



Developing a Deep Learning Model to Predict Mutation Probabilities at Single Base Resolution

Nimród Grandpierre

Master's Thesis

June 15th, 2025

ABSTRACT

Deep learning has undergone a resurgence over the past decade and a half, proving to be highly effective across a wide range of domains, including genomics. The mutational landscape - a topic within the field of genomics - has previously been investigated by applying both traditional statistical learning methods and neural networks. In this thesis, a new neural network-based approach is presented that aims to infer mutation outcome at single base resolution across the human genome. A key hypothesis in this domain that has gathered significant proof is that the most prevalent factor in a substitution being developed is the local context around the nucleotide site. This insight suggests that a k -mer representation of local nucleotide context captures the primary signal in mutation prediction. By utilizing both the local context and additional features that inform predictions about the characteristics of the broader region around the site, neural network models gain a comprehensive understanding of factors determining mutation outcome, leading to impressive predictive results.



CONTENTS

- 1 Introduction 1**
 - 1.1 Background and Motivation 1
 - 1.2 Research Problem and Objectives 3
 - 1.3 Neural Networks 4
 - 1.3.1 Fundamentals 4
 - 1.3.2 Deep Learning in Genomics 9
 - 1.4 Biological Background 10
 - 1.4.1 Sources of Data 12
 - 1.4.2 Existing Models around Mutation Rates 15
- 2 Methods 19**
 - 2.1 Data 19
 - 2.2 Preprocessing 20
 - 2.3 Benchmark Model 24
 - 2.4 Model Architecture 25
 - 2.4.1 Training Configuration 29
 - 2.5 Experimental Setup 30
 - 2.6 Evaluation Metrics 31
- 3 Results 33**
 - 3.1 Model Performance 33
 - 3.1.1 "A" Central Nucleotide 33
 - 3.1.2 "C" Central Nucleotide 37
 - 3.2 Best Model Selection 38
 - 3.3 Performance Across Different *k*-mer Lengths and Bin Sizes 38
 - 3.4 Feature Importance 39
 - 3.5 Summary of Key Results 39
- 4 Discussion 41**
 - 4.1 Interpretation of Results 41

4.2	Strengths and Limitations	45
5	Conclusion, Future Work	49
A	Appendix	57