

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

Abstract	i
Acknowledgement	ii
Chapter 1: Introduction	1
1.1 Overview	1
1.2 Research Questions	1
1.3 Research Aims	1
1.4 Scope and Delimitations	2
Chapter 2: Related Works	3
2.1 Machine Learning Models for Predicting Human Dicer Cleavage Sites	3
2.2 Deep Learning Models for RNA Cleavage Prediction	3
2.3 Computational Approaches to miRNA Precursor Prediction	4
2.4 Interpretable Machine Learning and Deep Learning Approaches	4
2.5 Summary and Research Gap	4
Chapter 3: Background Theory	6
3.1 Introduction to miRNA Biogenesis	6
3.2 Drosha Cleavage: Structural and Sequence Determinants	6
3.2.1 Architecture of the pri-miRNA Substrate	6
3.2.2 Conserved Sequence Motifs	7
3.2.3 Stem Length and Structural Flexibility	7
3.2.4 Cleavage Fidelity, isomiRs and Functional Consequences	8
3.2.5 Disease Relevance and Variant Impact	8
Chapter 4: Methods	10
4.1 Overview	10
4.2 Dataset and Data Sources	10
4.3 Variant-Specific Distributional Patterns	10
4.4 Target Variable Selection	11
4.4.1 Cleavage Efficiency	11
4.4.2 Cleavage Fidelity	12
4.5 Feature Extraction	12
4.5.1 Structure-Based Features	12
4.5.2 Sequence-Based Features	14
4.5.3 Combined Feature Set	16
4.5.4 Modelling Strategy	19
4.5.4.1 Structure-Based Modelling	19
4.5.4.2 Sequence-Based and Combined Modelling	20
4.5.4.3 Target Variables Across Pipelines	

4.6	Feature Preprocessing and Normalisation	21
4.6.1	Missing Value Handling	21
4.6.2	Near-Zero Variance Feature Removal	21
4.6.3	Feature Scaling	22
4.6.4	Target Variable Processing	22
4.6.5	Class Imbalance	23
4.6.6	Severity-Stratified Sampling and Case Weighting	23
4.7	Modelling Framework	24
4.7.1	Overview	24
4.7.2	Regression	24
4.7.3	Classification	25
4.7.4	Random Forest	26
4.7.5	XGBoost	27
4.8	Feature Importance and Interpretability Analysis	27
4.8.1	Model-Based Feature Importance	27
4.8.2	SHAP Analysis	28
Chapter 5: Results & Discussion		29
5.1	Dataset Overview and Per-miRNA Characteristics	29
5.1.1	Dataset Composition	29
5.1.2	Per-miRNA Distribution of Alternative Cleavage	29
5.1.3	Relationship Between Folding Energy, GC Content, and Alternative Cleavage	30
5.1.3.1	Folding Energy	30
5.1.3.2	GC Content	30
5.2	Structural Analysis of pri-miRNA Variants	31
5.2.1	Wild-Type Structure Identification	31
5.2.2	Structure-Grouped Variant Analysis	32
5.2.3	Relationship Between Cleavage Fidelity and Efficiency	33
5.2.4	Thermodynamic Stability and GC Content Differences	34
5.2.5	Effect of Internal Bulge Pairs on Alternative Cleavage	35
5.2.6	G·U Wobble Pair Enrichment Promotes Alternative Cleavage	36
5.3	Feature Correlation Analysis	37
5.3.1	Structural Feature Correlations with log2FC	37
5.3.2	Sequence-Based Feature Correlations	38
5.3.3	Combined Structure and Sequence Feature Correlations	38
5.4	Structure-Based Modelling Results	39
5.4.1	Regression	39
5.4.2	Classification	40
5.5	Sequence-Based Modelling Results	41
5.5.1	Regression	41

5.5.2 Classification	42
5.6 Combined Structure and Sequence Modelling Results	42
5.6.1 Regression	42
5.6.2 Classification	43
5.7 Feature Importance Analysis	44
5.7.1 Structure-Based Models	44
5.7.2 Sequence-Based Model	45
5.7.3 Combined Sequence and Structure Model	46
5.8 Cross-Dataset Generalisation	46
5.9 Discussion	47
5.10 Future Perspectives	47
Chapter 5: Conclusion	49
References	50
Appendix A — Data Curation and Feature Extraction	54
Appendix B — Supplementary Results	56