HIDDEN-RBP: Harnessing Intelligent Data Discovery to Explore Networks of RNA-Binding Proteins

Master's in Bioinformatics Thesis

Iresha De Silva

Supervisors:

Xavier Bofill De Ros and Thomas Bataillon

Spring 2025

Aarhus University

ABSTRACT

Understanding RNA-binding proteins (RBPs) is crucial for elucidating the regulatory mechanisms that control post-transcriptional gene expression and for identifying molecular pathways implicated in development and disease. In this study, I trained and compared supervised and unsupervised machine learning models to predict functional interactions involving RBPs using a combination of curated features and CRISPR-Cas9 derived phenotypic screening data. Two supervised models, Lasso and Random Forest, demonstrated strong classification performance on known interaction datasets, with the Random Forest model achieving 84% accuracy and a ROC-AUC of 0.90. In parallel, a Bayesian Latent Variable Model was employed to infer interaction likelihoods in an unsupervised manner, capturing hidden structure in the data by modeling uncertainty and leveraging prior knowledge.

The Bayesian model effectively identified latent signals associated with phenotypic similarity, particularly those derived from CRISPR screens, and generalized these patterns to unlabeled gene interaction pairs. Among 25,000 unverified pairs, 12 were identified with latent scores overlapping those of known interactions, including BRD4_TWIST2 and COX4I1_COX6B2, which show potential functional relevance based on existing literature. While posterior parameters convergence, referring to the stabilization of key latent variables or interaction strength estimates, was limited in some cases, particularly at lower training iterations or when key biological signals or features were missing, the findings nonetheless underline the value of probabilistic models in uncovering biologically meaningful patterns in sparse and noisy datasets.

These results highlight the critical role of CRISPR-Cas9 data in modeling RBP-related functions and suggest that incorporating additional features from RBPs datasets and structural, sequence, and pathway databases could further improve prediction accuracy. This integrative approach offers a framework for prioritizing novel RBP interactions and supports their downstream validation through experimental studies.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to everyone who has supported me throughout the completion of this thesis.

I am deeply grateful to my supervisors, Asst. Professor Xavier Bofill De Ros and Professor Thomas Bataillon, for your invaluable guidance, continuous support, and patience during my research. Your immense knowledge and experience have been influential in shaping this thesis. Your enthusiasm for the research is truly inspiring.

I would also like to thank the members of the XBR Lab Group and my DDSA mentor Myky Tran for the insightful discussions and help through the thesis for assistance in navigating and working with the dataset. Your contributions have significantly improved the quality of my work.

Special thanks to the Bioinformatics Research Center (BiRC) and the Department of Molecular Biology at Aarhus University for providing a stimulating environment. I am particularly thankful to my fellow Master's in Bioinformatics students for their moral support and the fun times we shared.

I would like to extend my heartfelt thanks to my family for their unwavering support and encouragement. To my husband and parents, thank you for your endless love and for believing in me.

ABBREVIATIONS

AUC	Area Under the Curve
Cas9	CRISPR-associated protein 9
CLIP-seq	Crosslinking and Immunoprecipitation sequencing
CRISPR	Clustered regularly interspaced palindromic repeats
DepMap	Dependency Map
DL	Deep learning
eCLIP	enhanced crosslinking and immunoprecipitation
ENCODE	Enhanced Crosslinking and Immunoprecipitation
ESS	Effective sample size
FN	False negative
FP	False positive
FPR	False Positive Rate
GI	Gene interaction
GO	Gene ontology
HITS-CLIP	High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation
iCLIP	individual-nucleotide resolution cross-linking and immunoprecipitation
Lasso	Least Absolute Shrinkage and Selection Operator for classification
lincRNAs	long intergenic noncoding RNA
lncRNAs	long non-coding RNAs
MCMC	Markov Chain Monte Carlo
miRNAs	MicroRNAs
ML	Machine learning
mRNA	Messenger RNA
n_eff	Effective sample size
NIR	No information rate
OOB	Out-of-Bag
PAM	protospacer adjacent motif
POSTAR	Post-transcriptional Regulation Database

RBDs	RNA binding domains
RBPs	RNA binding proteins
RBPDB	RNA-Binding Protein Database
RBS	Ribosome binding sites
RF	Random forest
RIP-seq	RNA Immunoprecipitation sequencing
RNAi	RNA interference
RNPs	Ribonucleoprotein
ROC	Receiver Operating Characteristic
RRM	RNA recognition motifs
scaRNA	small Cajal body-specific RNA
sd	Standard deviation
se	Standard error
sgRNA	single-guide RNA
shRNA	small hairpin RNA
siRNA	Small interfering RNA
snoRNAs	small nucleolar RNA
TN	True negative
TP	True positive
TPR	True Positive Rate
X_{bin}	Binary variable
X_{cont}	Continuous variable
Z_i	Latent score

CONTENTS