Recognition of ncRNA regions & RNA Families using Neural Networks, to reduce the search time of Infernal: inference of RNA alignments

Background Information

RNA

Ribonucleic acid (RNA) is a nucleic acid, present in most living cells. It is largely known for its role of acting as a messenger that carries the information stored in DNA for the synthesis of proteins.

Though once considered “junk”, the existence of ncRNAs was first discovered over 60 years ago (1).Since then he known functions and roles of ncRNAs is quickly expanding and includes the cleaving of other RNA (2), regulation of macromolecule structure (3), and cell-cycle regulation (4), its capacity for catalytic activity (2, 5, 6), and gene regulation at the transcriptional and post-transcriptional level (7). Just to name a few.

RNA is grouped into families based on two main criteria: sequences having homology by common descent (that is, evolution from a common ancestor) (8), and secondary structures. In particular, the use of secondary structure is advantageous because the secondary structure of functional ncRNAs is often more conserved than the nucleotide sequence (9).

Machine Learning: Neural Networks

Artificial Neural Networks (ANNs) have been around since 1943 (10). They are modelled after biological neurons found in animal cerebral cortexes (11). They excel at large and complex Machine Learning tasks, ergo, making them potentially good at classifying RNA.

Convolutional Neural Networks (CNNs) emerged from studying the brain’s *visual* cortex (11). It introduced two new ideas into the structure of ANNs, convolutional layers and pooling layers. CNNs are particularly useful in RNA classification because they can potentially retain spatial information.

Rfam & Infernal

Rfam, a database of RNA families has millions of aligned sequences making up 4094 families (as of May 2022). The families in Rfam are represented by a multiple sequence alignment of known RNA sequences (called the seed) and a Covariance Model (CM) (12).

These families/clans, and any new ones are found mostly using Infernal. By its own descriptions Infernal builds probabilistic profiles of the RNA sequence and secondary structure of RNA families. They call these profiles covariance models (CMs). Each RNA family will end up with a covariance model associated with it.

Infernal uses a program called cmscan to search a sequence against a CM database (e.g., Rfam). It takes a single query sequence and a CM database as input and will find the different known/detectable RNAs in a sequence.

The Problem

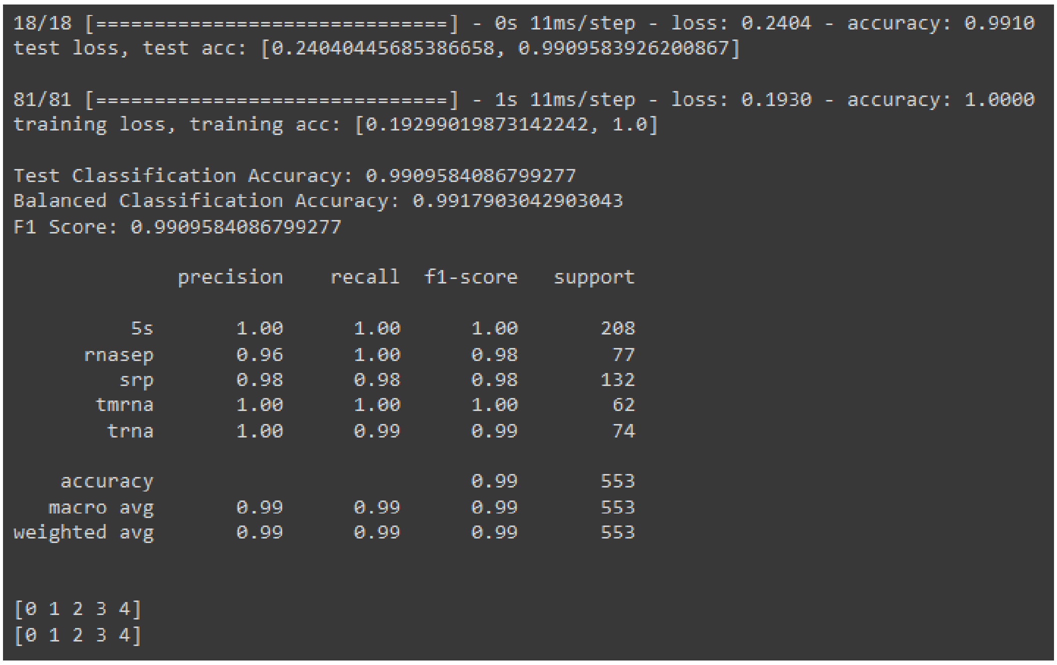
When Infernal’s cmscan searches a sequence against a CM database it will query ALL the different CMs in the database. The size of the search space for cmscan is double the length of the sequence (because the strand and its complement will be searched) multiplied by the number of models in the CM database (Rfam, for example, has 4094). (13). This process is exhaustive, but it is slow. Though improvements have been made on speed it would still take ~3h to search all Rfam models against the 1 Gb chicken genome, on a 100-CPU compute cluster (13). Finding a way to speed up this process would help greatly with the classification of ncRNAs into families.

Solving the Problem

One possible way of doing this is by building a neural network to classify sequences in RNA as either belonging to a pre-existing RNA family, or not belonging to a pre-existing RNA family. If a network can do this with a high sensitivity (~100%) we can feed these smaller sequences of the whole genome into cmscan, effectively cutting down the search space, improving its speed. Previous papers have even shown that a CNN can achieve a sensitivity of 88.04% (14), which is a promising result. ANNs are desirable for these purposes because they can balance between trying to obtain high sensitivity and precision.

Another way is by reducing the number of CMs that need to be run against the sequences. This might be possible by using a neural network to classify parts of the genome that belong to RNA families into different “bins”. Bins that would contain a grouping of RNA families. Having each bin contain a single RNA family is a bit idealistic and would be hard to achieve in the timeframe of this project. But possible avenues exist in Rfams own groupings of RNA families into clans, or personal home-brewed bins. Here, proof that such a concept is feasible is the goal. From the literature I could find any instances of trying to classify ncRNA, using ANNs, with all the families in Rfam (usually a handful of families are picked).

Preliminary results are promising with a small CNN showing 99% classification accuracy on ArchiveII (an RNA dataset consisting of 5 families).



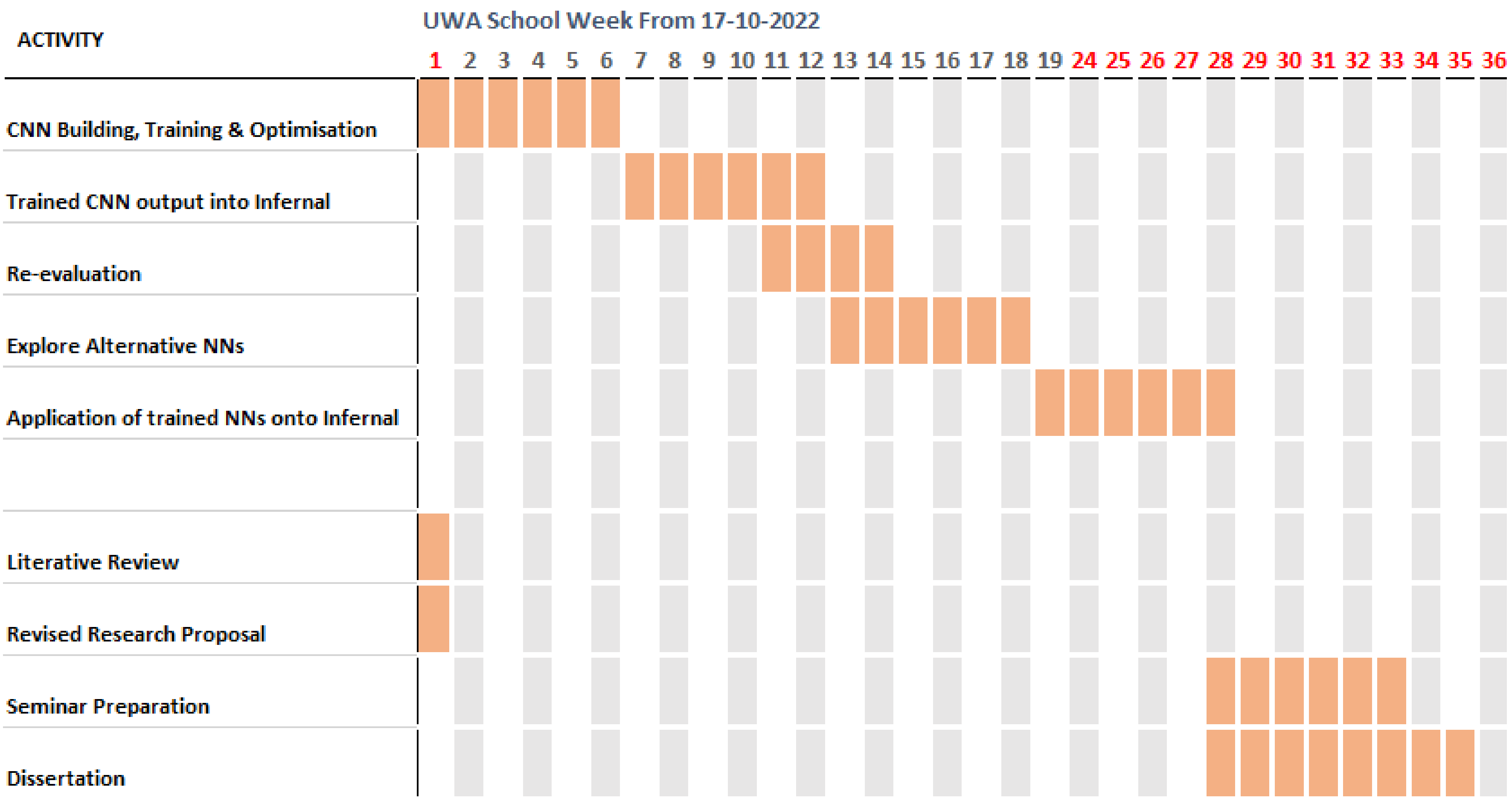
```
18/18 [==============================] - 0s 11ms/step - loss: 0.2404 - accuracy: 0.9910
test loss, test acc: [0.24040445685386658, 0.9909583926200867]

81/81 [==============================] - 1s 11ms/step - loss: 0.1930 - accuracy: 1.0000
training loss, training acc: [0.19299019873142242, 1.0]

Test Classification Accuracy: 0.9909584086799277
Balanced Classification Accuracy: 0.9917903042903043
F1 Score: 0.9909584086799277

              precision    recall  f1-score   support

          5s       1.00      1.00      1.00       208
      rnasep       0.96      1.00      0.98        77
         srp       0.98      0.98      0.98       132
       tmrna       1.00      1.00      1.00        62
        trna       1.00      0.99      0.99        74

    accuracy                           0.99       553
   macro avg       0.99      0.99      0.99       553
weighted avg       0.99      0.99      0.99       553


[0 1 2 3 4]
[0 1 2 3 4]
```

Aims

- Build Neural Networks to recognise the presence of RNA families in genomes, with high sensitivity
- Build a Neural Network to classify RNA families in genomes into Rfam families

Timeline

Project Timeline



ACTIVITY	UWA School Week From 17-10-2022																															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	24	25	26	27	28	29	30	31	32	33	34	35	36
CNN Building, Training & Optimisation	🟧	🟧	🟧	🟧	🟧	🟧																										
Trained CNN output into Infernal							🟧	🟧	🟧	🟧	🟧	🟧																				
Re-evaluation											🟧	🟧	🟧	🟧																		
Explore Alternative NNs													🟧	🟧	🟧	🟧	🟧	🟧														
Application of trained NNs onto Infernal																			🟧	🟧	🟧	🟧	🟧	🟧								
Literative Review	🟧																															
Revised Research Proposal	🟧																															
Seminar Preparation																								🟧	🟧	🟧	🟧	🟧	🟧			
Dissertation																								🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	

References

1. Palazzo AF, Lee ES. Non-coding RNA: What is functional and what is junk? Frontiers in genetics. 2015;5:2–2.

2. Braunwald E, Libby P, Bhatt D, Mann DL, Solomon SD, Bonow RO, et al. Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine. Elsevier; 2021.

3. Ferguson BS. Nutritional Epigenomics. San Diego: Elsevier Science & Technology; 2019.

4. Bertrand-Lehouillier V, Legault LM, McGraw S. Endocrine Epigenetics, Epigenetic Profiling and Biomarker Identification.

5. Peedicayil J. Non-coding RNAs and psychiatric disorders. Epigenetics in Psychiatry 2021 Jan 1 (pp. 321-333). Academic Press.

6. Stark BC, Kole R, Bowman EJ, Altman S. Ribonuclease P: An Enzyme with an Essential RNA Component. Proceedings of the National Academy of Sciences - PNAS. 1978;75(8):3717–21.

7. Yang VW, Lerner MR, Steitz JA, Flint SJ. A Small Nuclear Ribonucleoprotein is Required for Splicing of Adenoviral Early RNA Sequences. Proceedings of the National Academy of Sciences - PNAS. 1981;78(3):1371–5.

8. Dubitzky W, Wolkenhauer O, Cho KH, Yokota H, editors. Encyclopedia of systems biology. New York: Springer; 2013 Aug 17.

9. Soldà G, Makunin IV, Sezerman OU, Corradin A, Corti G, Guffanti A. An Ariadne's thread to the identification and annotation of noncoding RNAs in eukaryotes. Briefings in bioinformatics. 2009;10(5):475–89.

10. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics. 1943 Dec;5(4):115-33.

11. Géron A. Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems. Sebastopol, California: O'Reilly Media, Inc.; 2017.

12. Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Research. 2021;49(D1):D192–D200.

13. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. BIOINFORMATICS. 2013;29(22):2933–5.

14. Chantsalnyam T, Lim DY, Tayara H, Chong KT. ncRDeep: Non-coding RNA classification with convolutional neural network. Computational biology and chemistry. 2020;88:107364–107364.