University of Piraeus

Non Coding RNA Classifier

Author:
Konstantinos Vasilas
Student ID: ME2102
vasilas.cei@gmail.com

Supervisor:
Ilias Maglogianis

November 16, 2023

Contents

List of Figures

List of Tables

1 Introduction

Non-coding RNAs (ncRNAs) are a diverse class of RNA molecules that do not encode proteins. Despite their lack of coding potential, ncRNAs play important regulatory roles in various biological processes, including gene expression, DNA replication, and epigenetic regulation. The discovery of numerous unknown non-coding RNAs (ncRNAs) has presented significant challenges for researchers conducting functional studies on these molecules. Given the diverse functions associated with different families of ncRNAs, precise prediction of ncRNA families is essential for advancing research on their functions. Biological experimental methods for identifying ncRNA families are not only time-consuming and labor-intensive but also expensive, making them impractical for the demands of high-throughput technology. Consequently, the use of computational methods becomes inevitable for the efficient prediction of ncRNA families.

In this thesis, we present a machine learning approach for classifying ncRNAs using a variety of features derived from the primary sequence and secondary structure of the RNA molecules. Our approach utilizes a combination of supervised and unsupervised learning techniques to accurately predict the class of an ncRNA based on its sequence and structure. We evaluate the performance of our classifier on a diverse set of ncRNA datasets, and demonstrate its ability to accurately classify known ncRNA families as well as identify novel ncRNA classes. There are many different types of ncRNAs, including microRNAs (miRNAs), small nucleolar RNAs (snoRNAs), transfer RNAs (tRNAs), small interfering RNAs (siRNAs), piwi-interacting RNAs (piRNAs), and long non-coding RNAs (lncRNAs).

Overall, our machine learning approach provides a powerful tool for the identification and classification of ncRNAs, which will be valuable in understanding the functional roles of these molecules in various biological processes.

2 Theoretical Background

2.1 RNA

RNA (ribonucleic acid) is a type of nucleic acid that plays a central role in various biological processes, including protein synthesis, gene expression, and regulation of genetic information. It is a single-stranded molecule composed of nucleotides, each of which consists of a nitrogenous base (adenine, cytosine, guanine, or uracil), a sugar (ribose), and a phosphate group. RNA is synthesized from DNA through a process called transcription and is involved in the translation of the genetic code from DNA to protein. There are several different types of RNA, including messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), and small nuclear RNA (snRNA). Each type of RNA has a specific function in the cell and plays a crucial role in maintaining the overall health and function of the organism.

- Messenger RNA (mRNA) is synthesized from DNA and carries the genetic code from the nucleus to the ribosomes in the cytoplasm, where it is translated into protein.
- Transfer RNA (tRNA) is a small RNA molecule that brings amino acids to the ribosomes during protein synthesis. It does this by matching the appropriate amino acid to the mRNA codon (a sequence of three nucleotides that codes for a specific amino acid) and adding the amino acid to the growing polypeptide chain.
- Ribosomal RNA (rRNA) is a component of ribosomes, the cellular structures that carry out protein synthesis. rRNA plays a structural role in ribosomes and also helps to catalyze the chemical reactions involved in protein synthesis.
- Small nuclear RNA (snRNA) is a class of small RNA molecules that play a role in various cellular processes, including RNA splicing, in which introns (non-coding sequences) are removed from mRNA and the remaining exons (coding sequences) are joined together.

In addition to these types of RNA, there are also other small RNA molecules that play important regulatory roles in the cell, such as microRNA (miRNA) and small interfering RNA (siRNA). These molecules help to regulate gene expression by blocking the expression of specific genes or by degrading specific mRNA molecules.

Overall, RNA plays a vital role in the flow of genetic information and in the regulation of gene expression in cells. It is an essential molecule in the maintenance of life and is involved in many important biological processes.

2.2 Non-coding RNA

Non-coding RNA (ncRNA) refers to RNA molecules that do not encode proteins and do not have a role in protein synthesis. These RNA molecules can be divided into two main categories: long non-coding RNA (lncRNA) and small non-coding RNA (sncRNA).

Long non-coding RNA (lncRNA) are RNA molecules that are more than 200 nucleotides in length and do not encode proteins. They are involved in various cellular processes, including gene regulation, chromosome organization, and RNA splicing. Some examples of lncRNA include Xist, which is involved in the regulation of gene expression during development, and HOTAIR, which is involved in the regulation of gene expression in cancer cells.

Small non-coding RNA (sncRNA) are RNA molecules that are less than 200 nucleotides in length and do not encode proteins. They include microRNA (miRNA), small interfering RNA (siRNA), and piwi-interacting RNA (piRNA), all of which play important regulatory roles in the cell.

MiRNA are small RNA molecules that bind to specific mRNA molecules and inhibit their translation into protein, thereby regulating gene expression. SiRNA are small RNA molecules that are involved in the RNA interference (RNAi) pathway, a process that silences specific genes by degrading their mRNA. PiRNA are small RNA molecules that are involved in the silencing of transposable elements, which are sequences of DNA that can move within the genome.

Overall, non-coding RNA molecules play a variety of important roles in the regulation of gene expression and other cellular processes. They are an important area of research in molecular biology and have the potential to be used as therapeutic targets for various diseases.

Non-coding RNA (ncRNA) molecules, have been shown to play important roles in the development and progression of cancer. Dysregulation of ncRNA expression has been observed in various types of cancer, and ncRNA molecules have been shown to be involved in a number of cancer-related processes, including cell proliferation, survival, and migration.

For example, some lncRNA molecules have been found to be upregulated (expressed at higher levels) in cancer cells and to promote the growth and proliferation of these cells. One example is HOTAIR, a lncRNA that has been found to be upregulated in various types of cancer, including breast, ovarian, and lung cancer. HOTAIR has been shown to promote the migration and invasion of cancer cells and to inhibit the programmed cell death (apoptosis) of these cells.

On the other hand, some sncRNA molecules, such as microRNA (miRNA), have been found to be downregulated (expressed at lower levels) in cancer cells and to have tumor-suppressive effects. MiRNA molecules have been shown to regulate the expression of multiple target genes and to play a role in the regulation of various cellular processes, including cell proliferation and apoptosis. Dysregulation of miRNA expression has been observed in various types of cancer and has been linked to the development and progression of these diseases.

2.3 Small non-coding RNA

Various classes of small non-coding RNAs (sncRNAs) exhibit distinctions in nucleotide sequence length, folding, and function. Among the well-known ncRNAs are structural RNA types, such as ribosomal RNA (rRNA) and transfer RNA (tRNA), integral to translation events. MicroRNAs (miRNAs) constitute another intriguing class of ncRNAs, measuring 18–24 nucleotides, and function as regulatory RNA

molecules. Their impact can manifest as either tumor-suppressive or oncogenic, contingent upon their target and the molecular mechanisms they influence. MiRNAs interact directly with target genes, binding to complementary sequences, resulting in mRNA degradation or translational suppression, ultimately inhibiting protein production. The classification of a miRNA as an oncogene or tumor suppressor hinges on its ability to respectively down-regulate tumor suppressors or genes involved in cell differentiation, contributing to cancer formation by promoting proliferation, angiogenesis, and invasion.

Additional classes of ncRNAs encompass small nuclear RNAs (snRNA), long non-coding RNAs (lncRNA), silencing RNA (siRNA), riboswitches, and internal ribosome entry sites (IRES). Small nucleolar RNA (snoRNA), a subset of snRNA, partakes in post-transcriptional modifications of rRNA alongside small nucleolar ribonucleoproteins (snoRNPs) with which they form complexes. Studies, such as those by Dong and colleagues, have identified disruptions in these RNA molecules in various conditions and cancer diseases. Specifically, snoRNA U50 has been recognized as a significant factor in the development and/or progression of breast cancer.

In diagram 1

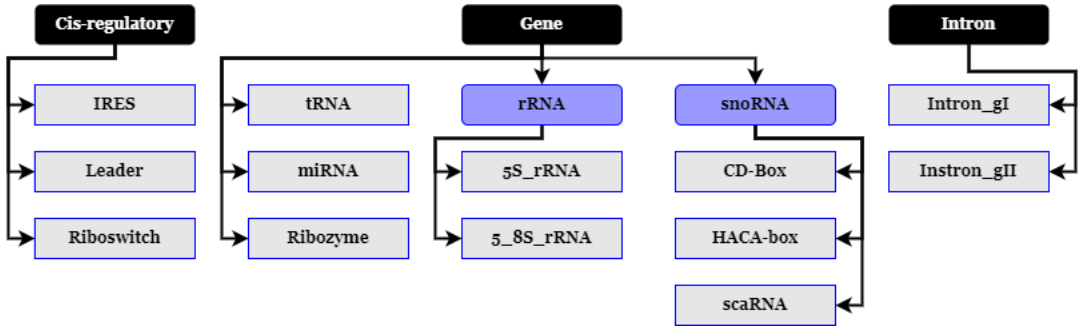

Figure 1: Small non-coding RNA Classes and sub-Classes

Given the extensive quantity and diverse functions of various non-coding RNAs (ncRNAs), accurately identifying and classifying them presents a novel and challenging bioinformatics scenario. With a significant portion of the "ncRNAome" yet to be uncovered and limited understanding of these non-coding molecules, their classification holds the potential to aid biologists and clinicians in comprehending the molecular mechanisms of this intricate regulatory system.

2.4 RNA Secondary Structure

The secondary structure of an RNA molecule refers to the arrangement of its nucleotides into base pairs and single strands, which determines its three-dimensional shape. RNA molecules can form a variety of secondary structures, including stem-loop structures, hairpin loops, and pseudoknots. These structures are formed by the formation of base pairs between complementary nucleotides (adenine with uracil and cytosine with guanine) and are stabilized by hydrogen bonds between the nucleotides.

The secondary structure of an RNA molecule is important because it plays a role in its function. For example, the secondary structure of transfer RNA (tRNA)

is important for its ability to bring amino acids to the ribosome during protein synthesis, while the secondary structure of ribosomal RNA (rRNA) is important for its structural and catalytic role in the ribosome. The secondary structure of other types of RNA, such as microRNA (miRNA) and long non-coding RNA (lncRNA), may also be important for their functions in the cell.

Predicting the secondary structure of an RNA molecule is important for understanding its function and for designing RNA molecules with specific functions. There are several computational tools that can be used to predict the secondary structure of an RNA molecule, including RNAcon and Mfold. These tools use various algorithms and models to predict the most likely secondary structure based on the nucleotide sequence of the molecule.

3 Machine Learning

TODO

4 Related Work

4.1 RNAcon

RNAcon [2] is a tool that includes two different prediction models: one for distinguishing non-coding and coding RNAs, and another for classifying predicted non-coding RNAs into various categories. The model for distinguishing between non-coding and coding RNAs uses a machine learning approach based on Support Vector Machines (SVMs) and nucleotide composition as input features. To optimize and evaluate the model, three different datasets were used and various kernels and parameters of the SVM were tested using a 10-fold cross-validation technique. The final model implemented in the RNAcon web server uses tri-nucleotide compositions (TNC) for discrimination between non-coding and coding RNA sequences. RNAcon algorithm is presented in figure 2

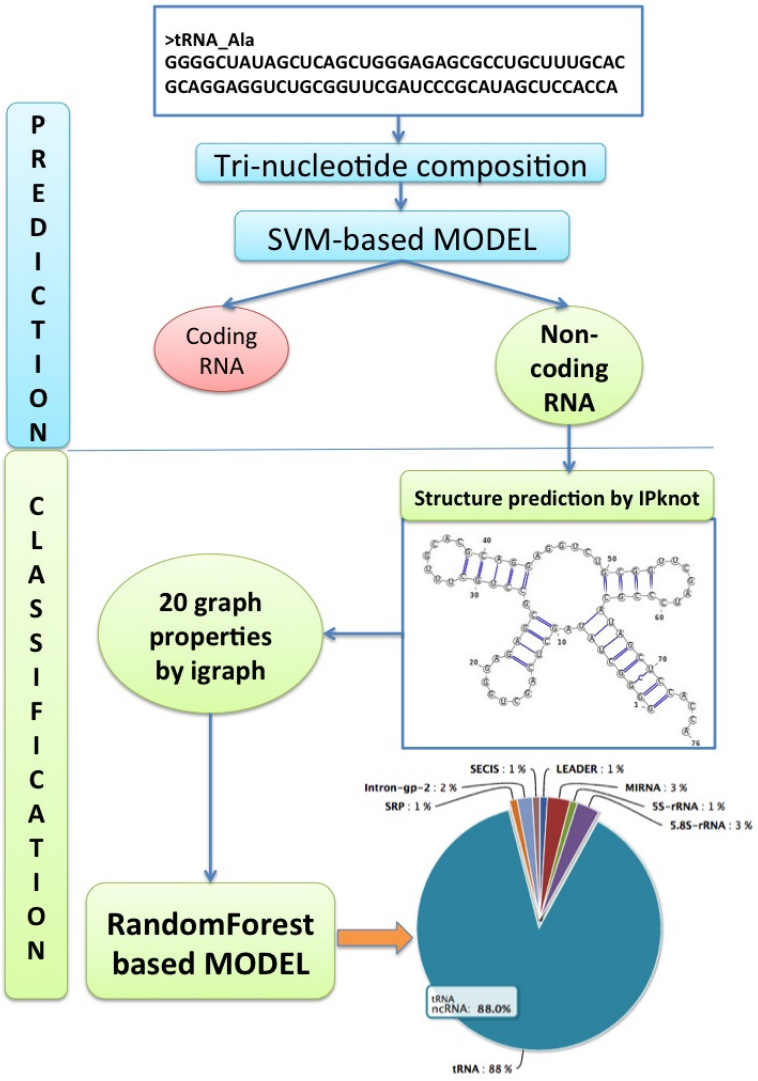

Figure 2: RNAcon Algorithm - a non-coding RNA Classifier

4.2 nRC

The nRC [1] tool is based on the extraction of features from the ncRNA secondary structure and a supervised classification algorithm using a deep learning architecture based on convolutional neural networks. The nRC tool was tested for the classification of 13 different ncRNA classes, similar to previous described RNAcon tool, and achieved an accuracy and sensitivity score of about 74%. The nRC tool outperformed other similar classification methods that have been developed until the year 2017 and were based on secondary structure features and machine learning algorithms, including the RNAcon tool, which was the reference classifier. Three steps are the basis of the proposed method:

1. The prediction of ncRNAs secondary structures, the extraction of frequent sub-structures as features and the classification of known ncRNA classes. To implement these processes, IPknot [3] algorithm was used to predict RNA secondary structures with pseudoknots,

2. the MoSS [4] decision tree pruning algorithm to obtain sub-structures

3. a deep learning network architecture, namely a convolutional neural network, as a supervised classifier.

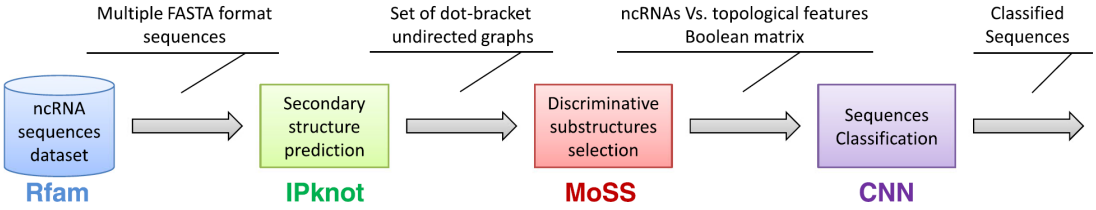

Figure 3: Pipeline of the nRC ncRNA sequence classification tool

4.3 GraPPLE

This study [5] investigates the use of specific properties of graphs that represent the predicted secondary structure of non-coding RNA (ncRNA) to reflect functional information. The authors developed a computational algorithm and a web-based tool called GraPPLE for classifying ncRNA molecules as functional and into Rfam families based on their graph properties. The tool was demonstrated to be more robust than sequence-similarity-based methods and covariance models with increasing sequence divergence and, when combined with existing methods, led to a significant improvement in prediction accuracy. The most informative graph properties were found to provide insight into the structural features that give ncRNA molecules functional properties. The GraPPLE tool may be useful for identifying potentially interesting ncRNA molecules among large candidate datasets.

4.4 ncRFP

Another approach [6], describes a method for predicting the family of non-coding RNAs (ncRNAs) called ncRFP. Traditional methods for ncRNA prediction involve predicting the secondary structure of the RNA and then identifying the ncRNA family based on the properties of the secondary structure. However, these methods can be complex and may not always be accurate due to errors that can accumulate in the multi-step process, particularly due to imperfections in tools used to predict the secondary structure of RNA. The ncRFP method is a novel approach that uses deep learning to predict the ncRNA family directly from the RNA sequence, bypassing the need to predict the secondary structure. This method simplifies the prediction process and improves accuracy compared to traditional methods.

4.5 NCodR

In this study [7], the authors developed eight classifiers for non-coding RNA (ncRNA) based on the sequence and structural measures of known ncRNAs. The multi-class support vector machine with a radial basis function kernel (SVM-RBF) model was found to be the best at discriminating ncRNAs and was implemented in a web server called NCodR for classifying ncRNAs in plants. The SVM-RBF model had an F-score of 0.96, with a specificity of 0.96 and a sensitivity of 0.99. This classifier can be used for genome-wide identification and classification of ncRNAs in various plant species, which will improve our understanding of gene regulation in plants at the transcriptional and post-transcriptional levels. The improved understanding of ncRNAs will enable the development of crops with improved yield, productivity, stress tolerance, and disease resistance through genome-editing technology. The classifier also has the ability to classify sequences in diverse taxonomic groups, advancing our knowledge of ncRNAs in general.

4.6 ncDLRES

The authors of this article [8] propose a novel method called ncDLRES for predicting the family of non-coding RNA (ncRNA) based on dynamic long short-term memory (LSTM) and residual neural network (ResNet). The method, called ncDLRES, extracts the features of ncRNA sequences using dynamic LSTM and then classifies them using ResNet. The authors compare ncDLRES to both the homologous sequence alignment method and other methods that predict ncRNA based on secondary structure, called ncRFP. The homologous sequence alignment method is currently the most accurate method, but it has limitations due to the need for consensus secondary structure annotation of ncRNA sequences and the inability to model pseudoknots. ncDLRES reduces the data requirements and expands the application scope compared to the homologous sequence alignment method, and its performance is greatly improved compared to the ncRFP methods.

4.7 ncDENSE

In this study [9], a method called ncDENSE, based on a deep learning model, was introduced. It predicts families of non-coding RNAs (ncRNAs) by analyzing the sequence features of these ncRNAs. The nucleotide bases in the sequences of ncRNAs were encoded using a one-hot coding scheme. These encoded sequences were then input into an ensemble deep learning model, which comprised three components: the dynamic bi-directional gated recurrent unit (Bi-GRU), the dense convolutional network (DenseNet), and the Attention Mechanism (AM). More specifically, the dynamic Bi-GRU was utilized to extract contextual feature information and capture long-term dependencies within the ncRNAs sequences. The AM was employed to assign varying weights to the features extracted by the Bi-GRU, focusing attention on information with higher weights. Meanwhile, DenseNet was employed to extract local feature information from the ncRNAs sequences and carry out classification using the fully connected layer.

4.8 MncR

Newest publication [10] on this subject in order to enhance the classification of ncRNAs, exploration of various methods involving the utilization of primary sequences and secondary structures, as well as their subsequent integration through machine learning models, including various neural network architectures. Employment of the latest version of RNAcentral, focusing on six ncRNA classes (lncRNA, rRNA, tRNA, miRNA, snRNA, and snoRNA) as input. The incorporation of graph-encoded structural features and primary sequences in our MncR classifier at a later stage resulted in an overall accuracy of >97%, which could not be further improved through finer subclassification. In comparison to the currently top-performing tool, ncRDense, observed a marginal increase of 0.5% across all four overlapping ncRNA classes on a similar test set of sequences. In summary, MncR not only surpasses existing ncRNA prediction tools in accuracy but also enables the prediction of long ncRNA classes (lncRNAs, specific rRNAs) of up to 12,000 nucleotides and is trained on a more diverse ncRNA dataset sourced from RNAcentral.

5 Technologies

5.1 RNA Secondary Structure Classifiers

5.1.1 IPknot

[3]

5.1.2 Knotify

[11]

5.2 Deep learning Frameworks

5.2.1 Tensorflow

TensorFlow is an open-source machine learning framework developed by Google. It is known for its computational graph paradigm, where computations are represented as a directed graph. Originally, TensorFlow used a static computation graph, meaning the entire structure of the graph had to be defined before computation could begin. However, TensorFlow 2.x introduced eager execution, which allows for more dynamic and intuitive development. TensorFlow supports a wide range of hardware, including CPUs, GPUs, and TPUs (Tensor Processing Units). It has a large and active community, offering extensive resources and libraries for various machine learning tasks.

5.2.2 Keras

Keras started as an independent high-level neural networks API, designed to be user-friendly and provide a simple interface for building and training models. It abstracts many low-level details, making it easy to create standard neural network architectures. Initially, Keras could run on top of different backends, such as TensorFlow, Theano, or Microsoft Cognitive Toolkit (CNTK). However, with TensorFlow 2.x, Keras was integrated as the official high-level API for building and training models. This allows users to leverage the strengths of both TensorFlow and Keras together.

5.2.3 PyTorch

PyTorch is an open-source deep learning framework developed by Facebook's AI Research lab (FAIR). It is known for its dynamic computation graph, which allows for more flexible and intuitive development. Unlike TensorFlow's earlier versions, PyTorch embraces a more Pythonic and imperative style, making it popular among researchers and practitioners who prefer a dynamic approach to model construction. PyTorch gained significant popularity in the research community due to its ease of use, strong support for GPU acceleration, and a highly active community.

5.2.4 Key Differences

Computational Graphs
TensorFlow originally used static computation graphs, while PyTorch employs dynamic computation graphs.Keras, when integrated with TensorFlow, follows TensorFlow's computational graph paradigm.

Ease of Use
TensorFlow had a steeper learning curve initially, but with TensorFlow 2.x and eager execution, it became more intuitive. Keras has always been designed for ease of use and user-friendliness. PyTorch is known for its intuitive and Pythonic programming style.

Flexibility and Control
TensorFlow provides both high-level and low-level APIs, offering a balance between abstraction and control. Keras abstracts many low-level details, providing less flexibility compared to TensorFlow's lower-level APIs. PyTorch offers a highly flexible and dynamic approach, allowing for a high degree of control over models.

Community and Ecosystem
TensorFlow has a large and established community with extensive resources and third-party libraries. Keras benefits from TensorFlow's ecosystem and has its own community as well. PyTorch gained rapid popularity in research communities, particularly for its dynamic computation graph.

Hardware Support
TensorFlow provides extensive support for various hardware, including CPUs, GPUs, and TPUs. PyTorch is well-suited for GPU acceleration and is adaptable to different hardware setups.

Ultimately, the choice between TensorFlow, Keras, and PyTorch depends on individual preferences, project requirements, and familiarity with the framework. All three are powerful tools with active communities, and they are widely used in both research and industry.

6 ncRNA data

6.1 ncRNA databases

A non-coding RNA database is a collection of ncRNA sequences and their corresponding functional and structural annotations. ncRNA databases provide a comprehensive resource for studying these molecules, as they allow researchers to easily access and analyze large amounts of high-quality data on ncRNAs. These databases often include multiple sequence alignments (MSAs) of ncRNA families, as well as functional and structural annotations for each ncRNA family. They may also provide cross-references to other databases and resources, such as databases of miRNA-disease associations. Here is a list of some popular non-coding RNA (ncRNA) databases:

- Rfam: A comprehensive database of ncRNA families and their annotated alignments, curated by the RNA Bioinformatics Group at the University of Cambridge.
- snoRNA-LBME-db: A database of small nucleolar RNAs (snoRNAs), which are a type of ncRNA that play a role in the modification of ribosomal RNAs (rRNAs) and small nuclear RNAs (snRNAs).
- miRBase: A database of microRNAs (miRNAs), which are small ncRNAs that play a role in the regulation of gene expression.
- tRNAscan-SE: A database of transfer RNAs (tRNAs), which are small ncRNAs that play a role in protein synthesis.
- piRNABank: A database of small interfering RNAs (siRNAs) and piwi-interacting RNAs (piRNAs), which are small ncRNAs that play a role in the regulation of gene expression and DNA methylation.
- lncRNApedia: A database of long non-coding RNAs (lncRNAs), which are a type of ncRNA that are longer than 200 nucleotides in length and are thought to play a role in various biological processes.
- ncRNA.org: A database of ncRNAs, including miRNAs, lncRNAs, and other types of ncRNAs.
- miR2Disease: A database of miRNA-disease associations, which provides information on the role of miRNAs in various diseases and disorders.

6.2 Rfam

Rfam is a database of non-coding RNA families and their annotated alignments, curated by the RNA Bioinformatics Group at the University of Cambridge. It is a comprehensive resource for ncRNA sequences and their corresponding functional and structural annotations.

The Rfam database is based on multiple sequence alignments (MSAs) of ncRNA families, which are groups of related ncRNA sequences that share a common ancestor

and are believed to have similar functions. These alignments are created using computational tools and are manually curated by expert annotators to ensure their quality and accuracy. In addition to the MSAs, Rfam also provides a variety of annotations for each ncRNA family, including functional descriptions, secondary structure predictions, and cross-references to other databases.

Rfam is an important resource for researchers studying ncRNAs, as it allows them to easily access and analyze large amounts of high-quality data on these molecules. It is particularly useful for identifying and classifying novel ncRNA sequences, as it provides a comprehensive set of known ncRNA families that can be used as a reference. Rfam is constantly updated with new ncRNA families and annotations, making it an invaluable resource for the RNA community.

6.2.1 Rfam MySQL database

Rfam provides a public read-only MySQL database containing the latest version of Rfam data. Details to access this database are provided in Table 1. The database core scheme is shown in diagram 4.

Parameter	Value
host	mysql-rfam-public.ebi.ac.uk
port	4497
user	rfamro
database	Rfam

Table 1: Rfam MySQL Database - Connection Details

Advanced examples of using the public Rfam database can be found in Current Protocols in Bioinformatics publication [12]. Some examples bellow:

```
1
2  -- Number of Sequences per file and ncRNA type
3  -- ------------------------------------------------------------
4  SELECT family.rfam_acc , family.type ,
5      COUNT(full_region.rfamseq_acc) as Number_of_Sequences
6  FROM full_region, family
7  Where family.rfam_acc = full_region.rfam_acc
8  GROUP BY family.rfam_acc;
9
10 -- For RF00014.fa fasta file show
11 -- sequence id, start and stop, description and type
12 -- This query will be used to construct fasta headers
13 -- ------------------------------------------------------------
14 SELECT full_region.rfam_acc, full_region.rfamseq_acc, seq_start, seq_end,
15     rfamseq.description, family.type
16 FROM full_region , rfamseq , family
17 WHERE full_region.rfamseq_acc = rfamseq.rfamseq_acc
18 AND family.rfam_acc = full_region.rfam_acc
19 AND full_region.rfam_acc = 'RF00014';
20
```

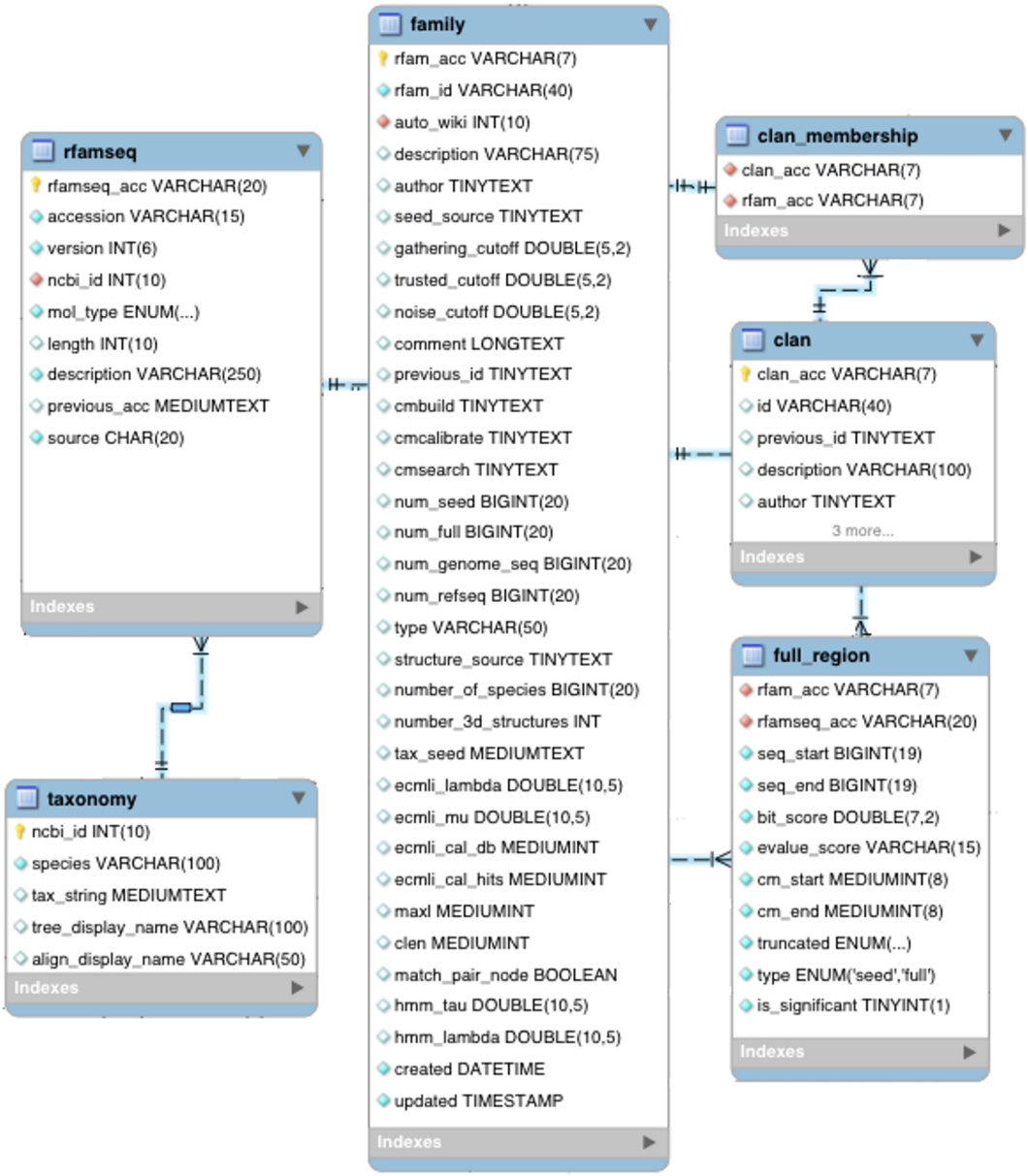

Figure 4: Rfam core database scheme diagram

6.3 Data Collection

To assemble a dataset of small non-coding RNA sequences, Rfam database was mainly used, The bellow SQL query was used to retrieve all .fasta file names that contain RNA sequences of the RNA families in interest.

```
-- Get all .fasta file names
-- ------------------------------------------------------------
SELECT rfam_acc FROM family
WHERE type = "Cis-reg; IRES;" OR type = "Cis-reg; leader;"
OR type = "Cis-reg; riboswitch;" OR type = "Gene; miRNA;"
OR type = "Gene; ribozyme;" OR type = "Gene; snRNA; snoRNA; CD-box;"
OR type = "Gene; snRNA; snoRNA; HACA-box;" OR type = "Gene; snRNA; snoRNA; scaRNA;"
```

```
OR type = "Gene; tRNA"
```

This query returns a list of all the fasta filenames:

```
# The response of the mySQL rfam server
# ------------------------------------------------------------
['RF00008', 'RF00009', 'RF00010', 'RF00011', ... , 'RF04235', 'RF04236']
```

Each file is downloaded from the SFTP server :

```
acc_id_list = ['RF00008', 'RF00009', 'RF00010', 'RF00011' , 'RF04235', 'RF04236']
for filename in acc_id_list:
    URL = "http://http.ebi.ac.uk/pub/databases/Rfam/CURRENT/fasta_files/"
    URL += filename + ".fa.gz"
    wget.download(URL, "../datasets/Rfam/"+filename+".fa.gz")
```

Example of RF00004.fa fasta file contents:

```
>CM001883.1/36104473-36104278 Theobroma cacao cultivar Matina 1-6 chromosome ...
ATACCTTTCTCGGCCTTTTGGCTAAGATCAAGTGTAGTATCTGTTCTTATCAGTTTAATATCTGATACGTGGGCCA ...
>JH795869.1/919604-919793 Dacryopinax sp. DJM-731 SS1 chromosome Unknown DAC ...
GCACCACTCTGGCCTTTTGGCTTAGATCAAGTGTAGTATCTGTTCTTATTAGTTTAACCACTAATATGGTCGCACC ...
>CM001769.1/9270458-9270652 Cicer arietinum chromosome Ca6, whole genome sho ...
ATACCTTTCTCGGCCTTTTGGCTAAGATCAAGTGTAGTATCTGTTTTTATCAGTTTAATATCTGATATGTGGTCCA ...
>FR853084.2/62466812-62466957 Gorilla gorilla gorilla genomic chromosome, ch ...
ATTACTTCTCAGCCTTTTGGCTAAGATCAAGTGTAATAAATCTCATTGTGCTTTATGCCTAATGTGTGCTTATATT ...
>KK088422.1/566554-566746 Aspergillus ruber CBS 135680 unplaced genomic scaf ...
CCAGCTCTCTTTGCCTTTTGGCTTAGATCAAGTGTAGTATCTGTTCTTTTCAGTTTAATCTCTGAAAGTGTTCTAA ...
>AACT01051284.1/529-401 Ciona savignyi cont_51284, whole genome shotgun sequ ...
ACAGCTGATGCCGCAGCTACACTATGTATTAATCGGATTTTTGAACTTGGAGTACGGTTCTGGAGCTTGCTCCACC ...
```

Explain how fasta files are formated

```
>IRES
ATACCTTTCTCGGCCTTTTGGCTAAGATCAAGTGTAGTATCTGTTCTTATCAGTTTAATATCTGATACGTGGGCCA ...
>tRNA
GCACCACTCTGGCCTTTTGGCTTAGATCAAGTGTAGTATCTGTTCTTATTAGTTTAACCACTAATATGGTCGCACC ...
>tRNA
ATACCTTTCTCGGCCTTTTGGCTAAGATCAAGTGTAGTATCTGTTTTTATCAGTTTAATATCTGATATGTGGTCCA ...
>riboswitch
ATTACTTCTCAGCCTTTTGGCTAAGATCAAGTGTAATAAATCTCATTGTGCTTTATGCCTAATGTGTGCTTATATT ...
>HACA-box
CCAGCTCTCTTTGCCTTTTGGCTTAGATCAAGTGTAGTATCTGTTCTTTTCAGTTTAATCTCTGAAAGTGTTCTAA ...
>tRNA
ACAGCTGATGCCGCAGCTACACTATGTATTAATCGGATTTTTGAACTTGGAGTACGGTTCTGGAGCTTGCTCCACC ...
```

Files of the same family were combined in order to generate one .fasta file per RNA family. IRES non-coding RNA family dataset was poor, so a second data source, IRESbase [13], dedicated to this family was used to extend the initial dataset. In table 2 the number of sequences of each family:

RNA Family	Source	Number of Sequences
IRES	Rfam	2800
IRES	IRESbase	1328
leader	Rfam	31662
riboswitch	Rfam	69465
miRNA	Rfam	387173
ribozyme	Rfam	220007
CD-box	Rfam	132915
HACA-box	Rfam	36938
scaRNA	Rfam	2962
tRNA	Rfam	1432442
5S_rRNA	Rfam	140644
5_8S_rRNA	Rfam	4940
Intron_gpI	Rfam	2611
Intron_gpII	Rfam	15729

Table 2: Number of Sequences per family per Data source

6.4 Final dataset

The final dataset has nearly 4-5 thousand sequences per family, with some exceptions. The distribution of our data is shown in diagram 5

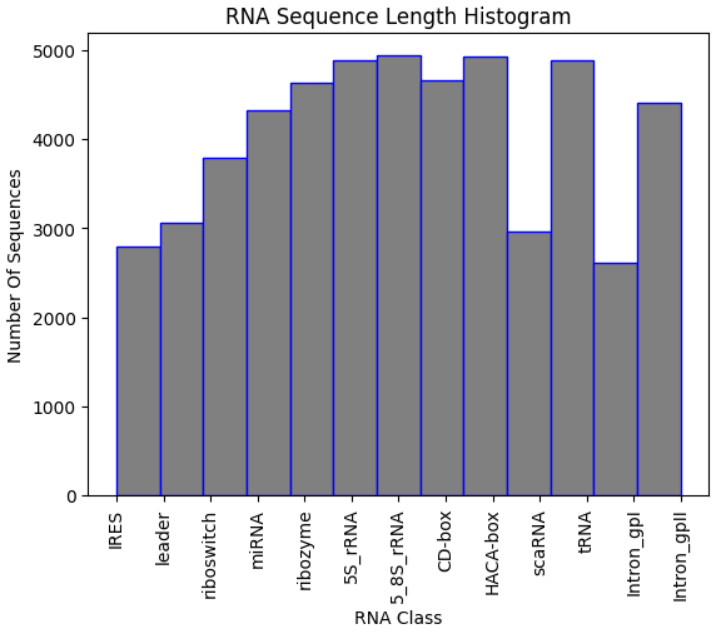

Figure 5: Number of sequences per class distribution of final dataset

7 Implementing the Prediction Mechanism.

7.1 Data preparation

The dataset used in this research consist of nearly 9000 number of non-coding RNA sequences. The same dataset that has been used in [1] as shown in Table 3

Train	6160 seq
Test	2529 seq

Table 3: nRC [1] dataset details

7.1.1 Sequence Padding

Answer these questions:

1. Find the maximum sequence length of the dataset
2. Why padding and not cutting ?
3. What is the padding length

7.1.2 One-hot encoding

Since there are no known relations between the bases of RNA One-hot encoding was used to encode the sequences before using them as input into the deep learning model.

8 Conclusion

References

[1] Antonino Fiannaca, Massimo La Rosa, Laura La Paglia, Riccardo Rizzo, and Alfonso Urso. nrc: non-coding rna classifier based on structural features. *BioData Mining*, 10, 2017. doi: https://doi.org/10.1186/s13040-017-0148-2. URL https://doi.org/10.1186/s13040-017-0148-2.

[2] Bharat Panwar, Amit Arora, and Gajendra PS Raghava. Prediction and classification of ncrnas using structural information. *BMC Genomics*, 15, 2014. ISSN 1471-2164. doi: https://doi.org/10.1186/1471-2164-15-127. URL https://doi.org/10.1186/1471-2164-15-127.

[3] Kengo Sato, Yuki Kato, Michiaki Hamada, Tatsuya Akutsu, and Kiyoshi Asai. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics*, 27(13):i85–i93, 06 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr215. URL https://doi.org/10.1093/bioinformatics/btr215.

[4] Christian Borgelt, Thorsten Meinl, and Michael Berthold. Moss: A program for molecular substructure mining. In *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, OSDM '05, page 6–15, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595932100. doi: 10.1145/1133905.1133908. URL https://doi.org/10.1145/1133905.1133908.

[5] Liam Childs, Zoran Nikoloski, Patrick May, and Dirk Walther. Identification and classification of ncRNA molecules using graph properties. *Nucleic Acids Research*, 37(9):e66–e66, 04 2009. ISSN 0305-1048. doi: 10.1093/nar/gkp206. URL https://doi.org/10.1093/nar/gkp206.

[6] Linyu Wang, Shaoge Zheng, Hao Zhang, Zhiyang Qiu, Xiaodan Zhong, Haiming Liu, and Yuanning Liu. ncrfp: A novel end-to-end method for non-coding rnas family prediction based on deep learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(2):784–789, 2021. doi: 10.1109/TCBB.2020.2982873.

[7] Chandran Nithin, Sunandan Mukherjee, Jolly Basak, and Ranjit Prasad Bahadur. Ncodr: A multi-class support vector machine classification to distinguish non-coding rnas in viridiplantae. *Quantitative Plant Biology*, 3:e23, 2022. doi: 10.1017/qpb.2022.18.

[8] Linyu Wang, Xiaodan Zhong, Shuo Wang, and Yuanning Liu. ncdlres: a novel method for non-coding rnas family prediction based on dynamic lstm and resnet. *BMC Bioinformatics*, 22, 09 2021. doi: 10.1186/s12859-021-04365-4. URL https://doi.org/10.1186/s12859-021-04365-4.

[9] Kai Chen, Xiaodong Zhu, Lei Hao, Jiahao Wang, Zhen Liu, and Yuanning Liu. ncdense: a novel computational method based on a deep learning framework for non-coding rnas family prediction. *BMC Genomics*, 12 2022. doi: 10.21203/rs.3.rs-2374139/v1.

[10] Heiko Dunkel, Henning Wehrmann, Lars R. Jensen, Andreas W. Kuss, and Stefan Simm. Mncr: Late integration machine learning model for classification of ncrna classes using sequence and structural encoding. *International Journal of Molecular Sciences*, 24(10), 2023. ISSN 1422-0067. doi: 10.3390/ijms24108884. URL https://www.mdpi.com/1422-0067/24/10/8884.

[11] Christos Andrikos, Evangelos Makris, Angelos Kolaitis, Georgios Rassias, Christos Pavlatos, and Panayiotis Tsanakas. Knotify: An efficient parallel platform for rna pseudoknot prediction using syntactic pattern recognition. *Methods and Protocols*, 5(1), 2022. ISSN 2409-9279. doi: 10.3390/mps5010014. URL https://www.mdpi.com/2409-9279/5/1/14.

[12] Ioanna Kalvari, Eric P. Nawrocki, Joanna Argasinska, Natalia Quinones-Olvera, Robert D. Finn, Alex Bateman, and Anton I. Petrov. Non-coding rna analysis using the rfam database. *Current Protocols in Bioinformatics*, 62(1):e51, 2018. doi: https://doi.org/10.1002/cpbi.51. URL https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpbi.51.

[13] Jian Zhao, Yan Li, Cong Wang, Haotian Zhang, Hao Zhang, Bin Jiang, Xuejiang Guo, and Xiaofeng Song. Iresbase: A comprehensive database of experimentally validated internal ribosome entry sites. *Genomics, Proteomics & Bioinformatics*, 18(2):129–139, 2020. ISSN 1672-0229. doi: https://doi.org/10.1016/j.gpb.2020.03.001. URL https://www.sciencedirect.com/science/article/pii/S1672022920300577. Special Issue:Bioinformatics Commons—2020.

A This is an appendix