CS434 MINI COMPETITION: PSEUDOKNOT DETECTION IN RNA SECONDARY STRUCTURES

1 Task overview

RNA molecules usually come as single strands but they fold themselves into specific structures, which are critical toward their functionality. Computational approaches for predicting the RNA structure is an active area of research. In particular, many algorithms have been developed to predict well-nested RNA secondary structures (pseudoknot-free), versus overlapped pseudoknot structures (see Figure 1). For this competition, you will be building predictive models for "Pseudoknot Detection", the task of predicting the presence or absence of pseudoknot (without identifying the specific location) given an RNA sequence.

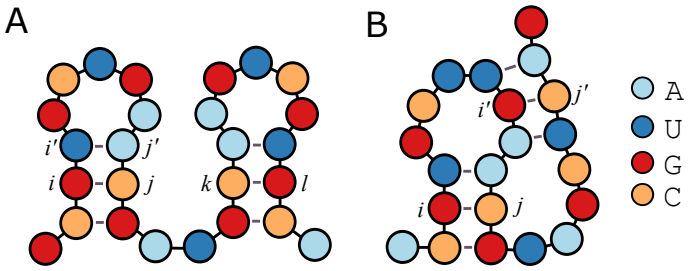

Figure 1: RNA secondary structures representation. **(A)** A secondary structure of an RNA without a pseudoknot that has base pair (i, j) with all other base pairs such that $i < i' < j' < j$ as with (i', j'), or $i < j < k < l$ as with (k, l). **(B)** A secondary structure with a pseudoknot and base pairs (i, j) and (i', j') such that $i < i' < j < j'$.

2 Data overview

Training data. For the competition you are provided training data in several formats. First, you are given the raw RNA sequences. The sequences are given in two different files, containing PK-free RNA sequences (pks_Train.fasta) and PK-present sequences (pk_Train.fasta) respectively. The files are in plain text format and each sequence is associated with an ID followed the sequence itself. Note that we have more PK-free sequences than PK-present sequences for training.

Second, you are given two text data files named 'featuresall_train.txt' and 'features103_train.txt'. Each line of these text files corresponds to one RNA sequence and contains the ID of the sequence and its target label (0 for PK-free and 1 for PK-present), and a set of numerical features representing the sequence, engineered and extracted from the sequence based on biological knowledge. The first data 'featuresall_train.txt' contains a total of 1053 features whereas the second data 'features103_train.txt' contains a much smaller subset of 103 features that are identified by the domain expert as particularly relevant to this task.

Testing data. For the competition, you will be required to make predictions for a set of test RNA sequences. Note that different from the training data, the test data is balanced, that is, it contains the same amount of PK-free and PK-present examples. This information is important as you tune your models.

The test data will come in the same format as the training data. You will be provided with three different files: features103_test.txt, featuresall_test.txt and sequence_test.fasta.

-

3 Required submission for competition

The goal of the competition is to build the most accurate model for detecting the presence/absence of Pseudoknot. We will have two mandatory competition tasks. For the first task, your model can only use the selected 103 features to make predictions. For the second task, your model can consider all 1053 features. As an additional optional task, you can also build model consider the sequences in addition to the provided features. For each task, each team will make up to 3 submissions, the best results among the 3 submissions will be used to rank the teams.

4 Evaluation Criterion

We will consider two different evaluation criteria: accuracy and the area under the ROC curve (AUC). Here is a detailed description of the concept of AUC: Understanding AUC - ROC curve