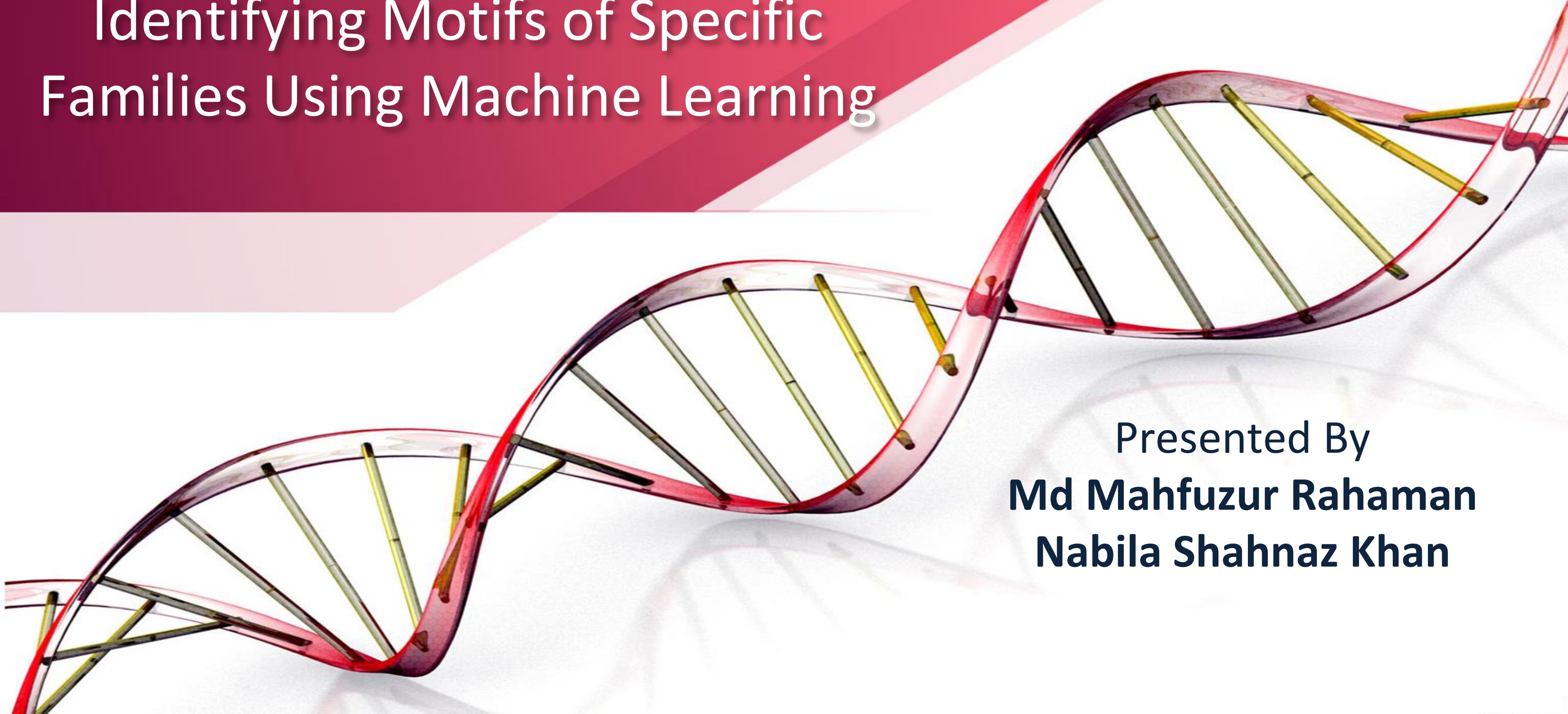
Identifying Motifs of Specific Families Using Machine Learning

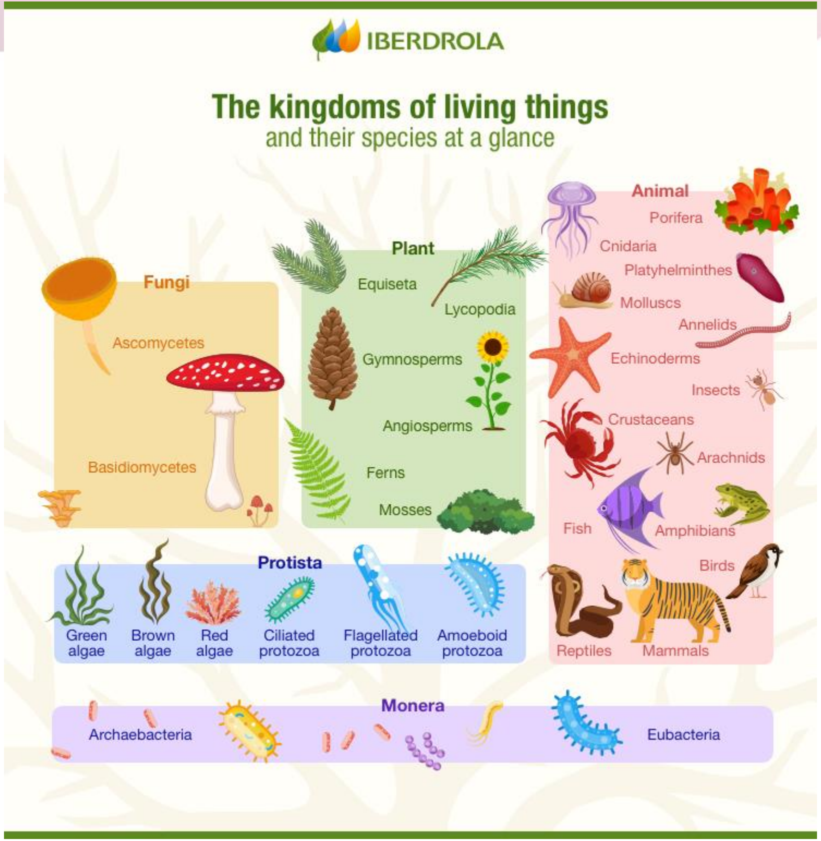


Presented By

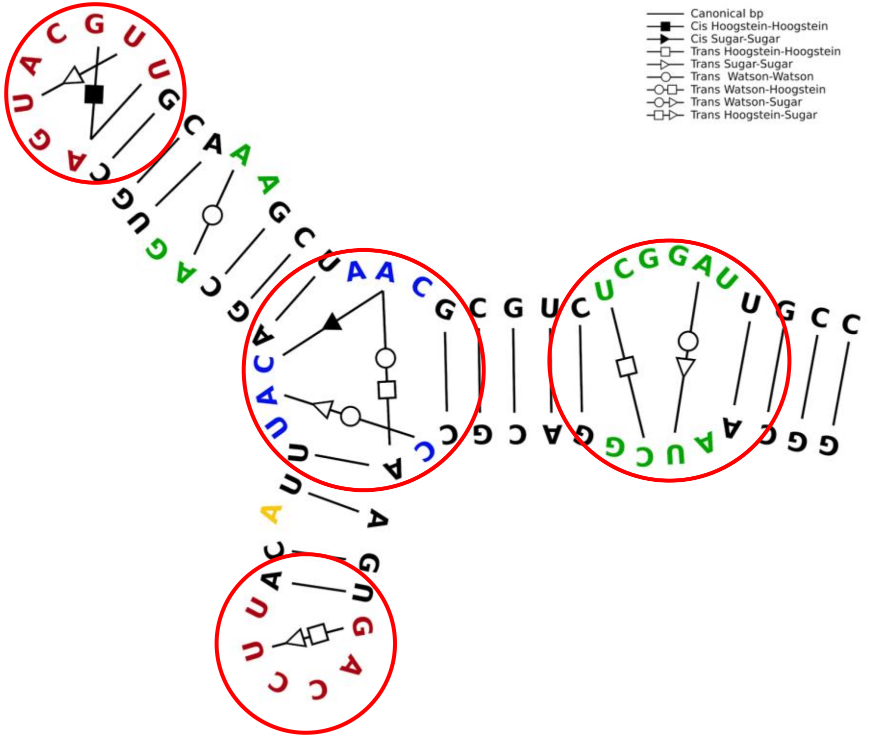
Md Mahfuzur Rahaman

Nabila Shahnaz Khan

RNA, Motif, and Motif Families

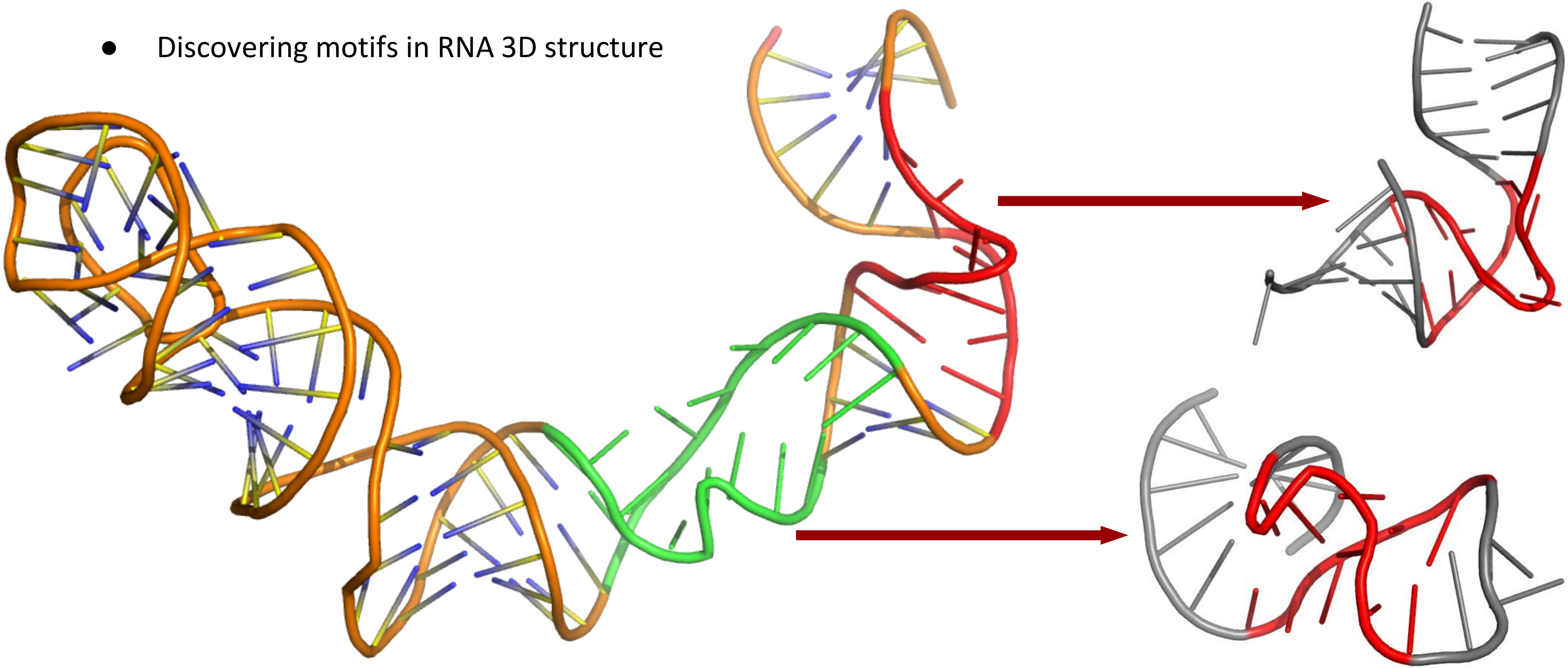


https://www.iberdrola.com/sustainability/biology-kingdoms-living-things-classification



Problem Statement

- Discovering motifs in RNA 3D structure



Motivation and Goal

Motivation

- Earn valuable insights on the instances of well known motif families
- Explore the feasibility of solving motif finding problem using ML

Goal

- Identify the instances of well known motif families in different RNA chains

Methodology

- Develop a model that can identify the instances of a specific family using ML techniques
- Considered families:
  - Kink-turn
  - reverse Kink-turn
  - Sarcin-ricin
  - Tandem-shear

- Steps followed:
  - ✓ Data Collection
  - ✓ Feature Extraction
  - ✓ Data Preprocessing and Normalization
  - ○ Model Training
  - ○ Testing and Model Evaluation

Data Collection

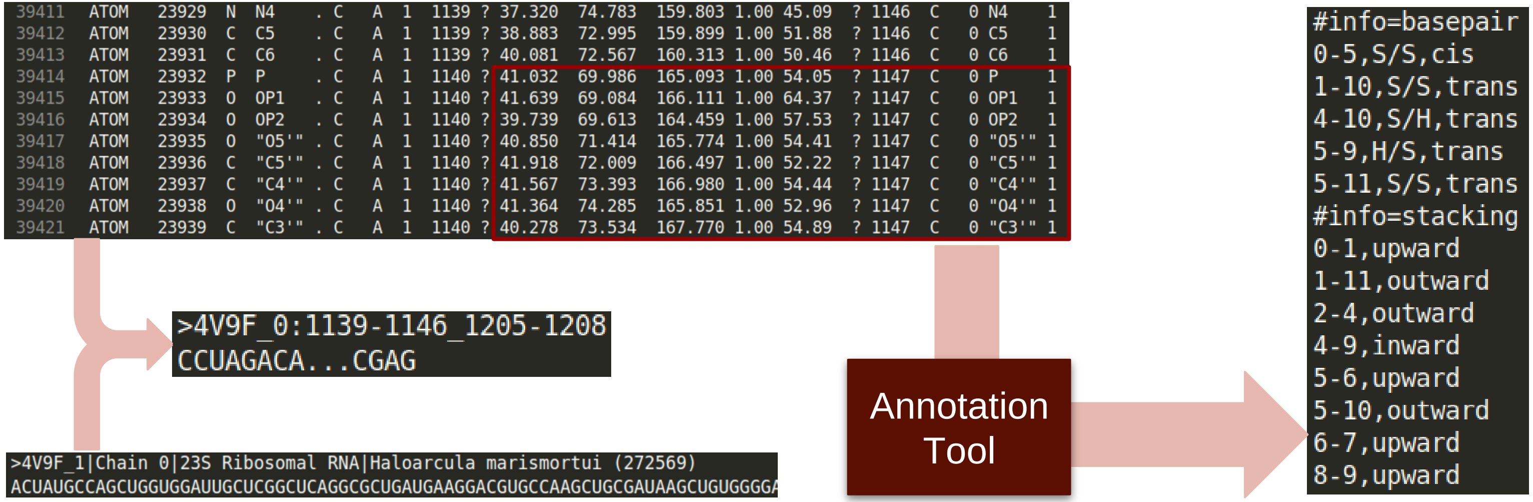
- Collecting motif locations

- Lilley et al. [4]
- Strobel et al. [5]
- Szewczak et al. [6]
- Zhong et al. [2]
- Ge et al. [8]
- Zhong et al. [7]
- Petrov et al. [9]
- Islam et al. [10]

$$\underbrace{\text{4V9F}}_{\text{PDB\_ID}}\_\underbrace{\text{0}}_{\text{Chain\_ID}}:\underbrace{\text{1147-1154\_1213-1216}}_{\text{Location}}$$

Data Collection

- Collecting motif sequence and 3D structure data



```
ATOM   23929  N  N4     . C   A  1   1139 ? 37.320   74.783   159.803 1.00 45.09   ? 1146   C    0 N4     1
ATOM   23930  C  C5     . C   A  1   1139 ? 38.883   72.995   159.899 1.00 51.88   ? 1146   C    0 C5     1
ATOM   23931  C  C6     . C   A  1   1139 ? 40.081   72.567   160.313 1.00 50.46   ? 1146   C    0 C6     1
ATOM   23932  P  P      . C   A  1   1140 ? 41.032   69.986   165.093 1.00 54.05   ? 1147   C    0 P      1
ATOM   23933  O  OP1    . C   A  1   1140 ? 41.639   69.084   166.111 1.00 64.37   ? 1147   C    0 OP1    1
ATOM   23934  O  OP2    . C   A  1   1140 ? 39.739   69.613   164.459 1.00 57.53   ? 1147   C    0 OP2    1
ATOM   23935  O  "O5'"  . C   A  1   1140 ? 40.850   71.414   165.774 1.00 54.41   ? 1147   C    0 "O5'"  1
ATOM   23936  C  "C5'"  . C   A  1   1140 ? 41.918   72.009   166.497 1.00 52.22   ? 1147   C    0 "C5'"  1
ATOM   23937  C  "C4'"  . C   A  1   1140 ? 41.567   73.393   166.980 1.00 54.44   ? 1147   C    0 "C4'"  1
ATOM   23938  O  "O4'"  . C   A  1   1140 ? 41.364   74.285   165.851 1.00 52.96   ? 1147   C    0 "O4'"  1
ATOM   23939  C  "C3'"  . C   A  1   1140 ? 40.278   73.534   167.770 1.00 54.89   ? 1147   C    0 "C3'"  1
```

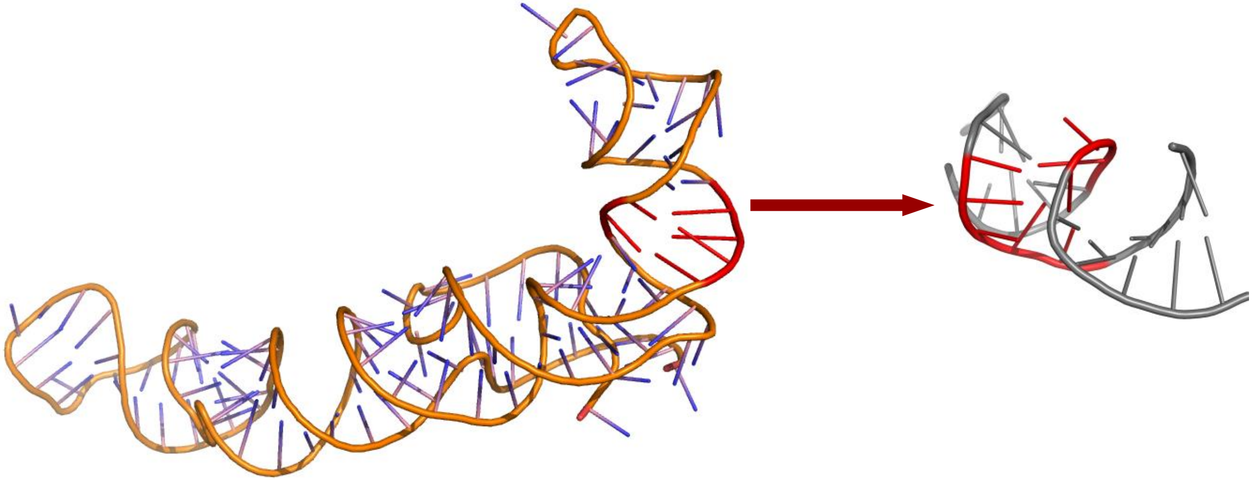
```
>4V9F_0:1139-1146_1205-1208
CCUAGACA...CGAG
```

```
>4V9F_1|Chain 0|23S Ribosomal RNA|Haloarcula marismortui (272569)
ACUAUGCCAGCUGGUGGAUUGCUCGGCUCAGGCGCUGAUGAAGGACGUGCCAAGCUGCGAUAAGCUGUGGGGA
```

```
#info=basepair
0-5,S/S,cis
1-10,S/S,trans
4-10,S/H,trans
5-9,H/S,trans
5-11,S/S,trans
#info=stacking
0-1,upward
1-11,outward
2-4,outward
4-9,inward
5-6,upward
5-10,outward
6-7,upward
8-9,upward
```

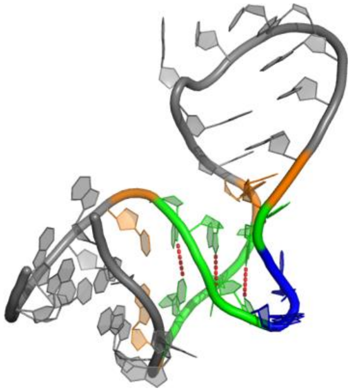
Data Collection

- Collecting testing data set
  - Cut loops from RNA chains
  - Follow the same procedure to get the structure information

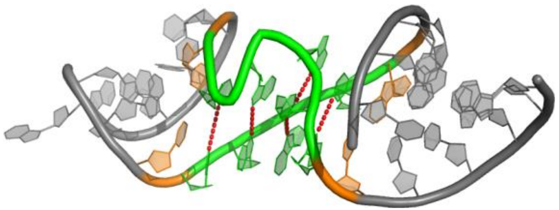


Representative Selection

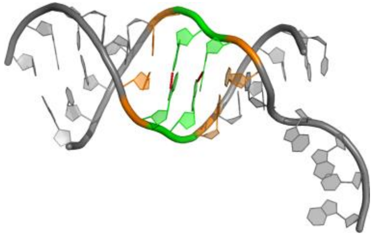
- Selecting Representatives using the tool RNAMotifContrast [10]



Kink-turn



Sarcin-ricin



Tandem-shear

Feature Extraction:

- **Motif Sequence and 3D Structure Collection:**

  Sequence Features:

□ Motif length, GC percentage

□ Sequence Alignment Score & Sequence Identity (Needleman–Wunsch algorithm [1])

Motif_str	Motif_length	GC_percentage	Seq_identity			Align_score			
			Kinkturn	Sarcin Ricin	Tandem Shear	Kinkturn	Sarcin Ricin	Tandem Shear	Family
1E7K_C:2-8_16-19	11	0.545	0.643	0.71	0.71	9	10	7	Kink-turn
1NBS_B:52-57_79-85	13	0.154	0,625	0.5	0.94	10	8	7	Sarcin-Ricin
1U6B_B:52-55_81-84	8	0.375	0.909	0.818	0.714	10	9	7	Tandem Shear
...	...	...	...	...	...	...	...	...	...

Feature Extraction:

- **Motif Sequence and 3D Structure Collection:**

3D Structural Features

□ 3D structure Alignment Score, Alignment Length

□ RMSD Value (using tool RNAMotifScanX [2])

□ Total base-pairs and base-stacks (using tools DSSR [3])

□ Matching base-pairs and base stacks with representative (using tool RNAMotifScanX [2])

Motif_str	Tot _BP	Tot _BS	3D_Alignment_Score			3D_Alignment_RMSD			Matching_basepair			Family
			Kinkturn	Sarcin Ricin	Tandem Shear	Kinkturn	Sarcin Ricin	Tandem Shear	Kinkturn	Sarcin Ricin	Tandem Shear	
1E7K_C:2-8_16-19	2	7	46.8	107	35	7.95	6.9	7.796	2	2	2	Kink-turn
1NBS_B:52-57_79-85	5	13	172.8	35.9	91.7	0.87	5.4	5.93	5	3	2	Sarcin -Ricin
1U6B_B:52-55_81-84	2	6	46.4	80.4	35.3	6.28	7.35	0.96	2	2	2	Tandem Shear
…	…	…	…	…	…	…	…	…	…	…	…	…

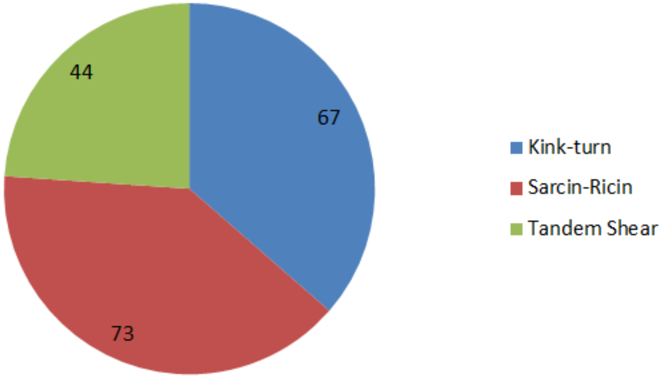
Data Preprocessing:

- **Motif Sequence and 3D Structure Collection:**

□ Total 33 features, all numerical

□ No missing values

□ Less-noisy dataset



RNA motif Family Data Ratio

□ Standard scaling and Normalization to bring all the attributes to a comparable level

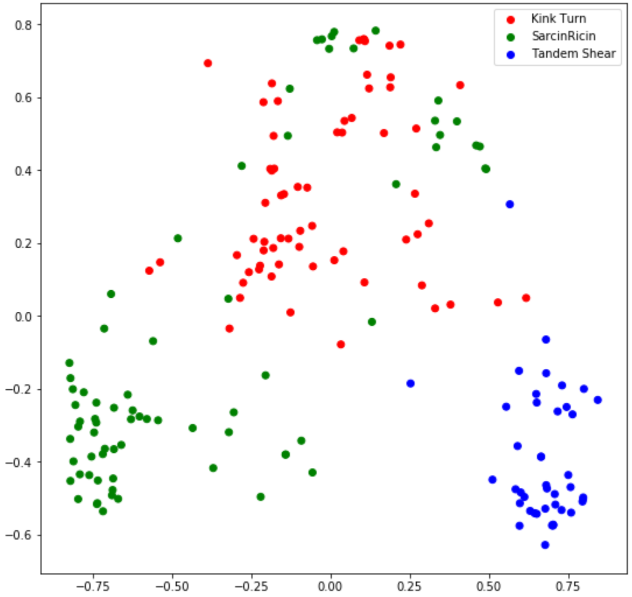
□ PCA(Principal component Analysis) for cluster implementation and visualization

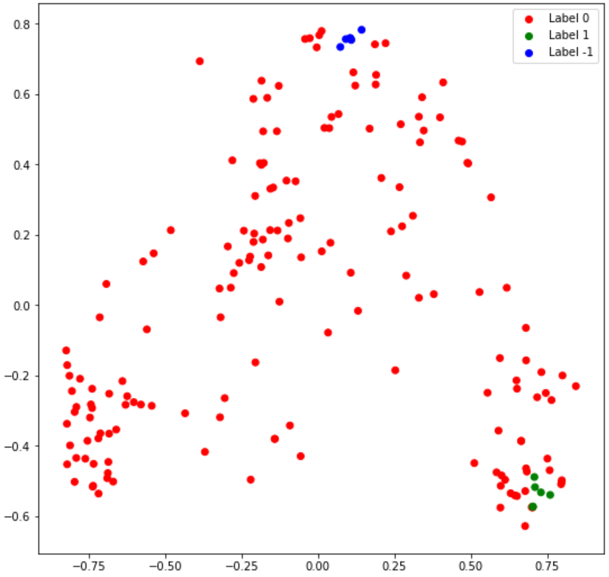
Model Training:

- **Unsupervised Learning (DBSCAN clustering)**

  Converted 31 features to two features (P1 and P2) using PCA

  Parameters used: Min distance = 0.05, Min Neighbors = 7



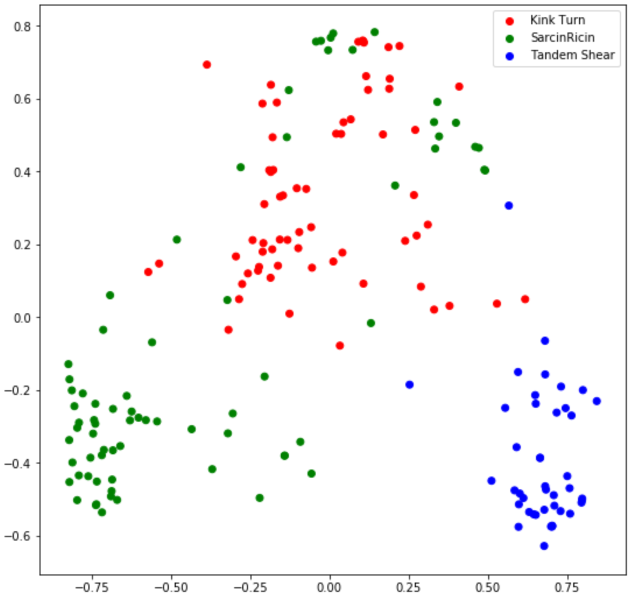


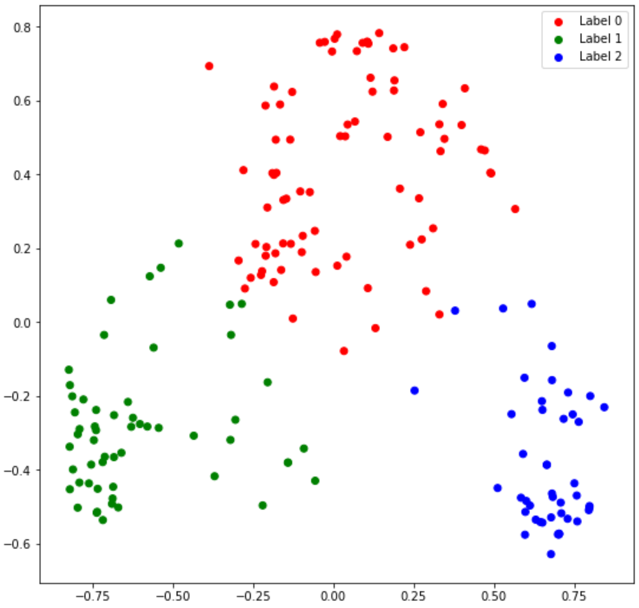
Model Training:

- **Unsupervised Learning (K-means clustering)**

  Converted 31 features to two features (P1 and P2) using PCA

  Parameters used: Initial Clusters = 3

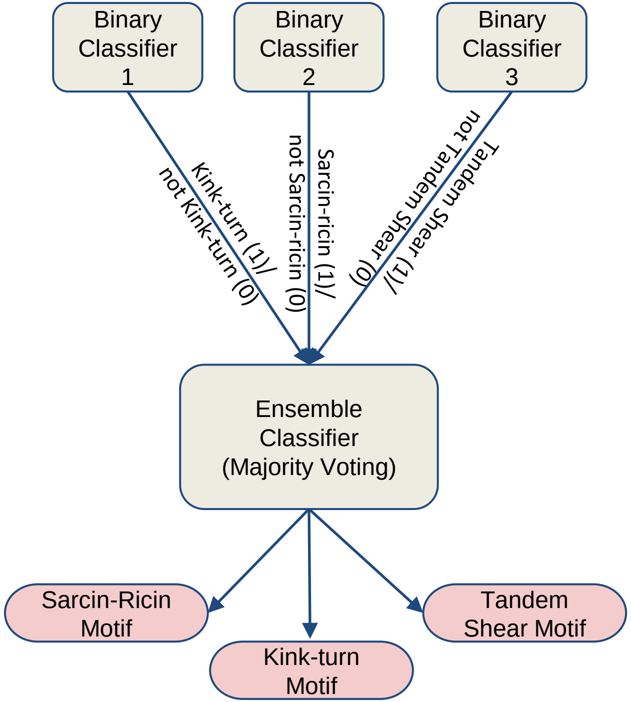




Model Training:

- **Supervised Learning (Binary Classifier):**
  - → Modeled 3 naïve bayes binary classifiers for 3 motif families
  - → Built an ensemble classifier
  - → Used maximum voting
  - → Model Evaluation

	Accuracy	Precision	F1-Score	Recall-sore
Binary Classifier 1	86.96	84.33	88.62	85.63
Binary Classifier 2	93.48	93.6	93.37	93.45
Binary Classifier 3	97.8	98.6	95.0	96.7
Ensemble Classifier	84.78	87.23	86.71	86.3


Binary Classifier 1
Binary Classifier 2
Binary Classifier 3
Kink-turn (1)/ not Kink-turn (0)
Sarcin-ricin (1)/ not Sarcin-ricin (0)
Tandem Shear (1)/ not Tandem Shear (0)
Ensemble Classifier (Majority Voting)
Sarcin-Ricin Motif
Kink-turn Motif
Tandem Shear Motif

Current Challenges:

- ➔ Tertiary structural information of motifs has not been taken into consideration before
- ➔ No pre-existing dataset, had to extract features and generate data
- ➔ Dataset clean but Really small in size
- ➔ Can’t guarantee model accuracy
- ➔ Motif sequences not same length, so couldn’t generate Position Weight Matrix as a feature yet

Future Work:

- Performing Multi-alignment in order to generate position weight matrix
- Feature importance generation
- Implementing other supervised and unsupervised learners to compare performance
- Performing Evaluation using both motif family members and non-member instances

Proposed Work Schedule

Timeline	Task to be Completed	Responsible Individual
Week 1-2	a) Motif Location Collection b) Motif Sequence and 3D Structure Collection	Mahfuz Nabila
Week 3-4	a) Feature Extraction and Sequence Alignment b) 3D structural Alignment and RMSD calculation	Nabila Mahfuz
Week 5-6	a) Data Preprocessing and Normalization b) Implementing DBSCAN Algorithm	Mahfuz Nabila
Week 7-8	a) Model Training b) Collection of Testing Dataset	Mahfuz and Nabila
Week 9-10	Testing and Model Evaluation	Mahfuz and Nabila
Week 11 (April 29th)	Final Report Preparation	Mahfuz and Nabila

References

[1] Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology, 48(3), 443-453.
[2] Zhong, Cuncong, and Shaojie Zhang. "RNAMotifScanX: a graph alignment approach for RNA structural motif identification." RNA 21.3 (2015): 333-346.
[3] Lu, X. J., Bussemaker, H. J., & Olson, W. K. (2015). DSSR: an integrated software tool for dissecting the spatial structure of RNA. Nucleic acids research, 43(21), e142-e142.
[4] Nucleic Acid Structure Research Group, University of Dundee, University of Dundee, Professor David M.J. Lilley FRS. http://www.lifesci.dundee.ac.uk/groups/nasg/
[5] Strobel, S. A., Adams, P. L., Stahley, M. R., & Wang, J. (2004). RNA kink turns to the left and to the right. RNA (New York, N.Y.), 10(12), 1852–1854. https://doi.org/10.1261/rna.7141504
[6] Szewczak, A. A., Moore, P. B., Chang, Y. L., & Wool, I. G. (1993). The conformation of the sarcin/ricin loop from 28S ribosomal RNA. Proceedings of the National Academy of Sciences of the United States of America, 90(20), 9581–9585. https://doi.org/10.1073/pnas.90.20.9581
[7] Zhong, C., Tang, H., & Zhang, S. (2010). RNAMotifScan: automatic identification of RNA structural motifs using secondary structural alignment. Nucleic acids research, 38(18), e176. https://doi.org/10.1093/nar/gkq672
[8] Ge P., Islam S., Zhong C., Zhang S. De novo discovery of structural motifs in RNA 3D structures through clustering. Nucleic Acids Res. 2018; 46:4783–4793.
[9] Petrov A.I., Zirbel C.L., Leontis N.B. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. RNA. 2013; 19:1327–1340.
[10] Shahidul Islam, Md Mahfuzur Rahaman, Shaojie Zhang, RNAMotifContrast: a method to discover and visualize RNA structural motif subfamilies, Nucleic Acids Research, 2021;, gkab131, https://doi.org/10.1093/nar/gkab131

Any Questions?