Final Project Proposal

CAP 6545: Spring, 2021

Machine Learning for Biomedical Data

Group Members:

1. Md. Mahfuzur Rahaman (4808016)
2. Nabila Shahnaz Khan (5067496)

Project Title: Identifying Motifs of Specific Families Using Unsupervised Machine Learning

Project Idea:

❖ Problem Definition:

Motifs are the recurring patterns in the genome and are also known to be the binding sites of transcription factors. However, the definition of motif might vary depending on the research context. 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, MEME-Chip, DREME use approaches based on comparative genomics, motif profile search, and statistical analysis. But they lack in both efficiency and accuracy. Also, tertiary structural information of motifs is not taken into consideration in most of the cases. But 3D structural information plays a vital role in RNA-protein binding, hence can be an important factor in motif identification.

❖ Proposed Method:

For our project, we will try to develop a model that will be able to identify motifs of a specific motif family using the machine learning technique. For now, we are only focusing on some well-known motif families such as Kink-turn, Reverse Kink-turn, Sarcin-ricin, etc. In order to develop the model, we will need to go through the following steps:

- **Data Collection:**
  1. **Motif Location Collection:** To get the locations of the motif instances of the Kink-turn family, we could follow the research work of Professor David M.J. Lilley [1]. They have provided several locations of Kink-turn motif instances in different RNAs. We will also be able to find such locations for reverse Kink-turn and Sarcin-ricin from [2] and [3] respectively. The motif instances described in these papers might not be 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. [4] 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 can use as our data source. Again, Petrov et. al. [5] have also worked on these motifs and provided a bunch of locations for such motif instances.
  2. **Motif Sequence and 3D Structure Collection:** We will collect the specific PDBx file from the PDB database [6] and extract the sequential and structural information

(e.g. co-ordinates, base-pairings, base-stackings, etc.) of the motifs from specific RNA chains.

3. **Collection of Testing Dataset:** To test our machine learning model with unseen data, we need to 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.

- **Feature Extraction:** Some possible features for training the model are Position Weight Matrix (PWM), Motif length, GC percentage, Alignment score, RMSD value, etc. In order to generate the identity score, we plan to use Needleman–Wunsch sequence alignment algorithm. For calculating the RMSD values between motifs, first, we plan to use the tool RNAMotifScanX to align 3D structures. Then using the Kabsch algorithm, we will generate the RMSD value for the aligned nucleotides of the motifs.
- **Data Preprocessing and Normalization:** As the range of values will differ for all the extracted continuous features, we will need to perform normalization and scaling on the dataset.
- **Model Training:** For training the model we plan to use the Density-based spatial clustering of applications with noise (DBSCAN) algorithm. Here we will divide the training motif instances into clusters where each cluster will represent a motif family.
- **Testing and Model Evaluation:** For testing, we will provide both motif and non-motif RNA loop instances to the model to evaluate if it can cluster them accordingly. Based on this result we will finally generate the accuracy and confusion matrix of the model.

❖ Significance of the Project:

We haven’t been able to find any such tool that can predict if a given RNA chain loop belongs to a certain renowned motif family. Using this proposed model, we will be able to identify the existence of members of well-known motif families in different RNA chains. Besides sequential information, the utilization of 3D structural information will make the model more sophisticated and hence it might improve accuracy significantly. If the tool can be built successfully then in the future, using this tool, it will be possible to identify the existence of all the known motif family instances in the whole RNA chain database. This can be a milestone achievement in the field of motif discovery and analysis.

Data Source:

We will collect the information about motif instances belonging to Kink-turn, Reverse Kink-turn and Sarcin-ricin families from the research works as mentioned above [1 - 3]. Further motif instances of these families can be collected from RNAMotifScan [4] and RNA 3D Motif Atlas [5] databases. Finally, we will collect the sequential and 3d structural information of these motif instances from the well-known PDB database [6] and extract features as per requirement.

Weekly Schedule:

Timeline	Task to be Completed	Responsible Individual
Week 1-2	a) Motif Location Collection b) Motif Sequence and 3D Structure Collection	a) Mahfuz b) Nabila
Week 3-4	a) Feature Extraction and Sequence Alignment b) 3D structural Alignment and RMSD calculation	a) Nabila b) Mahfuz
Week 5-6	a) Data Preprocessing and Normalization b) Implementing DBSCAN Algorithm	a) Mahfuz b) 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 (Till April 29th)	Final Report Preparation	Mahfuz and Nabila

Reference:

[1] 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/

[2] 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

[3] 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

[4] 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

[5] Petrov, A. I., Zirbel, C. L., & Leontis, N. B. (2013). Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *RNA (New York, N.Y.)*, *19*(10), 1327–1340. https://doi.org/10.1261/rna.039438.113

[6] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242. https://www.rcsb.org/