SECPred

1. Data collection:

For development of machine learning models to perform binary classification, that is, to predict whether any input RNA sequence can serve as a SECIS element or not, two datasets are need -- a positive dataset, containing known SECIS elements, and a negative dataset, containing known non-SECIS element RNA sequence. The datasets for the tool development were extracted from NCBI Nucleotide. While for the positive dataset (~300) NCBI Nucleotide was queried for SECIS elements, in case of the negative dataset (~1000), random RNA sequences (but not belonging to SECIS element) were downloaded.

1. Data cleaning:

The next step was to clean the datasets of any information that may adversely affect prediction accuracy. Sequences containing ambiguous bases were removed. The sequence lengths were studies as well, and sequences which were too long or too short relative to the average (~70 bases), were removed.

1. Data preprocessing:

Machine learning algorithms perform statistical operations and hence can process only numerical data. To train a model on sequence data would require certain transformation of that text information into numerical format. Count vectorisation was used to convert the sequence information into numerical matrices.

1. Data balancing:

For predictions to be accurate and unbiased, the amount of information in both datasets, that is, for both classes, should be equal. Therefore, to balance the skeewness in the size of datasets, SMOTE was used to oversample the positive dataset.

1. Multiple models were trained on the datasets, of which Bernoulli Naïve-Bayes classifier was found to be the best performing one.