Improving BacTermFinder: extending its field of vision and identifying type of terminators predicted

by

Mirza Baig

A thesis submitted to the

Department of Computer Science

to fulfill the requirements for the degree of

Bachelor of Science (Honour's).

Memorial University of Newfoundland

St. John's, NL, Canada

August 1, 2025

St. John's Newfoundland

Abstract

Transcription termination is a crucial phase of gene expression in bacteria, where the emerging transcript and the RNA polymerase are released from each other and from the DNA template. Transcription termination occurs through two mechanisms: intrinsic and factor-dependent terminators. Intrinsic terminators function independently of external proteins, relying instead on specific signals encoded in the DNA template and emerging transcript. In contrast, factor-dependent terminators require the Rho protein, which binds ribosome-free RNA at designated Rho utilization sites (RUT). BacTermFinder, a recently published computational tool, effectively predicts transcription terminators; however, it has two major limitations, first its training data mostly exclude the RUT region and second it is unable to classify terminators by their type. To address these limitations, this study presents an approach involving retraining BacTermFinder by extending its input sequences to 200 nucleotides, thus improving recall over a set of validation data. Additionally, we introduce a tree based classifier capable of distinguishing between intrinsic and factor-dependent terminators based on nucleotide-derived features. This classification model provides deeper insights into bacterial gene regulatory mechanisms and contributes to the understanding of transcription termination complexity.

Acknowledgements

I would like to thank my supervisor, Dr Lourdes, for their valuable advice and guidance throughout this research and helping to make this dissertation possible. I would also like to thank Amin for his guidance in understanding BacTermFinder and for advice and tips in using and retraining the model.

Contents

List of Figures

List of Tables

List of Abbreviations

AUPRC Area under the PR Curve

AUROC Area under the ROC Curve

BP Base Pair

BTF BacTermFinder

CNN Convolutional Neural Network

GB Gradient boosting

NT Nucleotide

NTs Nucleotides

RF Random Forest

ROI Region of Interest

RTP RhoTermPredict

RUT Rho Utilization Site

STD Standard Deviation

SVM Support Vector Machine

TTHP TransTermHP

TTS Transcription Termination Site

Chapter 1

Introduction

Transcription is the first step in gene expression and consists of three stages, initiation, elongation, and termination. Transcription termination is a critical part of gene expression that ensures proper gene function and prevents unnecessary RNA synthesis. The site where the transcription from DNA to RNA ends is called the Transcription Termination Site (TTS). The transcription ends due to specific regions in the DNA called terminators. There are two types of terminators: intrinsic (or Rho-independent) and factor-dependent (or Rho-dependent).

Intrinsic terminators are terminators that do not require the presence of external proteins to end transcription. Instead, they depend on a specific sequence. A canonical intrinsic terminator sequence is composed of GC-rich inverted repeat sequences or dyad symmetry elements followed by an oligo (T) sequence ("T stretch"). This repeat is important for the formation of a stable RNA hairpin structure which is essential

for termination. The transcribed RNA contains a stable hairpin followed by 7 to 9 U residues ("U stretch")[3].

One of the main differences between intrinsic and factor/Rho-dependent terminators is that the latter requires the Rho protein which binds to ribosome-free mRNA to terminate. Rho is a homohexameric ring protein that binds to the nascent RNA transcript[4]. Another key factor for factor-dependent terminators is the presence of a Rho utilization site (RUT). In some bacteria, Rho binds preferentially to unstructured and ribosome-free C-rich and G-poor nascent RNA, of at least 70–80 nucleotides, with regularly spaced cytosines, this site is known as the Rho utilization site. Depletion of G within a natural RUT site minimizes the formation of potentially interfering secondary structures, which generally inhibit Rho binding[2].

BacTermFinder[1] is a CNN ensemble model that can predict the location of TTS in various bacterial species and achieves better results than its predecessors such as TransTermHP[5], TermNN[6], and ITerm-PseKNC[7]. However, due to its region of interest (ROI) of 100 nucleotides, it cannot use features extracted from the Rho Utilization Site (RUT) to inform its predictions in identifying factor dependent termination. BacTermFinder is also unable to classify the predicted terminators as intrinsic or factor dependent. In fact, we could not find any tool in the literature which classifies terminators into intrinsic and factor-dependent.

Classifying terminators is important in gaining a deeper understanding of bacterial gene regulation, but it remains challenging due to their structural and functional

diversity[4, 8]. While these terminators show heterogeneity across different bacterial species and genomic contexts, certain features and patterns exist specifically for each type which can be useful in developing a classification model. Such terminator classification can provide further insight into their mechanisms of action.

To address the limitations of the current BacTermFinder model and the challenges of transcription terminator classification, we proposed the following steps:

1. **Retraining BacTermFinder**

   Expanding input sequences from 100 to 200 nucleotides and evaluating the effect of this change in BacTermFinder's predictive performance.

2. **Classifying Terminators**

   Developing a classifier to distinguish intrinsic terminators from factor-dependent terminators using features derived from nucleotide sequences.

The organization of this thesis is as follows, Chapter 2 examines prior work on computational prediction of transcription terminators. Chapter 3 details the methods we developed to overcome the limitations discussed above. Chapter 4 reports and interprets our results. Chapter 5 concludes by reflecting on the remaining limitations, and outlining future directions.

Chapter 2

Related Works

In this chapter we are going to discuss previous works related to prediction and classification of transcriptional terminators. To identify these previous works we searched for articles in multiple sources such as NCBI, Google Scholar, Connected Papers and other tools to find articles related to prediction and classification of terminators.

2.1 Predicting Terminators

Prediction of bacterial transcription terminators has progressed from motif-driven heuristics to machine learning and recently deep learning methods. Early intrinsic terminator tools such as TransTermHP[9] and RNIE[10] relied on stem-loop patterns and covariance models. Subsequent methods such as iTerm-PseKNC[7], iterb-PPse[11] and the Convolutional Neural Network (CNN) driven TermNN[6] introduced techniques such as k-mer frequencies, physiochemical descriptors and classifiers such as

support vector machine (SVM) and random forests. Tools such as RhoTermPredict[2], OPLS-DA[12], PASIFIC[13] and InterPin[14] have been taken into consideration by BacTermFinder[1]. Two recent neural models which are AMter[15] and BATTER[16] use transformer architecture to capture long range sequence dependencies.

2.1.1 BacTermFinder

BacTermFinder[1] is an ensemble model that combines different models to produce a result. It is a Convolutional Neural Network (CNN) model that combines different encoding techniques such as:

PS2 (Position-Specific 2-mer Composition)

Represents each sequence by the frequency of all 16 possible adjacent nucleotide pairs, capturing position-specific 2-mer patterns.

ENAC (Enhanced Nucleic Acid Composition)

Uses a sliding window over the sequence and counts the occurrence of each nucleotide within each window, concatenating these counts to form a feature vector that captures local compositional variation.

One-Hot Encoding

Maps each base to a 4-dimensional binary vector, so a sequence of length L becomes an L×4 binary matrix preserving positional information.

NCP (Nucleotide Chemical Property)

Encodes each nucleotide as a 3-dimensional vector reflecting intrinsic chemical attributes, yielding a 3×L matrix that embeds biochemical properties.

It combines these encoding methods to give an average result. BacTermFinder uses roughly 41 000 bacterial terminators (intrinsic and factor-dependent) of 22 species with varying GC-content (from 28% to 71%) from published studies that used RNA-seq technologies. BacTermFinder is evaluated on terminators of five bacterial species (not used for training BacTermFinder) and two archaeal species.

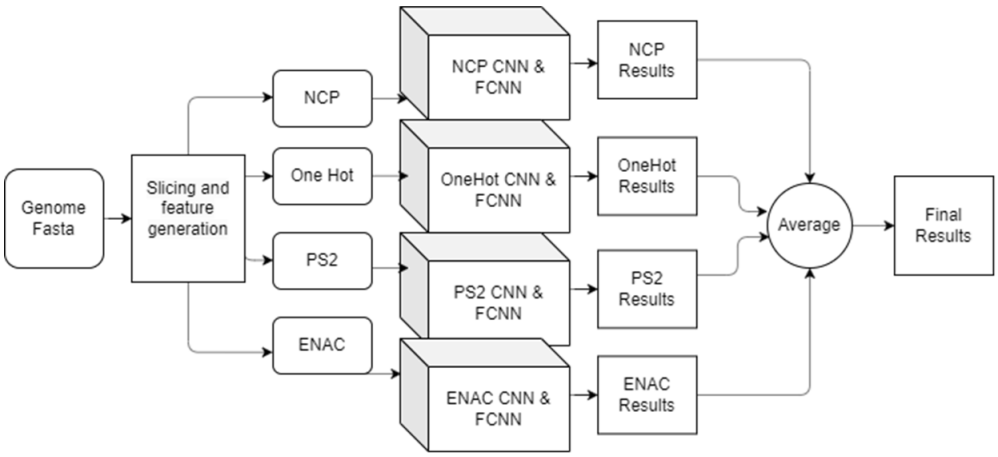


Figure 2.1: BacTermFinder architecture[1] (CC-BY-NC).

BacTermFinder outperforms in terms of recall other existing tools in an independent validation data set. BacTermFinder's average recall over five bacterial species is 0.57 ± 0.21, and TermNN's recall is 0.46 ± 0.23. This increase in recall is achieved by BacTermFinder while predicting, on average, two terminators less per gene than TermNN. BacTermFinder recall in all prokaryotic data (bacterial and archaeal) is

0.53 ± 0.20, and TermNN's is 0.40 ± 0.22.

2.1.2 TermNN

TermNN[6] is a tool to identify intrinsic transcription terminators in bacterial genomes. It uses a pre-training approach to implement the thermodynamic model for RNA folding into a Deep Learning framework, through which sequence and structure motif of intrinsic terminators are introduced. The inverse folding technique in DNA involves computationally predicting the nucleotide sequence that, when folded, forms a desired three-dimensional structure or shape.

2.1.3 ITT prediction

In a genome-wide survey of intrinsic transcription terminators (ITTs), the authors examined inter-operon regions in 13 phylogenetically diverse bacterial genomes for which high quality public RNA-seq data exist[17]. By scanning downstream of stop codons for single and paired RNA hairpins. They identified locations and RNA-seq-derived locations overlap with an accuracy of 72%, with 98% of sites being located ≤ 80 bases downstream of the translational stop codon.

2.1.4 iterb-PPse

iterb-PPse[11] incorporates 47 nucleotide properties into PseKNC-I/II[18], it utilizes extreme gradient boosting to predict terminators based on *Escherichia coli* and *Bacil-*

lus subtilis. It employs three new feature extraction methods K-pwm, Base-content, Nucleotidepro to represent sequences. The two-step method is applied to selected features.

2.1.5 iTerm-PseKNC

iTerm-PseKNC[7] is a support vector machine (SVM) based model designed to identify the transcription terminators using pseudo-k-tuple nucleotide composition (PseKNC) as features. PseKNC is a feature-generation technique. Their training data consisted of 280 terminator and 560 non-terminator sequences from *E. coli*.

2.1.6 RhoTermPredict

RhoTermPredict[2] specializes in predicting Rho-dependent termination sites through the identification of Rho utilization (RUT) sites. It does so by detecting various factors such as GC ratio. This study was conducted across three genomes *Escherichia coli K-12*, *Bacillus subtilis 168* and *Salmonella enterica LT2*.

2.1.7 OPLS - DA

In this article the authors used a prediction method based on Orthogonal Projections to Latent Structures Discriminant Analysis [OPLS-DA] of a large set of in vitro termination data[12]. Using previously uncharacterized genomic sequences for biochemical evaluation and OPLS-DA new factor-dependent signals and quantitative sequence

descriptors with significant predictive value were identified.

2.1.8 PASIFIC

PASIFIC[13] (Prediction of Alternative Structures for the Identification of Cis-regulation) is a machine-learning based approach, which, given a 5 UTR of a gene, predicts whether it can form the two alternative structures typical to riboregulators employing conditional termination. It uses a large positive training set of riboregulators which are derived from 89 human microbiome bacteria.

2.1.9 RNIE

RNIE[10] is a probabilistic approach for predicting RIT (Rho Independent Terminators). The method is based upon covariance models. It focuses on intrinsic terminators by putting emphasis on the stem loop and the poly U stretch. It uses 485 terminators from *E.coli* and *B.subtilis*.

2.1.10 TransTermHP

TransTermHP[9] is a dynamic programming model that is designed to detect terminators based on classic intrinsic terminator motif which is a hairpin stem followed by a poly-U tail. It predicts the location of terminators in 343 prokaryotic genomes. In *Bacillus subtilis*, it can detect 93% of known terminators with a false positive rate of just 6%.

2.1.11 AMter

AMter[15] differs from other models by using an end to end model for the prediction of transcription terminators based on the attention model. It uses two main attention methods that are the Frequency Attention model which focuses on the relative importance of different k-mer frequency features and Allkmer-Attention model. Frequency-Attention is designed to identify and extract frequency features, which are essential for understanding the significance of DNA sequences. The authors aggregate the representations of these informative frequency features to form an FA vector.

The Allkmer-Attention model model recasts the entire DNA sequence as a stream of k-mer indices and applies a self-attention mechanism with Query = Key = Value, which allows the model to uncover relations between any two k-mers, regardless of separation. This removes window size restrictions and helps learning of both local and global correlations.

The features of Frequency-Attention and Allkmer-Attention are concatenated along the feature dimension to create a Fully Concatenated Features (FCF) which is used as an input for a Multi-Layer Perceptron (MLP).

The training and validation dataset comprises of 280 Terminators sequences and 560 non-terminators sequences from iTerm-PseKNC[7]. It also uses 147 Terminators from *Escherichia coli* and 452 Terminators from *Bacillus subtilis*.

2.1.12 BATTER

BATTER (Bacteria Transcript Three prime End Recognizer)[16] is a context-sensitive transformer-based BERT-CRF (Bidirectional Encoder Representations from Transformers-Conditional Random Field) model to predict both intrinsic terminators and factor-dependent terminators. BERT-CRF is a combination of two powerful components for sequence labelling. BATTER is specifically designed for metagenome-scale applications across diverse bacterial species. It utilizes data augmentation techniques from high throughput 3' end mapping data in 17 bacteria species, and a large collection of 42,905 species-level representative bacteria genomes.

2.2 Classifying Terminators

After exploring various sources we have not found any paper that discusses a model to classify terminators into Intrinsic or factor dependent terminators. However, a relevant resource was found through a GitHub repository, PredictTerm[19], that addresses the classification of terminators.

2.2.1 PredictTerm

PredictTerm uses a Random Forest for its classification task, by training on both known processed ends and gene termini extracts to score factor-dependent terminators (RDT) and intrinsic terminators (IT) signals in sliding windows and then classifies

each candidate as intrinsic, factor dependent, both or unclassified. It uses two random forest classifiers one for each RDT and IT and then computes two scores per window using each of the random forests. It then uses the scores to classify the terminators into IT, RDT, both IT and RDT and unclassified.

2.3 Summary

We catalogued both the techniques examined by BacTermFinder and the more recent predictors that have since emerged (Table 2.1). Among all the sources examined, PredictTerm is the only available tool that does terminator-type classification.

Table 2.1: CM is covariance model, DA = discriminant analysis, DL = deep learning, DP = Dynamic programming, ML = machine learning, and SM = Structure matching

Method	Year	Terminator type	Method	# terminators	# species
BacTermFinder[1]	2024	Both	DL	41000	22
AMter[15]	2024	Both	DL	879	2
BATTER[16]	2023	Both	DL	42,905	17
InterPin[14]	2023	Intrinsic	SM	N/A	N/A
TermNN[6]	2022	Intrinsic	DL	1175	2
ITT pred[17]	2021	Intrinsic	Statistical	137	1
iterb-PPse[11]	2020	Both	ML	928	2
iTerm-PseKNC[7]	2019	Both	ML	852	1
RhoTermPredict[2]	2019	Factor-dep.	DP	1298	3
OPLS-DA[12]	2018	Factor-dep.	DA	104	2
PASIFIC[13]	2017	Intrinsic	SM	330	89
RNIE[10]	2011	Intrinsic	CM	1062	2
TransTermHP[9]	2007	Intrinsic	DP	N/A	N/A

Chapter 3

Methodology

In this section, we describe the methodology adopted to retrain the model, classify terminators, and make the tool accessible to users. We initially describe the steps taken for data collection and the different ways of encoding that data. We also describe ways of evaluating the data.

3.1 Retraining the model

We retrained the existing BacTermFinder[1] model on longer sequences to include the sequence containing the Rho binding site. The dataset present in the BacTermFinder was used, which contained roughly 41000 bacterial terminators from 21 genomes (Table 3.1).

Species	**Strain**	**Genome ID**	**GC%**	**Terminators**
Bacillus subtilis	168	AL009126.3	39.50	1715
Bacillus subtilis	168	NC_000964.3	40.63	3285
Caulobacter vibrioides	NA1000	CP001340.1	64.12	341
Clostridioides difficile	630	CP010905.2	23.59	1646
Dickeya dadantii	3937	NC_014500.1	52.92	1786
Escherichia coli	K12-BW25113	CP009273.1	47.34	1095
Escherichia coli	K12-MG1655	NC_000913.3	48.85	3191
Pseudomonas aeruginosa	PAO1	NC_002516.2	63.41	805
Staphylococcus aureus	JKD6009	LR027876.1	29.40	978
Staphylococcus aureus	NCTC8325	NC_007795.1	30.00	566
Streptococcus pneumoniae	D39V	CP027540.1	35.40	747
Streptococcus pneumoniae	TIGR4	NC_003028.3	36.85	1810
Streptomyces avermitilis	MA-4680	BA000030.4	68.38	1838
Streptomyces clavuligerus	ATCC27064(chr)	CP027858.1	70.30	1374
Streptomyces coelicolor	M145	NC_003888.3	70.62	1308
Streptomyces griseus	NBRC13350	NC_010572.1	71.15	2302
Streptomyces lividans	TK24	CP009124.1	69.54	1849
Streptomyces tsukubaensis	NBRC108819	CP020700.1	69.72	1283
Synechococcus elongatus	PCC7942	CP000100.1	53.47	1431
Vibrio natriegens	ATCC14048	CP009977.1	40.79	905
Vibrio natriegens	ATCC14048	CP009978.1	40.69	257
Vibrio parahaemolyticus	RIMD2210633	NC_004603.1	42.65	1852
Zymomonas mobilis	ATCC31821(ZM4)	CP023715.1	45.41	2040

Table 3.1: Genomes and Corresponding GC Content used for training

3.1.1 Data Processing

We used the scripts provided in BacTermFinder and modified it to accommodate 200 nts. The sequences were expanded by 50 nts on either side using the `bedtools slop` function in BEDTools[20], resulting in a final length of 200 nts for the Region of interest (ROI)[1]. After expansion the sequences were sorted and merged using `bedtools sort` and `bedtools merge` functions respectively. Later, we obtained the fasta files using the `bedtools getfasta` command.

3.1.2 Visualizing the RUT region

To emphasize the importance of extending to 200bp, we visualized such RUT regions in certain genomes, by summarizing the nucleotide composition of every residue within a 200bp window showing each candidate site. For every position in the 200bp we determined the proportion of sequences that contained a potential RUT region, that is a region with a presence of atleast 40% C, and at most 10% G, as RUT regions are rich in C and poor in G. We then visualized these regions by calculating the relative nucleotide frequency for a specific position which was done by dividing the total count of each nucleotide in that position by the number of terminators in that set.

3.1.3 Negative-Sample Generation

Negative controls were sampled by randomly selecting genomic regions that excluded annotated genes using `bedtools shuffle`. To reduce the false-positive rate in down-

stream genome scans, we used a 1:10 ratio of positive to negative to train the model. This sampling scheme reflects the class imbalance in bacterial genomes, experimentally defined terminator sites comprise only a tiny fraction of all possible length-matched genomic sequences[1].

3.1.4 Feature Encoding

The sequences present are in ACTG string format, to make it suitable for machine learning models various encoding methods were used such as ENAC, PS2, NCP and binary (one hot encoding). We used iLearnPlus[21] to generate these features which were later fed into a CNN model separately.

3.1.5 Model Description

We used the same training model as in BacTermFinder that is a CNN ensemble model and train the model on different encodings such as ENAC, PS2, NCP and binary. Finally, we averaged the results from the four CNNs to get a probability of the given sequence being a terminator.

BacTermFinder's convolutional neural network begins with a one-dimensional convolutional layer, `Conv1D` with 64 filters, a kernel size of 10, and a PReLU activation, followed by an `AveragePooling1D` layer with pool size 2. This `Conv1D`–`AveragePooling1D` block is repeated once more. A `BatchNormalization` layer normalizes the activations before passing them into a third `Conv1D` block (64 filters, kernel size 10, PReLU).

To reduce overfitting, a `Dropout` layer with rate 0.1 is applied next. The feature maps are then flattened (`Flatten`) and fed through a stack of fully connected layers (`Dense`) with PReLU activations, first 500 units (with 0.3 dropout), then 600 units (0.3 dropout), another 600 units (0.3 dropout), followed by two layers of 200 units each (0.4 dropout). Finally, a single-unit `Dense` layer with a Sigmoid activation produces the binary terminator/non-terminator prediction[1] (Figure 2.1).

The output of BacTermFinder is the average of the output of the four single-encoding CNNs. BacTermFinder output is a numerical score indicating the probability of a given sequence being a terminator. We retrained our model on 10 epochs and saw that the validation loss stagnated after 6 epochs (Figure 3.1).

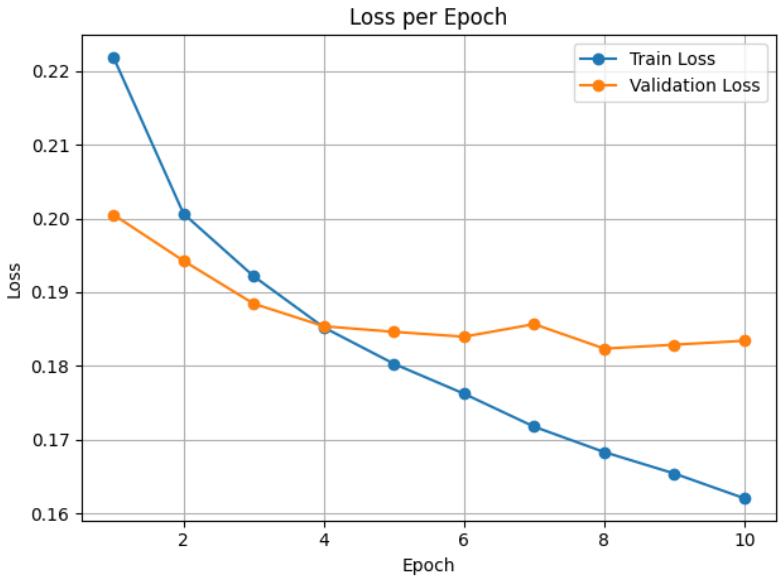


Figure 3.1: Training loss and Validation loss plotted against Epochs

3.1.6 Model Assessment

BacTermFinder with a ROI of 200 nts was compared against the previous BacTermFinder with a ROI of 100 nts on five genomes not used for training (Table 3.2) the models and on 2 archaeal terminators (Table 3.3).

To evaluate the performance of the updated BacTermFinder, we did comparative assessment and tested recall across a set of genomes. We tested recall as the experimentally identified terminators by sequencing methods in our validation data are not exhaustive, absence of a experimentally-detected terminator in a specific region does not prove that there is not a terminator in that region[1].

We compared results against existing terminator prediction tools and with the previous version of BacTermFinder to evaluate the results.

Species	**Strain**	**Genome Accession**	**GC (%)**	**# of Term**
Mycobacterium tuberculosis	H37Rv	AL123456.3	65.60	2070
Streptococcus agalactiae	NEM316	NC_004368.1	30.54	627
Streptomyces venezuelae	ATCC 15439	CP059991.1	68.80	786
Synechocystis	PCC 6803	CP054306.1	46.22	323
Synechocystis	PCC 7338	NC_000911.1	46.19	499

Table 3.2: Genomes with GC Content and Number of Terminators used for comparative assessment

Species	Strain	Genome Accession	GC (%)	# of Term
Haloferaxvolcanii	DS2	NC_013967.1	65.60	966
Methanococcus maripaludis	S2	NC_00579	30.54	965

Table 3.3: Some archaeal Terminators that were not used for training but were used for comparative assessment

3.2 Classification of Terminators

We developed a tree-based classification model to distinguish between intrinsic (Rho-independent) and factor-dependent (Rho-dependent) transcription terminators. To achieve this, we employed Random Forest (RF)[22] and Gradient Boosting Classifier (GBC)[23], two widely-used ensemble learning techniques known for their robustness, interpretability, and strong predictive performance with genomic sequence data.

3.2.1 Data Collection

We collected Rho-dependent terminators by running each genomes in the BacTermFinder[1] dataset through RhoTermPredict[2]. Similarly we used the intrinsic database INTERPIN[14] to collect intrinsic terminators. We found a total of 11 genomes in BacTermFinder that were present in INTERPIN and generated factor dependent terminator predictions for those genomes using RhoTermPredict. For *Escherichia coli* we used experimentally validated intrinsic and factor dependent terminators from the RegulonDB[24].

We initially obtained a total of 215785 factor dependent terminators from RhoTermPredict and a total of 27696 intrinsic terminators from INTERPIN and RegulonDB. To address this class imbalance we filtered the factor dependent terminators from RhoTermPredict to only consider those terminators with C>G ratio greater than 2. This resulted in a total of 89487 factor dependent terminators reducing the class imbalance (Table 3.4).

Specie–Strain	**Genome ID**	**# of IT**	**# of RDT**	**Total**
Bacillus subtilis - 168	NC_000964.3	2243	9851	12093
Escherichia coli - K-12 - MG1655	NC_000913.3	2383	37	2420
Pseudomonas aeruginosa - PAO1	NC_002516.2	2944	13831	16775
Staphylococcus aureus - NCTC 8325	NC_007795.1	1581	4225	5806
Streptococcus agalactiae - NEM316	NC_004368.1	1066	3870	4936
Streptococcus pneumoniae - D39V	CP027540.1	1038	4477	5515
Streptococcus pneumoniae - TIGR4 (ATCC BAA-334)	NC_003028.3	1146	4762	5908
Streptomyces clavuligerus - ATCC 27064	CP027858.1	3936	14258	18194
Streptomyces griseus - NBRC 13350	NC_010572.1	4774	14410	19184
Streptomyces lividans - TK24	CP009124.1	4815	14330	19145
Vibrio natriegens - NBRC 15636 / ATCC 14048 / DSM 759	CP009977.1	1771	5436	7207

Table 3.4: Genomes used for classification of terminators

3.2.2 Formatting Data

The data obtained from both methods was initially in Comma Separated Values (CSV). Each record was converted and mapped into a Browser Extensible Data (BED) dropping unwanted columns and adding the required ones using python scripts, we extended the predicted sequences to 78bp for consistency. These genomic coordinates were later converted into FASTA sequences using BEDTools[20] `getfasta` command. This helped preserve positional information while producing nucleotide strings that are compatible with sequence-based feature encoders.

3.2.3 Feature Encoding

To capture the underlying biological signals in each terminator sequence, we transformed the FASTA files into five feature sets **ENAC**, **PS2**, **NCP**, **Binary**, and **2-mer encodings** before feeding them to the tree-based classifiers:

1. **Enhanced Nucleotide Composition (ENAC)**

   ENAC slides a fixed-width window along the sequence and records single-nucleotide frequencies at every window position. For a four-letter DNA alphabet this yields $4 \times$ window length features.

2. **Position-Specific 2-mer (PS2).**

   PS2 encodes the ordered dinucleotide at every position: a sequence of length L produces $L - 1$ slots, each represented by a one-hot vector over the 16 possible

2-mers. This high-dimensional representation retains exact positional information about neighbouring bases, allowing the models to exploit precise motifs such as CG islands or pause sites.

3. **Nucleotide Chemical Properties (NCP)**

   Instead of treating A,C,G,T as arbitrary symbols, NCP maps each base to a triad describing its ring structure, chemical functionality, and hydrogen-bond count. The sequence becomes an $L \times 3$ binary matrix reflecting fundamental physicochemical characteristics.

4. **Binary (one-hot) encoding.**

   The canonical one-hot scheme represents each position with four bits. It is model-agnostic and serves as a baseline that preserves the full sequence without engineered abstractions.

5. **2-mer composition ($k = 2$)**

   In 2-mer, we counted the occurrences of every dinucleotide across the entire fragment and normalise by sequence length. The resulting 16-dimensional vector summarises global tendencies such as GC bias or repetitive "TT" tracts that distinguish intrinsic from factor-dependent terminators.

Comparing these encodings (Table 4.6) allows us to figure out whether raw positional detail (PS2), local compositional context (ENAC), biochemical semantics (NCP), or simpler global statistics (2-mer, Binary) contribute most to predictive

performance.

3.2.4 Models Considered

We chose Random Forests as it construct numerous decision trees independently, aggregating their results through majority voting, thereby reducing over fitting and improving generalization. Gradient Boosting, iteratively builds a sequence of shallow decision trees, with each new tree focused specifically on correcting errors made by the previous ensemble.

3.2.5 Model Evaluation

To assess the predictive performance of the classification model, we employed several evaluation metrics, including accuracy, recall and precision. As we are doing multi-class classification, we included macro f1 to check the average of the results. Additionally, we analysed feature importance derived from the 2-mer encoding to identify dinucleotides influencing the model's predictions (Figure 4.6). Certain features that we looked for in the model were:

1. **Hairpin Structure**

   The presence of Hairpin is a strong indicator of the terminator being an intrinsic terminator. A hairpin typically contains a GC-rich stem of 5–9 nt that is closed by a short 3–5 nt loop and immediately followed by a 7–9 nt U-rich tail[25]. Hairpin stability is also an important factor in intrinsic termination. Although

the U stretch is important in certain bacteria[26] as it helps disrupt stable transcription elongation complexes, it is found that the U stretch is absent in a lot of other bacteria.

2. **Presence of RUT**

   The RUT tends to have a C/G >1 (Figure 3.2). The higher the ratio of C/G the stronger the likelihood of the terminator being a Factor dependent terminator[2]. A pattern showing high similarity with RUT may not be sufficient to determine the terminator type.

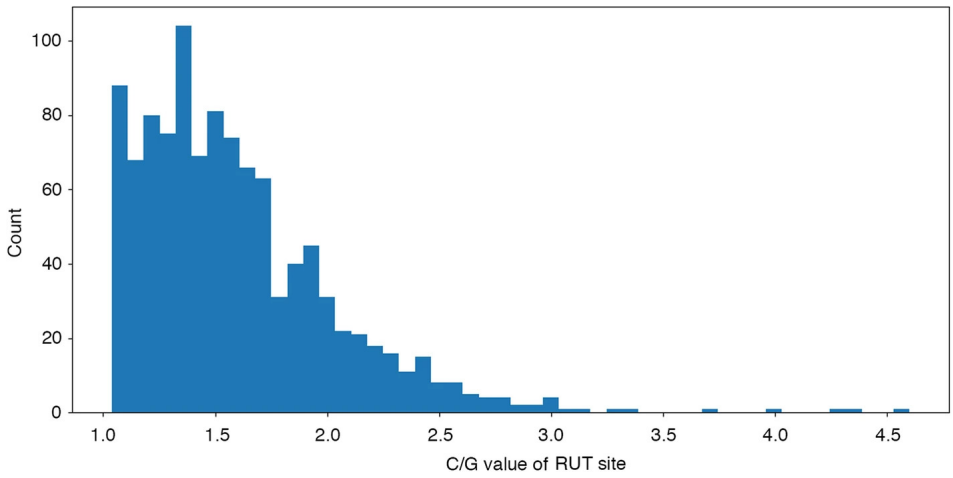


Figure 3.2: Distribution of C/G content of predicted terminators RUT sites[2] (CC-BY 4.0).

Chapter 4

Results and Discussion

In this section, we present and analyze the results of our data preprocessing, feature remodelling, and classification approach for bacterial transcription terminators. We compare the performance of our retrained model against the original BacTermFinder tool. We provide biological insight into the distribution of predicted Rho-independent and factor-dependent terminators across genomic contexts.

4.1 Retraining the Model

We observed the results of extending our region of interest on BacTermFinder to check whether it would produce better results.

4.1.1 Potential RUT Regions

To examine the potential presence of Rho utilization (RUT) sites, we applied a sliding window-based heuristic to each 200bp sequence, identifying windows of 20bp with cytosine content exceeding 40% and guanine content below 10% and a minimum C>G ratio of 1. These windows are indicative of C-rich, G-poor regions that may function as RUT sites. We visualized the location of these predicted regions across all sequences, with particular attention to those located outside the central 100bp region which is 50bp–150bp as we extended it by 50bp on either side. We found a total of 5534 terminators in the training set that contained potential RUT regions outside the central region (Table 4.2). In the validation set we found 753 terminators that contained similar regions (Table 4.1).

Species	**Strain**	**Genome Accession**	**GC (%)**	**# of Hits**
Mycobacterium tuberculosis	H37Rv	AL123456.3	65.60	268
Streptococcus agalactiae	NEM316	NC_004368.1	30.54	54
Streptomyces venezuelae	ATCC 15439	CP059991.1	68.80	116
Synechocystis	PCC 6803	CP054306.1	46.22	118
Synechocystis	PCC 7338	NC_000911.1	46.19	197

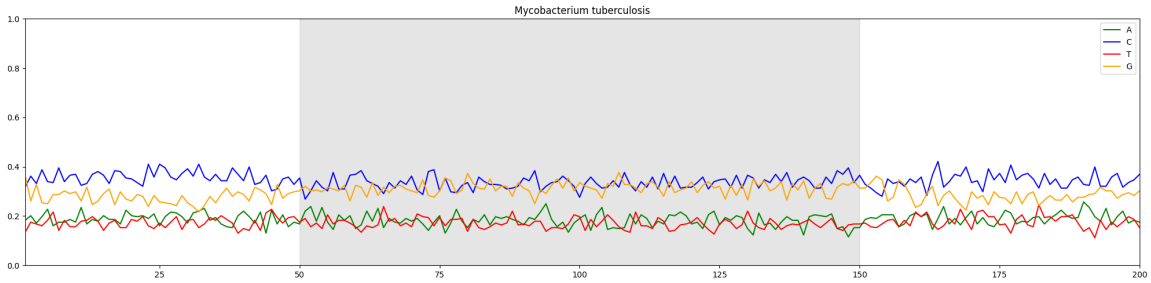
Table 4.1: Genomes with GC Content and Number of potential RUT sides outside of 50bp-150bp, used for comparative assessment

Species	**Strain**	**Genome ID**	**GC%**	**Hits**
Bacillus subtilis	168	AL009126.3	39.50	215
Bacillus subtilis	168	NC_000964.3	40.63	541
Caulobacter vibrioides	NA1000	CP001340.1	64.12	59
Clostridioides difficile	630	CP010905.2	23.59	96
Dickeya dadantii	3937	NC_014500.1	52.92	382
Escherichia coli	K12-BW25113	CP009273.1	47.34	255
Escherichia coli	K12-MG1655	NC_000913.3	48.85	726
Pseudomonas aeruginosa	PAO1	NC_002516.2	63.41	170
Staphylococcus aureus	JKD6009	LR027876.1	29.40	59
Staphylococcus aureus	NCTC8325	NC_007795.1	30.00	40
Streptococcus pneumoniae	D39V	CP027540.1	35.40	109
Streptococcus pneumoniae	TIGR4	NC_003028.3	36.85	269
Streptomyces avermitilis	MA-4680	BA000030.4	68.38	238
Streptomyces clavuligerus	ATCC27064(chr)	CP027858.1	70.30	187
Streptomyces coelicolor	M145	NC_003888.3	70.62	152
Streptomyces griseus	NBRC13350	NC_010572.1	71.15	301
Streptomyces lividans	TK24	CP009124.1	69.54	260
Streptomyces tsukubaensis	NBRC108819	CP020700.1	69.72	182
Synechococcus elongatus	PCC7942	CP000100.1	53.47	351
Vibrio natriegens	ATCC14048	CP009977.1	40.79	203
Vibrio natriegens	ATCC14048	CP009978.1	40.69	59
Vibrio parahaemolyticus	RIMD2210633	NC_004603.1	42.65	363
Zymomonas mobilis	ATCC31821(ZM4)	CP023715.1	45.41	317

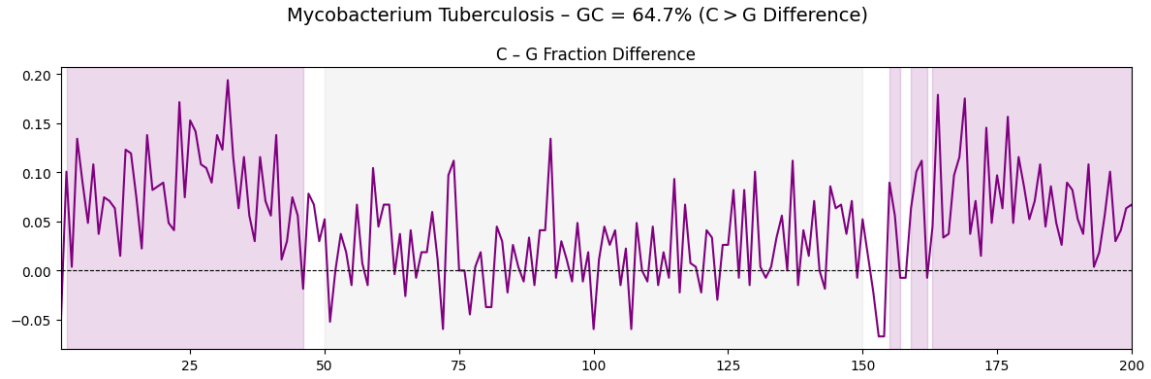
Table 4.2: Species with corresponding genome ID, GC content, and number of hits, that is the potential RUT sites.

4.1.2 Visualization of Terminator regions

To further understand the existence of RUT regions outside the central region, we computed the positional frequency of each nucleotide across sequences from each genome. We found that there was a rise of C over G in the regions outside the middle 50bp and 150bp. This rise was seen in *Mycobacterium tuberculosis* (Figure 4.1) and similarly in *Streptomyces venezuelae* (Figure 4.2).

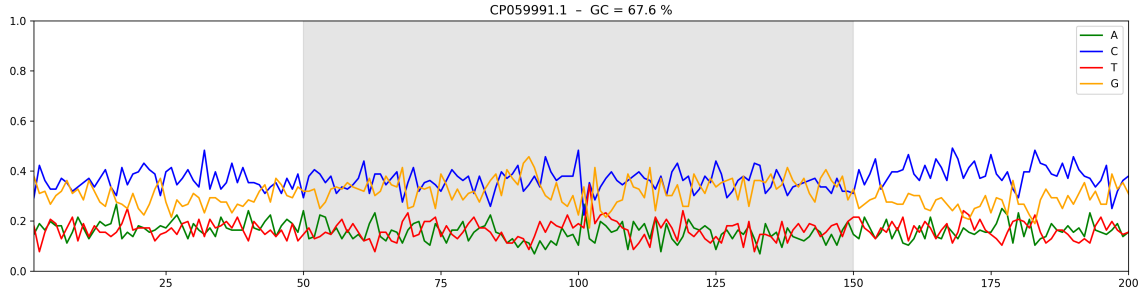


(a) Positional nucleotide frequency of *Mycobacterium tuberculosis*

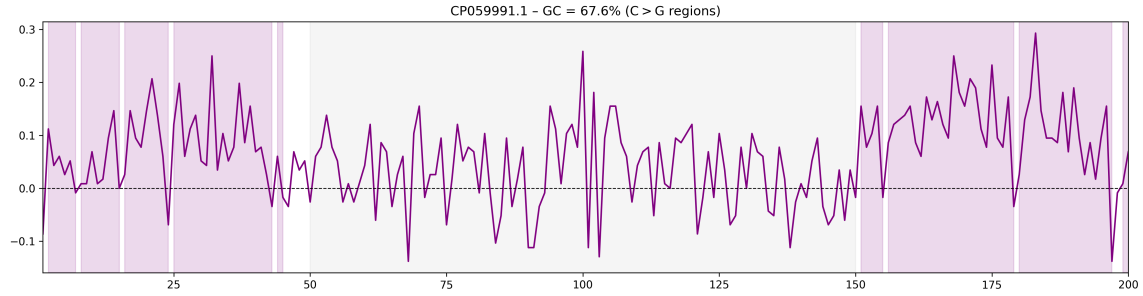


(b) Plot based on the difference of C and G to show more presence of C over G in the 200bp region

Figure 4.1: 268 *Mycobacterium Tuberculosis* terminators were used for this visualization



(a) Positional nucleotide frequency of *Streptomyces venezuelae*

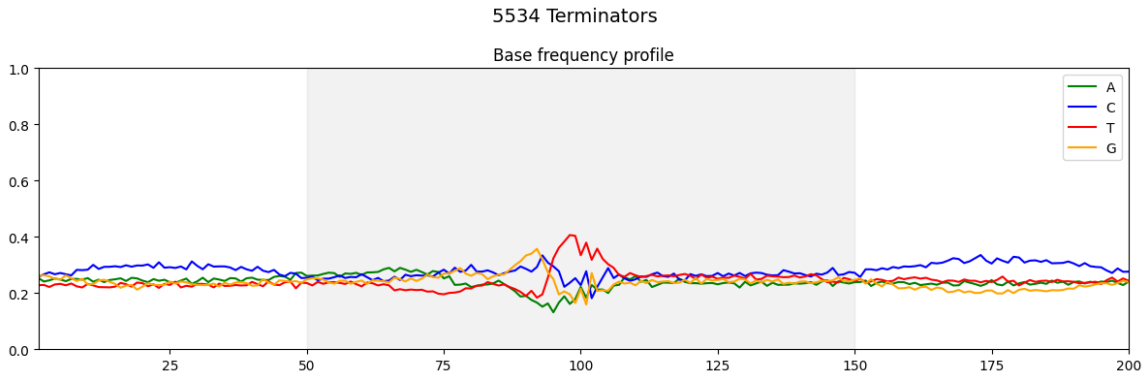


(b) Plot based on the difference of C and G to show more presence of C over G in the 200bp region

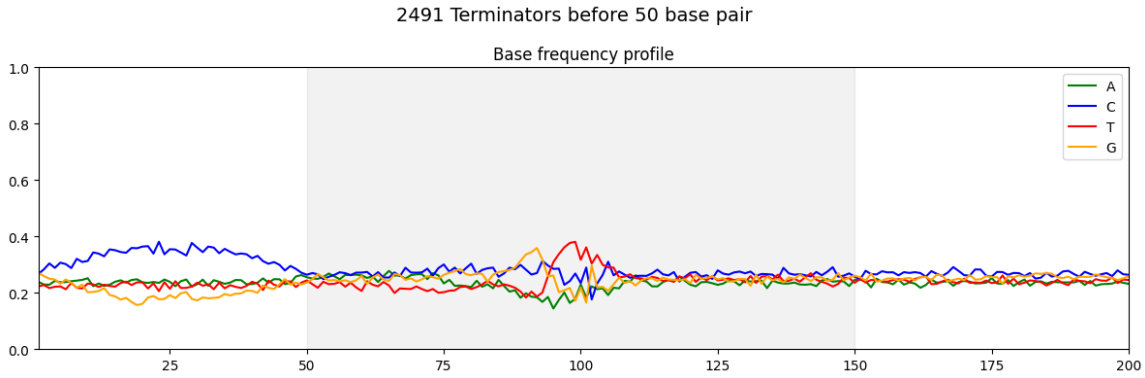
Figure 4.2: 116 *Streptomyces venezuelae* terminators were used for this visualization

To check this rise across all genomes, we took all the genomes in the training set and further filtered the regions to see which were occurring before the 50bp and which of the hits occurred after the 150bp. We saw patterns before and after the transcription terminators. We found a total of 2491 terminators that occurred before 50bp and 3330 that occurred after 150bp or 50bp after termination. The ones before termination are the potential RUT sites, we are unaware of the rise of Cytosine 50bp after termination. A paper does mention rise of Cytosine downstream of RNA terminated

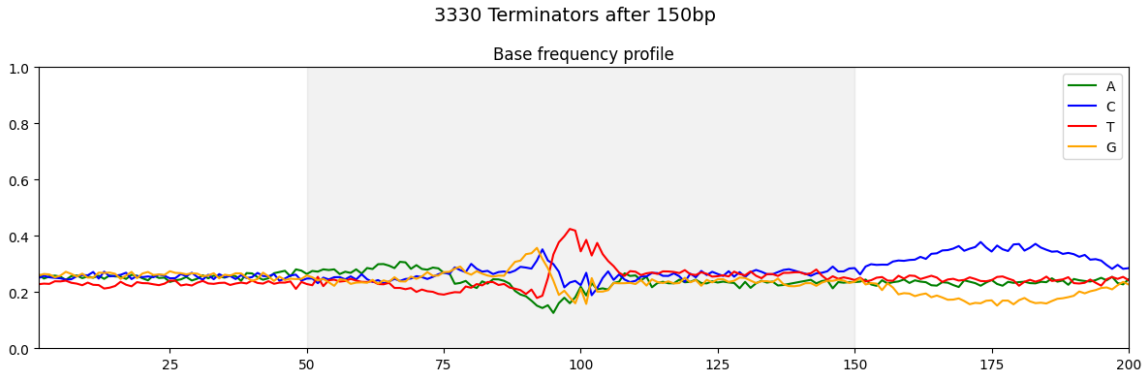
Rho dependent transcripts[27], but the rise we see is 50bp after termination.



(a) Combined frequency of the filtered terminators before 50bp and after 150bp



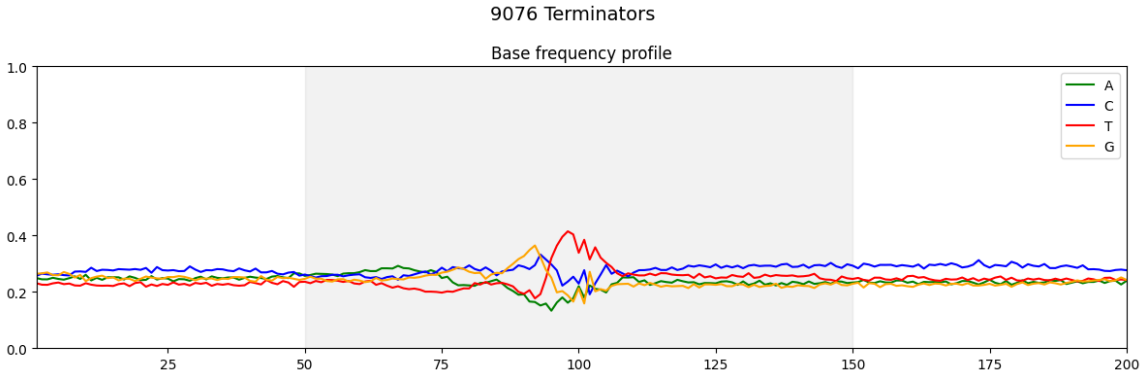
(b) 2491 Terminators showing Potential RUT region as before 50 base pairs



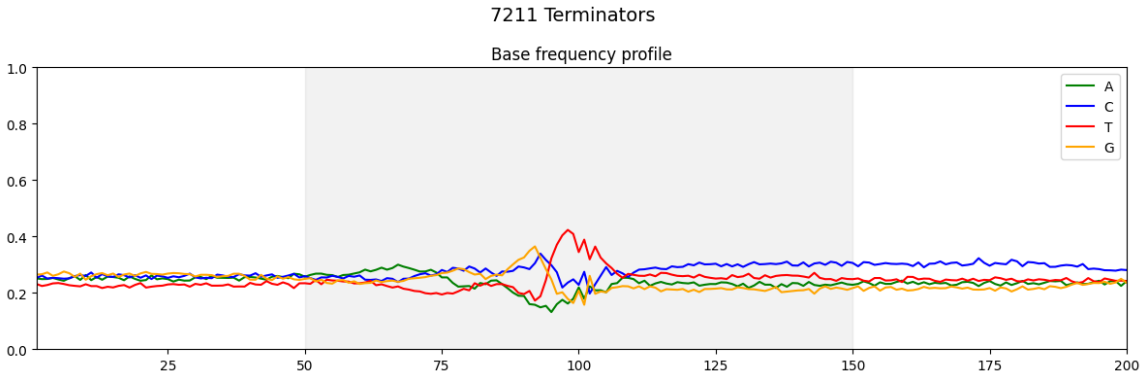
(c) 3330 Terminators showing Potential RUT region as after 150 base pairs

Figure 4.3: Filtering across all genome present in training dataset

To further check, we filtered sequences that show rise of Cytosine after 100bp and found around 7221 terminators that showed such presence which could be potentially in line with known literature[27]. Including rise in cytosine that occurs 50 bp before termination and immediately after termination we found a total of 9076 such terminators.



(a) Combined frequency of the filtered terminators before 50bp and after 100bp



(b) Filtering with cytosine spike immediately after 100bp

Figure 4.4: Filtering across all genome present in training dataset

4.1.3 Comparative Assessment

Here we present the recall performance of five bacterial genomes across four terminator prediction tools RhoTermPredict (RTP)[2], TransTermHP (TTHP)[9], BacTermFinder (BTF)[1], and the extended version of BacTermFinder using 200bp sequences in Table 4.3, and in Table 4.4 we mention the recall across archaeal terminators. Overall, BacTermFinder outperforms both RTP and TTHP in all cases, with the 200bp variant showing further improvements. Notably, for *Streptococcus agalactiae* NEM316, BacTermFinder (200bp) achieves the highest recall of 0.91 ± 0.23, demonstrating the effectiveness of longer sequence context. This increase in recall is mainly due to the observance of RUT structures that were present outside the previous 100 bp length (Figure 4.3).

We observed a similar declining trend when we move towards stricter thresholds as seen in BacTermFinder and TermNN, but the area under curve and the recall was higher than the other models. The decline in recall at higher thresholds reflects the increased strictness of matching criteria, and the area under each curve summarizes the overall strength of terminator predictions (Figure 4.5).

Bacterium	Genome (GC%)	RTP	TTHP	BTF	BTF (200bp)
Streptomyces gardneri ATCC 15439	CP059991.1 (70.7%)	0.25 ± 0.13	0.01 ± 0.004	0.60 ± 0.22	**0.72±0.21**
Mycobacterium tuberculosis H37Rv	AL123456.3 (64.7%)	0.24 ± 0.16	0.003 ± 0.002	0.23 ± 0.12	**0.39±0.21**
Synechocystis sp. PCC 7338	CP054306.1 (47.1%)	0.10 ± 0.04	0.28 ± 0.15	0.64 ± 0.22	**0.76±0.20**
Synechocystis sp. PCC 6803	NC_000911.1 (47.0%)	0.12 ± 0.08	0.22 ± 0.10	0.54 ± 0.18	**0.72±0.21**
Streptococcus agalactiae NEM316	NC_004368.1 (35.1%)	0.02 ± 0.006	0.56 ± 0.27	0.82 ± 0.30	**0.91±0.23**
Overall mean recall		0.15 ± 0.10	0.25 ± 0.20	0.57 ± 0.21	**0.70±0.21**

Table 4.3: Comparison of recall scores (± std) across prediction tools. The prediction tools are RhoTermPredict (RTP), TransTermHP (TTHP), BacTermFinder (BTF) and BacTermFinder with 200bp (BTF (200bp)). GC content and genome accession provided for each bacterium and the highest recall per Genome is highlighted

Bacterium	Genome (GC%)	TermNN	BacTermFinder	BacTermFinder (200bp)
Haloferax volcanii NC_013967 (65.7%)	Archaea	0.15 ± 0.11	0.32 ± 0.20	**0.55±0.25**
Methanococcus maripaludis NC_00579 (32.6%)	Archaea	0.40 ± 0.22	0.57 ± 0.25	**0.77±0.25**

Table 4.4: Comparison of recall scores (± std) for archaeal genomes using TermNN, BacTermFinder, and BacTermFinder (200bp). The highest recall per Archaea is highlighted

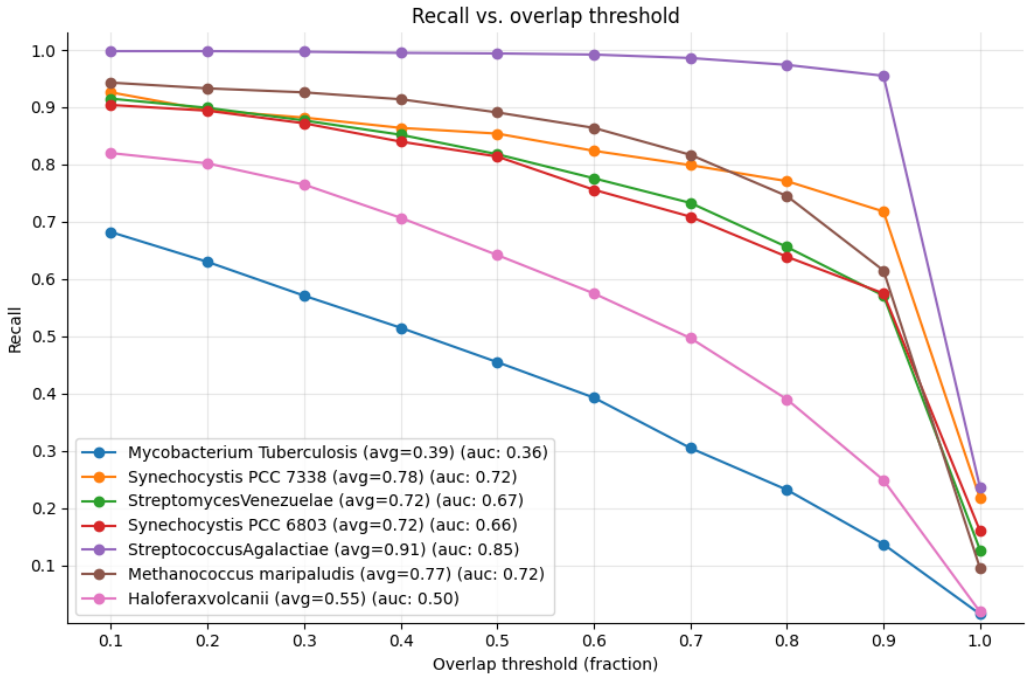


Figure 4.5: Recall as a function of minimum overlap threshold between predicted and known terminator regions. Curves are shown for each genome accession, with thresholds varying from 0.1 (10% overlap) to 1.0 (100% overlap).

4.1.4 Comparing with older BacTermFinder

We predicted and compared the results of the current extended BacTermFinder and the older BacTermFinder. We predicted the genome *Streptomyces gardneri* using both the models, the older model predicted 66201 terminators, while the newer model predicted 74110 terminators. Upon sorting and merging the terminators, it resulted in the older model having 37664 and the newer model having 20748 terminators, both were filtered with a threshold of score greater than 0.3. This along with higher recall

suggests that the model with extended nts is repeatedly predicting the same spots meaning it has greater positional precision. Table 4.5 shows the results across all genomes.

Genome accession	**BacTermFinder**	**BacTermFinder (200 bp)**	**Recall on 200bp**
AL123456.3	16 076	9 575	0.39 ± 0.21
CP054306.1	16 071	10 619	0.76 ± 0.20
CP059991.1	37 664	20 748	0.72 ± 0.21
NC_000911.1	15 099	9 995	0.72 ± 0.21
NC_004368.1	7 408	5 867	0.91 ± 0.23

Table 4.5: Non-redundant terminator counts after merging overlapping predictions for the original 100 bp model and the extended 200 bp model, together with the mean recall ($\pm$ standard deviation) achieved by the 200 bp BacTermFinder on each genome.

4.2 Classification of Terminators

To distinguish between factor-dependent and intrinsic terminators, we encoded our sequences using a variety of feature representations including 2-mer frequency vectors, Enhanced Nucleotide Composition (ENAC), Binary profiles, Position-Specific Scoring (PS2), and Nucleotide Chemical Properties (NCP). These encoding schemes were designed to capture both local and global sequence patterns relevant to terminator

functionality.

4.2.1 Model Results

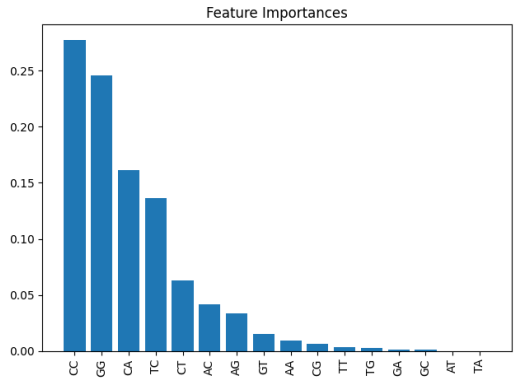
We looked at the results of different of models and how they compared with each other (Table 4.6). We analysed the results of 2-mer as they are more interpretable, they show the importance of each dinucleotide and its involvement in the prediction process (Figure 4.6). The classifier was evaluated using stratified 5-fold cross-validation to mitigate any class imbalance effects. The classifier with the best encoding was PS2 which obtained a f1 macro score of 0.853 using the Gradient Boosting Classifier.

4.2.1.1 Gradient Boosting Classifier

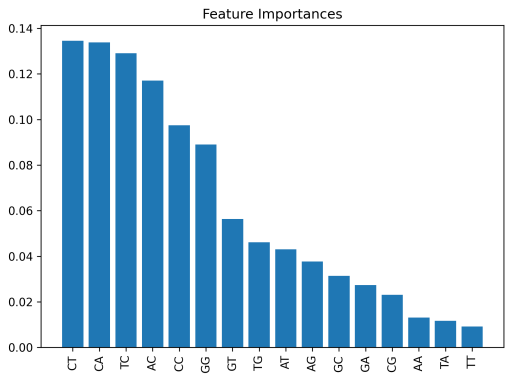
We employed a Gradient Boosting Classifier (GBC) using the various mentioned encoding methods and received various results. Majority of the models worked very well on factor dependent terminators and most worked decent on intrinsic terminators.

4.2.1.2 Feature Importance of 2 mer

The most important features the 2-mer models considered were dinucleotides such as CC and GG which are known to be significant in forming rho-dependent structures such as Rho Utilization Sites (RUT) (Figure 4.6a). RUT's are structures which are C rich and poor in G. Similarly intrinsic structures such as hairpin, that are crucial in initiating intrinsic termination, consist of a GC-rich dyad[28].



(a) Feature importance scores for 2-mer - GBC



(b) Feature importance scores for 2-mer - RF

Figure 4.6: Comparison of feature importance scores between (a) the Gradient Boosting Classifier on 2-mer encoding and (b) the Random Forest model.

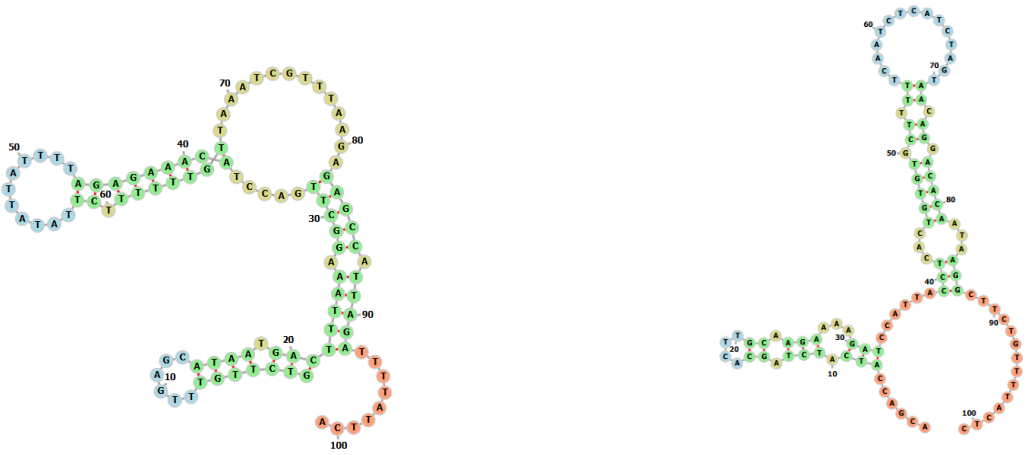
From the results in (Table 4.6), we can see that most of the models work well on factor dependent terminators, achieving high precision, recall and accuracy. But some of them fall short when it comes to correctly classifying intrinsic terminators which can be seen with lower F_1-macro scores. This is a result of data limitation as there were fewer predicted intrinsic terminators to train the model on.

Encoding	Model	Accuracy	Precision	Recall	F_1-score	F_1-macro
PS2	RF	0.8637	0.9200 ± 0.0019	0.9059 ± 0.0017	0.9129 ± 0.0014	0.7810 ± 0.0039
PS2	GBC	0.8573	0.8565 ± 0.0013	0.9779 ± 0.0006	0.9131 ± 0.0006	**0.8531 ± 0.0023**
ENAC	GBC	0.8993	0.8959 ± 0.0021	0.9830 ± 0.0013	0.9374 ± 0.0014	0.8503 ± 0.0021
ENAC	RF	0.8920	0.8930 ± 0.0021	0.9769 ± 0.0010	0.9331 ± 0.0016	0.7758 ± 0.0036
NCP	GBC	0.8672	0.8683 ± 0.0013	0.9766 ± 0.0011	0.9193 ± 0.0011	0.8049 ± 0.0040
NCP	RF	0.8430	0.8350 ± 0.0018	0.9960 ± 0.0006	0.9084 ± 0.0010	0.6790 ± 0.0036
Binary	GBC	0.8761	0.8773 ± 0.0018	0.9778 ± 0.0008	0.9248 ± 0.0010	0.8235 ± 0.0032
Binary	RF	0.8872	0.8799 ± 0.0013	0.9863 ± 0.0004	0.9301 ± 0.0008	0.7264 ± 0.0047
2-mer	RF	0.9484	0.8988 ± 0.0030	0.9538 ± 0.0008	0.9255 ± 0.0018	0.8264 ± 0.0030
2-mer	GBC	0.8869	0.8852 ± 0.0020	0.9763 ± 0.0005	0.9285 ± 0.0011	0.8177 ± 0.0034

Table 4.6: Performance of RF and GBC models across feature encodings. Metrics shown are accuracy, precision, recall, F_1-score, and F_1-macro (mean ± SD; 5-fold CV).

4.2.2 Visualizing classified sequences

We classified various sequences and visualized them using Vienna RNA[29]. Some of the sequence we chose were those terminators which were a result of BacTermFinder's predictions, we chose a low GC% *Streptococcus agalactiae* from the validation set to visualize, to emphasize the intrinsic terminators. In the structure on the left that we have visualized (Figure 4.7a), we can see a stretch of a poly T tail and GC stretches which gives us a strong indication of this terminator being an intrinsic terminator.



(a) Intrinsic terminator.

(b) Factor-dependent terminator

Figure 4.7: Comparison of RNA secondary structures for intrinsic and factor-dependent terminators predicted by BacTermFinder.

The structure on the right (Figure 4.7b) lacked the characteristic downstream T-rich region, instead it showed the extensive unstructured region representing the RUT indicative of a factor-dependent terminator.

Chapter 5

Conclusion

In this chapter we will conclude by summarizing this research and discuss the limitations of this project.

5.1 Contributions

The results show that the extension of BacTermFinder from 100bp to 200bp was useful in increasing recall and higher positional precision. The average recall of the previous BacTermFinder over the validation dataset was 0.57±0.21 and the newer extended version resulted in an average recall of 0.70±0.21. The average recall over bacterial and archaeal genomes previously was 0.53±0.19 and with the updated version was 0.68±0.22, observing an increase of roughly 15% recall. This recall was also higher in high GC bacteria than RhoTermPredict on our validation data such as *Mycobacterium tuberculosis*.

We also made a classification model based on 11 genomes which was able to classify intrinsic terminators by detecting the presence of hairpins and factor dependent terminators by detecting the presence of RUT. The weighted F_1 macro of our best model which was Gradient Boosting Classifier with PS2 encoding was 85%.

5.2 Limitations

When doing classification of terminators, we were limited with our data due to limited datasets of experimentally verified intrinsic and factor-dependent terminators so we had to rely on predicted sequences.

5.3 Future Work

Extend it asymmetrically upstream instead of symmetric on either side to include the RUT only instead of post terminator fluctuations present in the nucleotide frequency plot. That is instead of extending 50 bp on either side, we extend only before the termination.

Bibliography

[1] Seyed Mohammad Amin Taheri Ghahfarokhi and Lourdes Peña-Castillo. Bactermfinder: a comprehensive and general bacterial terminator finder using a CNN ensemble. *NAR genomics and bioinformatics*, 7(1):lqaf016, March 2025.

[2] M Di Salvo, S Puccio, C Peano, et al. Rhotermpredict: an algorithm for predicting rho-dependent transcription terminators based on *Escherichia coli, Bacillus subtilis* and *Salmonella enterica* databases. *BMC Bioinformatics*, 20(1):117, 2019.

[3] Ivan Gusarov and Evgeny Nudler. The mechanism of intrinsic transcription termination. *Molecular Cell*, 3(4):495–504, 1999.

[4] A Ray-Soni, MJ Bellecourt, and R Landick. Mechanisms of bacterial transcription termination: all good things must end. *Annual Review of Biochemistry*, 85:319–347, 2016.

[5] Carleton L. Kingsford, Kunmi Ayanbule, and Steven L. Salzberg. Rapid, accurate, computational discovery of rho-independent transcription terminators

illuminates their relationship to DNA uptake. *Genome Biology*, 8(2):R22, 2007.

[6] Vivian B. Brandenburg, Franz Narberhaus, and Axel Mosig. Inverse folding based pre-training for the reliable identification of intrinsic transcription terminators. *PLoS Computational Biology*, 18(7):e1010240, 2022.

[7] Chun-Quan Feng, Zuo-Yu Zhang, Xiao-Jun Zhu, Yu Lin, Wei Chen, Hao Tang, and Hao Lin. iterm-pseknc: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics*, 35(9):1469–1477, 2019.

[8] Zachary F. Mandell, Dani Zemba, and Paul Babitzke and. Factor-stimulated intrinsic termination: getting by with a little help from some friends. *Transcription*, 13(4-5):96–108, 2022. PMID: 36154805.

[9] Carleton L Kingsford, Kunmi Ayanbule, and Steven L Salzberg. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biology*, 8(2):R22, February 2007.

[10] P. P. Gardner, L. Barquist, A. Bateman, E. P. Nawrocki, and Z. Weinberg. RNIE: genome-wide prediction of bacterial intrinsic terminators. *Nucleic Acids Research*, 39(14):5845–5852, August 2011.

[11] Yongxian Fan, Wanru Wang, and Qingqi Zhu. iterb-PPse: Identification of transcriptional terminators in bacterial by incorporating nucleotide properties into PseKNC. *PLOS ONE*, 15(5):e0228479, May 2020.

[12] Cédric Nadiras, Eric Eveno, Annie Schwartz, Nara Figueroa-Bossi, and Marc Boudvillain. A multivariate prediction model for Rho-dependent termination of transcription. *Nucleic Acids Research*, 46(16):8245–8260, September 2018.

[13] Adi Millman, Daniel Dar, Maya Shamir, and Rotem Sorek. Computational prediction of regulatory, premature transcription termination in bacteria. *Nucleic Acids Research*, 45(2):886–893, January 2017.

[14] Swati Gupta, Namrata Padmashali, and Debnath Pal. Interpin: A repository for intrinsic transcription termination hairpins in bacteria. *Biochimie*, 214:228–236, 2023.

[15] Haotian Zhang, Jinzhe Li, Fang Hu, Haobo Lin, and Jiali Ma. An end-to-end model for transcriptional terminators prediction by extracting semantic feature automatically based on attention mechanism. *Concurrency and Computation: Practice and Experience*, 36(13):e8056, June 2024.

[16] Yunfan Jin, Hongli Ma, Zhenjiang Zech Xu, and Zhi John Lu. Batter: Accurate prediction of rho-dependent and rho-independent transcription terminators in metagenomes. *bioRxiv*, 2023.

[17] Swati Gupta and Debnath Pal. Clusters of hairpins induce intrinsic transcription termination in bacteria. *Scientific Reports*, 11(1):16194, August 2021.

[18] Wei Chen, Tian-Yu Lei, Dian-Chuan Jin, Hao Lin, and Kuo-Chen Chou. PseKNC: A flexible web server for generating pseudo K-tuple nucleotide composition. *Analytical Biochemistry*, 456:53–60, July 2014.

[19] David Chyou. PredictTerm: A tool for classifying bacterial transcription terminators. `https://github.com/davidchyou/PredictTerm`, n.d. Accessed: 2025-06-17.

[20] Aaron R. Quinlan and Ira M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010.

[21] Zhen Chen, Pei Zhao, Chen Li, Fuyi Li, Dongxu Xiang, Yong-Zi Chen, Tatsuya Akutsu, Roger J. Daly, Geoffrey I. Webb, Quanzhi Zhao, Lukasz Kurgan, and Jiangning Song. iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic acids research*, 49(10):e60, June 2021.

[22] Hasan Salman, Ali Kalakech, and Amani Steiti. Random forest algorithm overview. *Babylonian Journal of Machine Learning*, 2024:69–79, 06 2024.

[23] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, 12 2013.

[24] Heladia Salgado, Socorro Gama-Castro, Paloma Lara, Citlalli Mejia-Almonte, Gabriel Alarcón-Carranza, Andrés G López-Almazo, Felipe Betancourt-

Figueroa, Pablo Peña-Loredo, Shirley Alquicira-Hernández, Daniela Ledezma-Tejeida, Lizeth Arizmendi-Zagal, Francisco Mendez-Hernandez, Ana K Diaz-Gomez, Elizabeth Ochoa-Praxedis, Luis J Muñiz-Rascado, Jair S García-Sotelo, Fanny A Flores-Gallegos, Laura Gómez, César Bonavides-Martínez, Víctor M del Moral-Chávez, Alfredo J Hernández-Alvarez, Alberto Santos-Zavaleta, Salvador Capella-Gutierrez, Josep Lluis Gelpi, and Julio Collado-Vides. RegulonDB v12.0: a comprehensive resource of transcriptional regulation in *E. coli* K-12. *Nucleic Acids Research*, 52(D1):D255–D264, November 2023.

[25] Guillaume Cambray, Joao C. Guimaraes, Vivek K. Mutalik, Colin Lam, Quynh-Anh Mai, Tim Thimmaiah, James M. Carothers, Adam P. Arkin, and Drew Endy. Measurement and modeling of intrinsic transcription terminators. *Nucleic Acids Research*, 41(9):5139–5148, 03 2013.

[26] Agata Czyz, Rachel A. Mooney, Ala Iaconi, and Robert Landick. Mycobacterial RNA polymerase requires a u-tract at intrinsic terminators and is aided by nusg at suboptimal terminators. *mBio*, 5(2):10.1128/mbio.00931–14, 2014.

[27] Daniel Dar and Rotem Sorek. High-resolution RNA 3-ends mapping of bacterial Rho-dependent transcripts. *Nucleic Acids Research*, 46(13):6797–6805, July 2018.

[28] JM Peters, AD Vangeloff, and R Landick. Bacterial transcription terminators: the RNA 3'-end chronicles. *Journal of Molecular Biology*, 412(5):793–813, 2011.

[29] Ivo L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431, July 2003.