Understanding the Sequence and Structural determinants of microRNA processing using large datasets

Master's in Bioinformatics Thesis

Aakriti Singh

Supervisors:

Xavier Bofill De Ros and Christian Nørgaard Storm Pedersen

Spring 2025

Aarhus University

ABSTRACT

MicroRNAs (miRNAs) are small non-coding RNAs that regulate gene expression post-transcriptionally and play important roles in development and disease. Their biogenesis depends on precise cleavage of primary miRNA transcripts (pri-miRNAs) by the nuclear Microprocessor complex, composed of Drosha and DGCR8. Although several structural and sequence features influencing pri-miRNA processing have been identified, a comprehensive quantitative understanding of their relative contributions remains incomplete.

In this master thesis, I analyzed published high-throughput datasets comprising over 210,000 variants from three pri-miRNA substrates, capturing both cleavage efficiency and cleavage fidelity. I performed exploratory data analysis to examine the distribution of cleavage outcomes and identify substrate-specific patterns. I then engineered biologically motivated features capturing both structural properties of pri-miRNA secondary structure, such as paired fraction, stem length, stem-to-loop ratio, bulge size, and minimum free energy, and sequence features including GC content, sequence length, mutation burden, motif-related signals, and positional mutation patterns.

Using these features, I trained and evaluated Random Forest and XGBoost models for both regression and classification tasks to predict Drosha-mediated processing outcomes. Model performance was assessed using cross-validation, and feature importance analysis was conducted to identify the most influential predictors. Classification outperformed regression, with Random Forest achieving an AUC of 0.938 and XGBoost achieving an AUC of 0.924. The results show that both models captured meaningful biological signals, with structural flexibility and sequence motifs emerging as key determinants of processing outcomes.

This work provides a computational framework for predicting the impact of sequence and structure-based features on Drosha-mediated miRNA biogenesis. It may help interpret disease-associated variants and inform therapeutic design.

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude to my supervisors, Assistant Professor Xavier Bofill De Ros and Professor Christian Storm Pedersen, for their invaluable guidance, continuous support, and patience throughout this project. Their expertise and enthusiasm for the research have been a constant source of inspiration and have greatly shaped the direction and quality of this thesis.

I would also like to extend my thanks to the Bioinformatics Research Centre (BiRC) and the Department of Molecular Biology at Aarhus University for providing a stimulating and supportive research environment. I am particularly grateful to my fellow Master's students in Bioinformatics for their moral support, encouragement, and the memorable moments we shared throughout this journey.

Finally, and most importantly, I owe my deepest gratitude to my family. To my husband and my parents — thank you for your unconditional love, unwavering belief in me, and endless encouragement. This work would not have been possible without you.

TABLE OF CONTENT