs provided in the same work as a benchmark dataset. Figure 1 shows the representative loops for Kink-turn, Sarcin-ricin, and Tandem-shear motif families.

From the representative motifs discussed above, we have calculated sequence identity, sequence alignment score, 3D alignment score, and alignment length (3D) for each input loops with respect to each of the representative motifs. From the 3D structure alignment of RNAMotifScanX [14], we have collected Root Mean Square Deviation (RMSD), matching base-pairs and base-stackings to use as features. We have also kept the count of total number of base-pairs and base-stackings of a loop as a feature.

2.3 Data Analysis

2.3.1 Data Description

In total, the generated dataset has 33 independent features and one dependent feature ‘motif_family’ which is the label column. It has 1877 data objects which come from five separate classes: ‘Kink-turn’, ‘Sarcin-ricin’, ‘Tandem-shear’, ‘Known-motif’, ‘Unknown-motif’. Here ‘Kink-turn’, ‘Sarcin-ricin’, ‘Tandem-shear’ classes represent motifs from motif families Kink-turn, Sarcin-ricin and Tandem-shear respectively. Class ‘Known-motif’ contains motifs from some other motif families such

as reverse Kink-turn, C-loop, E-loop, Hook-turn, Tetraloop-receptor, L1-complex and Rope-sling [16]. And class 'Unknown-motif' contains loops from RNA chains that do not belong to any predefined motif families yet. All the attributes here are numerical except the identifier and the label.

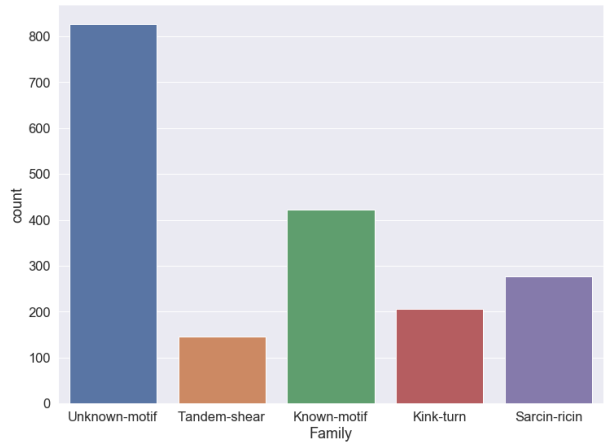

Fig. 2. Number of loops in known and unknown motifs family dataset.

2.3.2 Data Bias Check

In order to avoid confirmation data bias, the dataset was analyzed and it was found that the data was highly biased towards the class 'Unknown-motif'. The distribution of instances in the five classes is shown in Figure 2 There are 826 data objects of class 'Unknown-motif', 422 in 'Known-motif', 277 in 'Sarcin-ricin', 206 in Kink-turn' and lastly only 146 in class 'Tandem-shear'.

2.3.3 Feature Analysis

In order to understand the impact of the features over the motif family predictor model, two types of feature analysis have been done. They are: (1) Feature Correlation Analysis (2) Feature Importance Analysis.

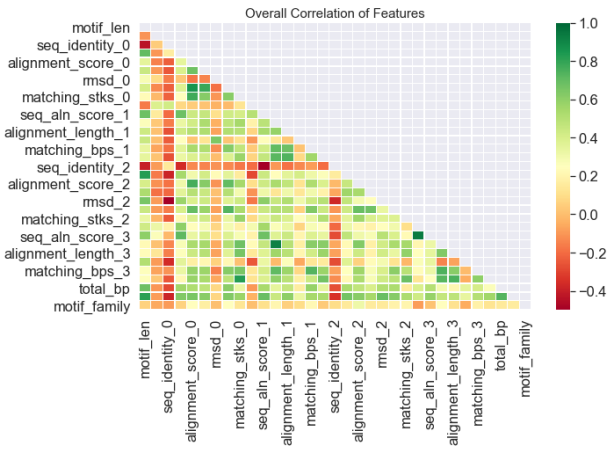

Fig. 3. Correlation of extracted features.

1. Feature Correlation Analysis: The overall correlation of all the features have been analyzed and depicted in Figure 3. As we can see from the figure, though there were some features showing high correlation value, but no significant correlation was found of the dependent label feature 'motif_family' with the other independent features. So, based on the feature correlation analysis, no effectual decision could be made.

2. Feature Importance Analysis: In order to figure out the less important features among the 32 numerical features, a feature importance analysis was performed using different machine learning models. The models that have been used in this analysis are: Logistic Regression, Decision Tree, Random Forest, Extreme Gradient Boosting Model (XGBM) and Light Gradient Boosting Model (LGBM). Figure 4 shows the ranking of the features based on feature score generated by Logistic Regression Model. The ranking of the features based on models Decision Tree and Random Forest has been shown in Figure 11 and similarly the ranking has been shown for models XGBoost and LGBM in Figure 12. From the list of important features, a list of commonly least important features were generated. Later, dropping those features improved the accuracy of the model significantly.

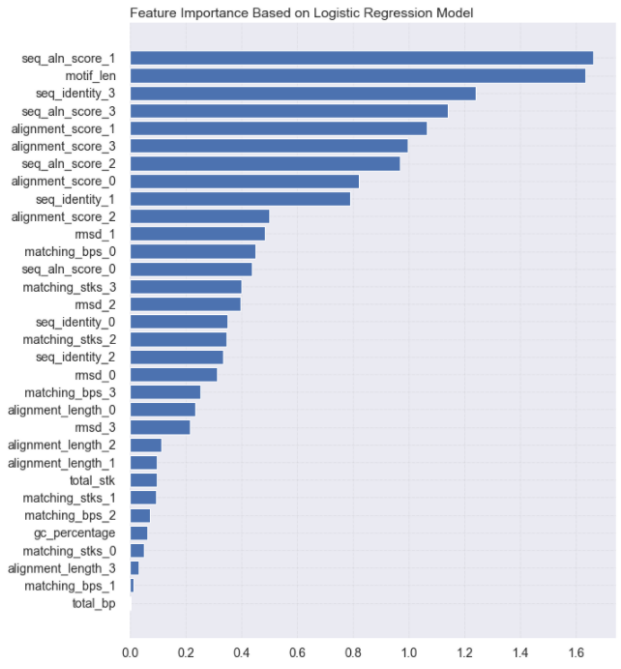

Fig. 4. Feature importance analysis based on logistic regression model.

2.4 Data Preprocessing and Normalization

As the range of values differ for all the extracted continuous features, standard scaling has been preformed on the dataset which improved the model accuracy. Apart from that, skewness of all the features were checked. The normalization of the features were analyzed and most of the features were found to be symmetric except the features 'alignment_score_0', 'seq_aln_score_2', 'alignment_length_2'. These three features were slightly right skewed with a skew score of 2.06, 2.02 and 2.45 respectively. The data distribution of these three features have been shown in the Figure 5. However, after performing Logarithmic Transform on these three slightly right skewed features, the model accuracy didn't improve. So, the features haven't been transformed in the final model pipeline.

2.5 Implementing Supervised Learner

For supervised learning, initially we implemented a Random Forest classifier and a XGBoost classifier. We tested the accuracy of these two models for various conditions and combinations of the dataset to figure

out which situation works best for the predictive model. The following conditions were considered while testing the models:

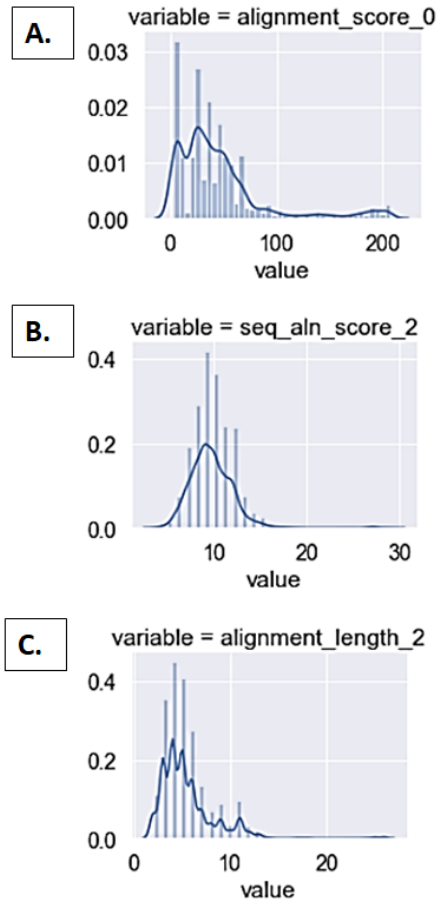

Fig. 5. Data distribution of sequence alignment score, 3d alignment score, and alignment length.

- Condition 1: Considering class 'Kink-turn', 'Sarcin-ricin', 'Tandem-shear', 'Known-motif', 'Unknown-motif' without bias handling
- Condition 2: Considering class 'Kink-turn', 'Sarcin-ricin', 'Tandem-shear', 'Known-motif', 'Unknown-motif' with bias handling
- Condition 3: Considering class 'Kink-turn', 'Sarcin-ricin', 'Tandem-shear', 'Known-motif' without bias handling
- Condition 4: Considering class 'Kink-turn', 'Sarcin-ricin', 'Tandem-shear', 'Known-motif' with bias handling
- Condition 5: Considering class 'Kink-turn', 'Sarcin-ricin', 'Tandem-shear', 'Unknown-motif' without bias handling
- Condition 6: Considering class 'Kink-turn', 'Sarcin-ricin', 'Tandem-shear', 'Unknown-motif' with bias handling
- Condition 7: Considering class 'Kink-turn', 'Sarcin-ricin', 'Tandem-shear' without bias handling
- Condition 8: Considering class 'Kink-turn', 'Sarcin-ricin', 'Tandem-shear' with bias handling

Here, different combination of data classes were considered to see how well the model can recognize patterns of different classes. 'With bias handling' means removing data bias towards specific labels by randomly sampling data of each class where each class contains only 146 members. As shown before in Figure 2, the class Tandem-shear is the lowest populated class containing only 146 number of data objects, hence only 146 members were selected randomly for each class. The 10-fold accuracy comparison of the two models Random Forest classifier and XGBoost classifier in different conditions are shown in Table 1:

Table 1. Accuracy comparison of Random Forest and XGBoost in different conditions

Condition	10-fold Accuracy	
	Random Forest	XGBoost
Condition 1	76.93%	73.94%
Condition 2	84.79%	83.56%
Condition 3	82.60%	81.93%
Condition 4	88.00%	87.67%
Condition 5	87.63%	86.32%
Condition 6	87.49%	87.14%
Condition 7	90.28%	89.64%
Condition 8	**92.70%**	**92.00%**

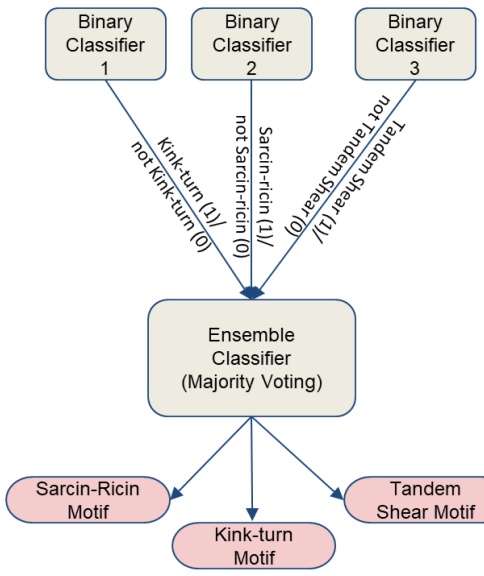

Fig. 6. Overall architecture of the ensemble learner.

As we can see from Table 1, both the models show best performance under condition 8. So, finally we implemented our ensemble learner based on Naive Bayes classifier using condition 8. In the ensemble learner, we first implemented three binary classifiers (Binary Classifier 1, Binary Classifier 2 and Binary Classifier 3). Binary Classifier 1 outputs if the given motif is kink-turn or not, Classifier 2 predicts if it's Sarcin-Ricin or not and Classifier 3 outputs if it's Tandem-shear or not. Finally, combining the three classifiers, we built an ensemble classifier which predicts if the motif belongs to Kink-turn, Sarcin-ricin or Tandem Shear family. The overall architecture of the ensemble learner has been shown in Figure 6. Though the accuracy of the ensemble classifier is slightly less than the Random forest and XGBoost classifier, but overall the accuracy is more than 90% which is pretty good. Also, in future, modifying the ensemble learner, we might be able to identify motif instances which don't belong to any of the

three classes and hence can be assigned to a fourth class. So, the ensemble classifier has more flexibility and future work scope.

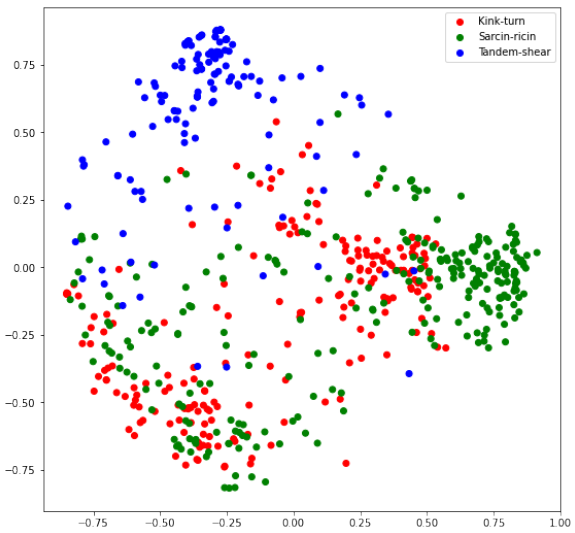

Fig. 7. Distribution of input dataset of unsupervised learning.

2.6 Implementing Unsupervised Learner

For unsupervised learning we have used several clustering algorithms to cluster the motifs. At first, we have used the Density-based spatial clustering of applications with noise (DBSCAN) algorithm. It was supposed to divide the training motif instances into clusters where each cluster represents a motif family. We have also applied K-means and Agglomerative clustering model, BIRCH (balanced iterative reducing and clustering using hierarchies), and Gaussian-Mixture model to compare the performance. Besides, we have tried T-distributed stochastic neighbor embedding, or T-SNE to visualize the original dataset. Figure 7 shows the distribution of our input dataset and Figure 8 depicts the visualization using T-SNE.

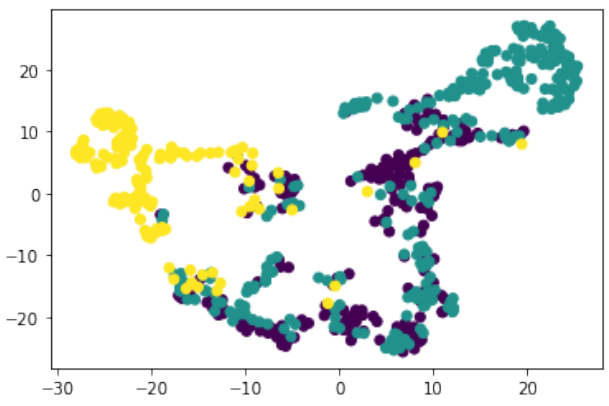

Fig. 8. Visualization of our input dataset using T-SNE.

3 Result Analysis

3.1 Supervised Model Evaluation

After balancing the dataset and only considering class ‘Kink-turn’, ‘Sarcin-ricin’, ‘Tandem-shear’, the maximum accuracy was gained for all the three models (Random forest, XGBoost, ELBC). Further improvement has been made on the supervised learners by excluding the least important sets of features while training the models. As previously mentioned in the Feature Analysis Section, using feature importance score, a list of least important features were generated. As we can see from Table 2, the maximum accuracy could be obtained after excluding the least important seven features ‘seq_aln_score_1’, ‘alignment_length_3’, ‘matching_stks_3’, ‘seq_aln_score_3’, ‘matching_bps_1’, ‘matching_bps_3’, ‘matching_stks_2’. In this table, the column ’Feature Set’ holds the number of features as they appear originally in the dataset.

Table 2. Accuracy after excluding different sets of least important features

Excluded Feature No	Feature Set	Accuracy RF	XG	ELBC
0	-	92.70%	92.00%	87.12%
6	16, 24, 26, 28, 30, 31	93.37%	93.14%	87.56%
7	12, 16, 24, 26, 28, 30, 31	**94.29**%	**93.84**%	**92.42**%
10	12, 14, 16, 19, 21, 24, 26, 28, 30, 31	92.77%	92.10%	88.63%

*RF: Random Forest
†XG: Extreme Gradient Boosting Model
‡ELBC: Ensemble Learner Binary Classifier

Figure 9 shows the confusion matrix for the Ensemble Learner Binary Classifier with data balancing and excluded less important features. As we can see from the figure, not only the overall accuracy, but also the probability of predicting the three class Kink-turn (class 1), Sarcin-ricin (class 2) and Tandem-shear (class 3) correctly are pretty high. In fact, the class Kink-turn can be identified with 100% accuracy. However, the model sometimes confuses between the classes Sarcin-ricin and Tandem-shear but the frequency of misclassification is pretty low.

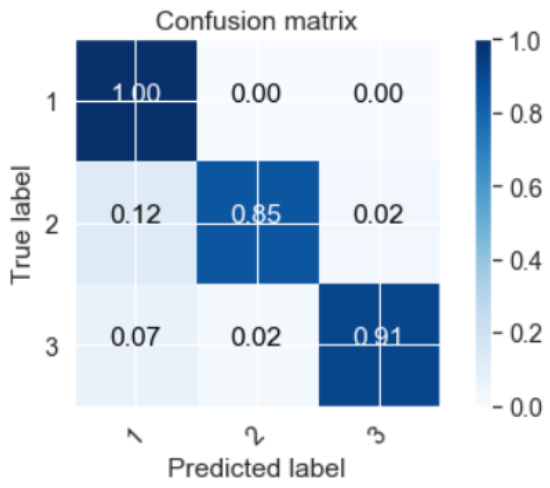

Fig. 9. Confusion matrix for the ELBC model after data balancing and excluding six less important features.

3.2 Unsupervised Model Evaluation

From Figure 7, it is clear that, based on our extracted features, the input dataset is separable in some areas but there are some overlaps. The boundary of the motif features is not clear enough to separate them into different clusters. As a result, DBSCAN, being a well established clustering model, could not separate the motif properly and marked most

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Table 3. Performance comparison of unsupervised learning models.

Clustering	RI*	NMI†	FMI‡
DBSCAN	0.397	0.092	0.591
K-means	0.666	0.311	0.523
Agglomerative	0.648	0.251	0.490
BIRCH	0.661	0.313	0.525
GMM	0.665	0.320	0.529

*RI: Random Index
†NMI: Normalized Mutual Information
‡FMI: Fowlkes-Mallows Index

of the motifs as noise in the outcome. The same effect can be seen in T-SNE visualization (Figure 8).

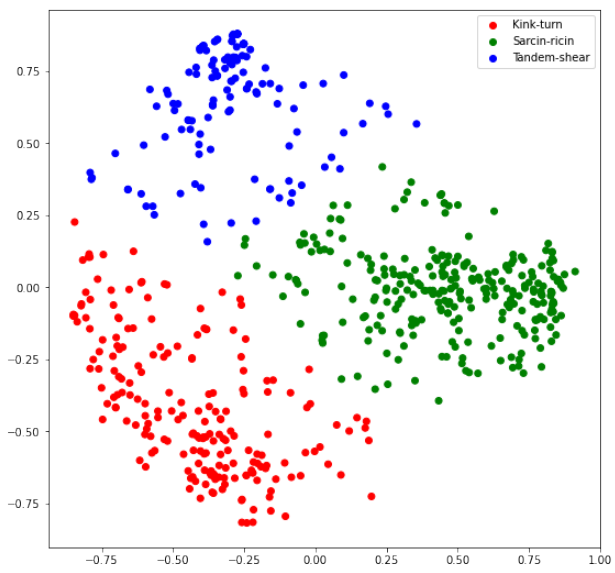

Fig. 10. Clustering our input dataset using Gaussian-Mixture model

To evaluate the performance of our unsupervised models, we have used three clustering evaluation metrics: Random Index (RI), Normalized Mutual Information (NMI), and Fowlkes-Mallows Index (FMI). These metrics are used when the ground truth knowledge is available. As we have the original labels for our input dataset, we utilized these metric to compare the performance of the unsupervised techniques. Table 3 shows the values of all these scores for different models. From this table, we can notice that the performance of DBSCAN is too bad for our dataset while the performance of other models are moderate. The best performed model for unsupervised learning seems to be Gaussian-Mixture model and Figure 10 shows the result of GMM.

After evaluating both supervised and unsupervised learners, it was pretty evident that the supervised learners perform well compared to the unsupervised learners.

4 Discussion

Currently, the implemented model can determine if a given motif instance belongs to Kink-turn, Sarcin-ricin or Tandem-shear family. However, when we tried to train model using motif instances from other known and unknown motif families, the output accuracy was less than 90% which might not be that reliable. The reason behind using the classes 'Known-motif' and 'Unknown-motif' to train our model was to identify motif instances not belonging to Kink-turn, Sarcin-ricin or Tandem-shear family as either 'Known-motif' or 'Unknown motif'. In that way, for any given motif instance, we will easily be able to say if that belongs to any of the three considered families or it belongs to some other known motif family or it is a motif instance which hasn't been assigned to any of the known motif families yet. But the current model couldn't accomplish that goal with high accuray. The reason behind this is, there is no significant pattern among the motif instances that were assigned to 'Known-motif' class as they belong from separate motif families. Similarly, the motif instances assigned to the class 'Unknown-motif' also came from very different sources, so it's hard to find a pattern among them as well. But, as we can see, there was a pattern among the motif instances of Kink-turn motif family, Sarcin-ricin-motif family and Tandem-shear motif family. Hence, the supervised learners were successfully able to learn from that pattern and identify the motif instances. However, instead of having some similar pattern among those families, many instances of the families behave slightly differently and generate a large number of outliers. This in turn puzzles the unsupervised learner models and hampers their performance, resulting in a low Random Score Value.

Table 4. Work Distribution

Timeline	Completed Tasks	Done By
Week 1-2	Motif Location Collection	Mahfuz
	Motif Sequence, 3D Structure Collection	Nabila
Week 3-4	Feature Extraction, Sequence Alignment	Nabila
	3D structural Alignment, RMSD calculation	Mahfuz
Week 5-6	Testing Data Coolection, Data Preprocessing	Mahfuz
	Implementing Kmeans, DBSCAN Algorithm	Nabila
Week 7-8	Implementing Unsupervised Models	Mahfuz
	Implementing Supervised Models	Nabila
Week 9-10	Testing and Model Evaluation	Both
Week 11	Final Report Preparation	Both

5 Conclusion

Using this proposed model, it might be possible to identify the existence of members of well-known motif families in different RNA chains. Besides sequential information, the utilization of 3D structural information has made the model more sophisticated and the accuracy of the supervised model has improved significantly. However, initially only three motif families have been considered for training the model. But by making the model known to all the existing RNA motif families, it might be converted into a tool which can successfully identify a motif instance belonging to any of the known families. Another limitation of the current model is that, it can not always accurately tell if the given instance belong to any of the known motif families at all. The reason behind this might be the lack of motif instance data and the lack of tuning among such data. One solution to this problem could be providing non-motif RNA chain instances. Another approach can be modifying our Ensemble Learner Binary Classifier (ELBC). Currently it can successfully identify if a motif is Kin-turn or not, Sarcin-ricin or not and Tandem-shear not. But while merging the results, it assigns that instance to one of the three classes with the highest probablity. But by setting a threshold value, a fourth class can be introduced in the model which will indicate that the given motif instance doesn't belong to any of the three classes. As we can see, our

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

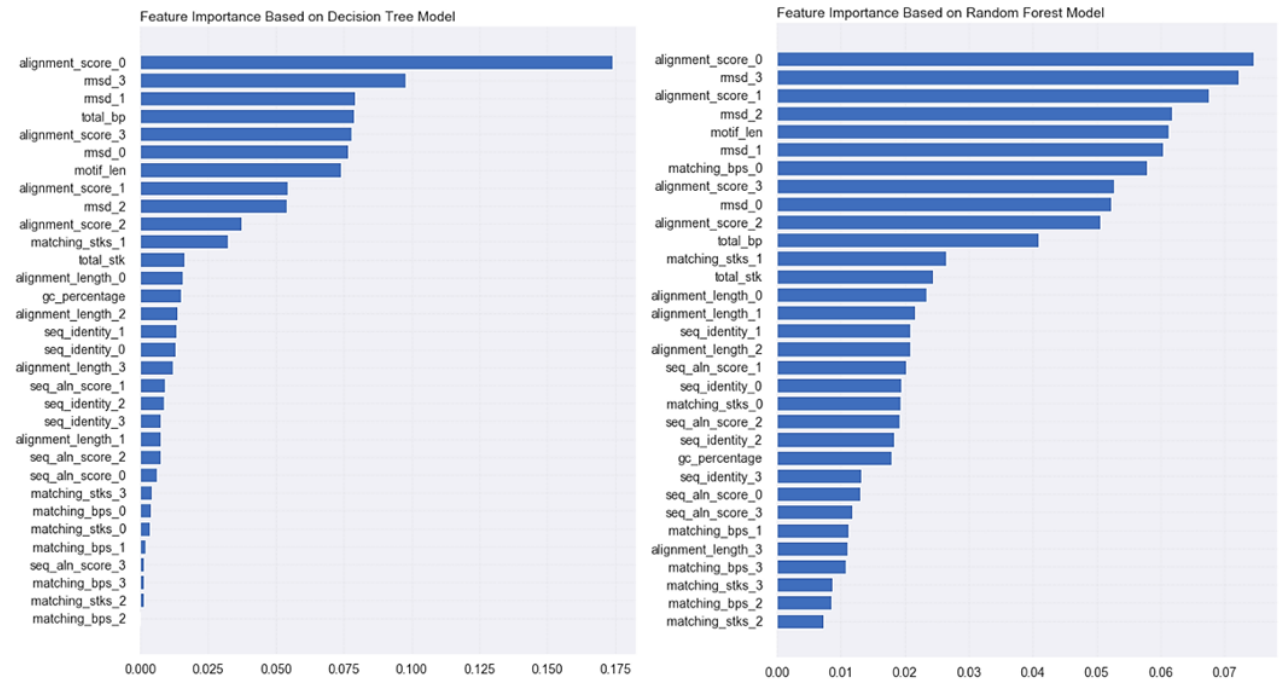

Fig. 11. Feature importance analysis based on decision tree (left) and random forest (right) model.

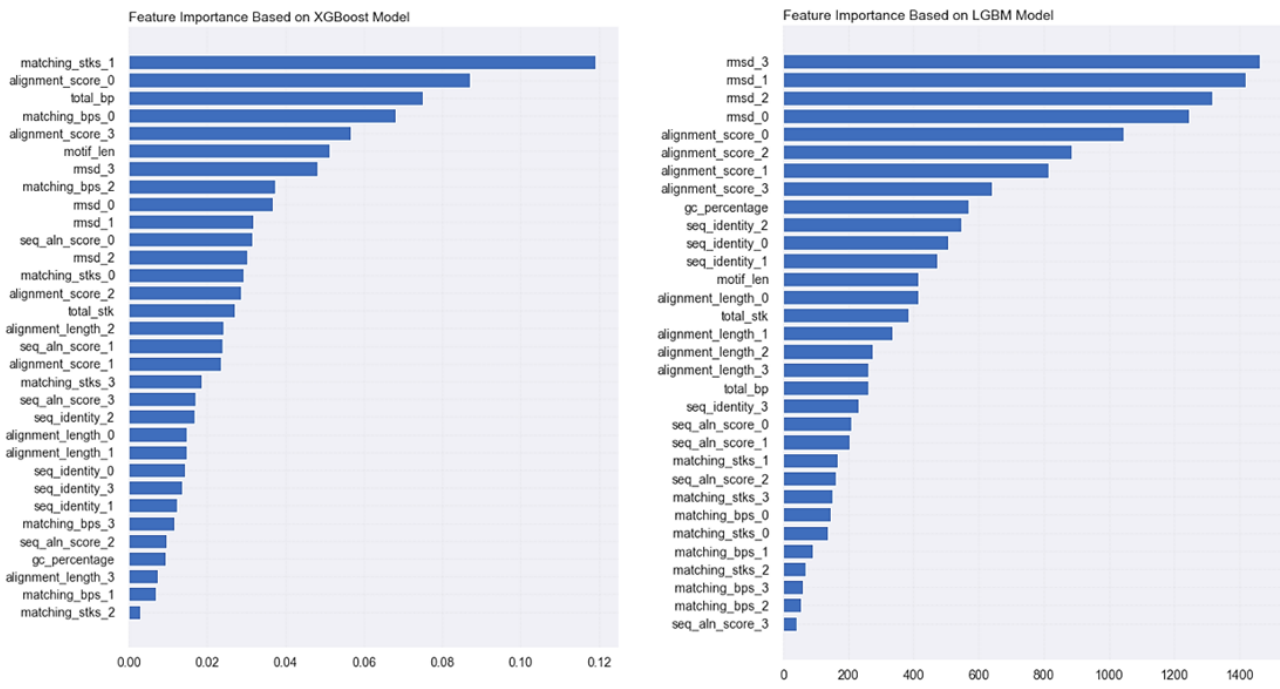

Fig. 12. Feature importance analysis based on XGBoost (left) and LGBM (right) model.

proposed Ensemble Learner Binary classifier has much more flexibility compared to other pre-existing supervised algorithms. Hence, inspite of its accuracy being slightly lower compared to the models Random Forest and XGBoost, we are still focusing on this model for future improvements. In conclusion it can be said that, if this tool can be built successfully in the future then using this tool, it will be possible to identify the existence of all the known motif family instances in the whole RNA chain database with much less effort and time. Hence, this can be a milestone achievement in the field of motif discovery and analysis.

6 Work Distribution

The work distribution between the two team members have been shown elaborately in Table 4

References

[1]Jürgen Brosius and Carsten A Raabe. What is an rna? a top layer for rna classification. *RNA biology*, 13(2):140–144, 2016.

[2]Neocles B Leontis, Aurelie Lescoute, and Eric Westhof. The building blocks and motifs of rna architecture. *Current opinion in structural biology*, 16(3):279–287, 2006.

[3]Aurelie Lescoute, Neocles B Leontis, Christian Massire, and Eric Westhof. Recurrent structural rna motifs, isostericity matrices and sequence alignments. *Nucleic acids research*, 33(8):2395–2409, 2005.

[4]Ping Ge, Shahidul Islam, Cuncong Zhong, and Shaojie Zhang. De novo discovery of structural motifs in rna 3d structures through clustering. *Nucleic acids research*, 46(9):4783–4793, 2018.

[5]Paul P Gardner and Hisham Eldai. Annotating rna motifs in sequences and alignments. *Nucleic acids research*, 43(2):691–698, 2015.

[6]Timothy L Bailey, James Johnson, Charles E Grant, and William S Noble. The meme suite. *Nucleic acids research*, 43(W1):W39–W49, 2015.

[7]Wenxiu Ma, William S Noble, and Timothy L Bailey. Motif-based analysis of large nucleotide data sets using meme-chip. *Nature protocols*, 9(6):1428–1450, 2014.

[8]Timothy L Bailey. Dreme: motif discovery in transcription factor chip-seq data. *Bioinformatics*, 27(12):1653–1659, 2011.

[9]Ying Wang, Craig L Zirbel, Neocles B Leontis, and Biao Ding. Rna 3-dimensional structural motifs as a critical constraint of viroid rna evolution. *PLoS pathogens*, 14(2):e1006801, 2018.

[10]Nucleic Acid Structure Research Group.

[11]Scott A Strobel, Peter L Adams, Mary R Stahley, and Jimin Wang. Rna kink turns to the left and to the right. *Rna*, 10(12):1852–1854, 2004.

[12]Alexander A Szewczak, Peter B Moore, YL Chang, and Ira G Wool. The conformation of the sarcin/ricin loop from 28s ribosomal rna. *Proceedings of the National Academy of Sciences*, 90(20):9581–9585, 1993.

[13]Cuncong Zhong, Haixu Tang, and Shaojie Zhang. Rnamotifscan: automatic identification of rna structural motifs using secondary structural alignment. *Nucleic acids research*, 38(18):e176–e176, 2010.

[14]Cuncong Zhong and Shaojie Zhang. Rnamotifscanx: a graph alignment approach for rna structural motif identification. *RNA*, 21(3):333–346, 2015.

[15]Anton I Petrov, Craig L Zirbel, and Neocles B Leontis. Automated classification of rna 3d motifs and the rna 3d motif atlas. *Rna*, 19(10):1327–1340, 2013.

[16]Shahidul Islam, Md Mahfuzur Rahaman, and Shaojie Zhang. Rnamotifcontrast: a method to discover and visualize rna structural motif subfamilies. *Nucleic Acids Research*, 2021.

[17]Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[18]Xiang-Jun Lu, Harmen J Bussemaker, and Wilma K Olson. Dssr: an integrated software tool for dissecting the spatial structure of rna. *Nucleic acids research*, 43(21):e142–e142, 2015.

[19]Michael Sarver, Craig L Zirbel, Jesse Stombaugh, Ali Mokdad, and Neocles B Leontis. Fr3d: finding local and composite recurrent structural motifs in rna 3d structures. *Journal of mathematical biology*, 56(1):215–252, 2008.