Characterisation of VapC in *Mycobacterium smegmatis*

A thesis

submitted in fulfilment

of the requirements for the Degree

of

Master of Science (Research)

at the

University of Waikato

by

Duncan Mark Koorey Willcock

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

University of Waikato

2016

Abstract

Members of the VapC family of proteins cleave RNA at specific sites in order to regulate biological processes with a cell. Characterization of the sites targeted by a specific protein using conventional biochemical techniques is resource intensive. This study explores the potential use of computational models to characterize the sites targeted by VapC in *Mycobacterium smegmatis*. Previous work has reported the impact of VapC upon each gene in the *M. smegmatis* genome and produced a hypothesis model for the specific motif targeted by the enzyme [1]. However, this model has been shown to be insufficient for the differentiation of sites cleaved by VapC from those not cleaved. This study aims to extend this model to accurately describe the features which influence VapC activity at a site. A model capable of accurately predicting the VapC target sites could supplement the existing biochemical techniques. Furthermore, a process developed to train such a model could potentially be generalized and applied to other proteins and species.

This thesis explores increasingly complex representations of RNA sites and a suite of supervised learning techniques to train models that predict the efficiency with which sites are cleaved by VapC. The simplest representations of RNA sites consider only the RNA sequence. More detail is added to the representation in the form of secondary structures and the potential influences of tertiary structures are discussed. No model is produced that is capable of accurate, meaningful predictions. This suggests that the construction of a successful model requires significant alterations to the representation of RNA sites or that the data available is insufficient for training an accurate model.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

This thesis outlines a study of the use of computational methods to characterize sites where VapC cleaves RNA in *Mycobacterium smegmatis*. VapC and its complement VapB form a toxin-antitoxin system. As a toxin, VapC targets specific RNA transcripts inhibiting the expression of the gene or genes contained on that transcript. VapB, the antitoxin, neutralises VapC. The quantity of VapC relative to the quantity of VapB present in a cell determines the extent to which genes targeted by VapC are downregulated. This control of biological processes within an organism has been proposed as serving a variety of uses for an organism or population of organisms [2]. VapBC systems have been observed in many species, including multiple instances in *Mycobacterium tuberculosis*. As a major human pathogen, the biology of *M. tuberculosis* is worthy of special interest. The study of VapC in the closely related species *M. smegmatis* is useful because *M. smegmatis* contains a single VapBC system as opposed to the over 40 putative instances in *M. tuberculosis* [3]. Modelling the activity of a single VapC protein in isolation is a simple and approachable step towards understanding the internal cellular mechanisms of *M. tuberculosis*.

The basis for this study are the results reported by McKenzie et al. [1] which include the effects of VapC on the expression of each gene in *M. smeg-*

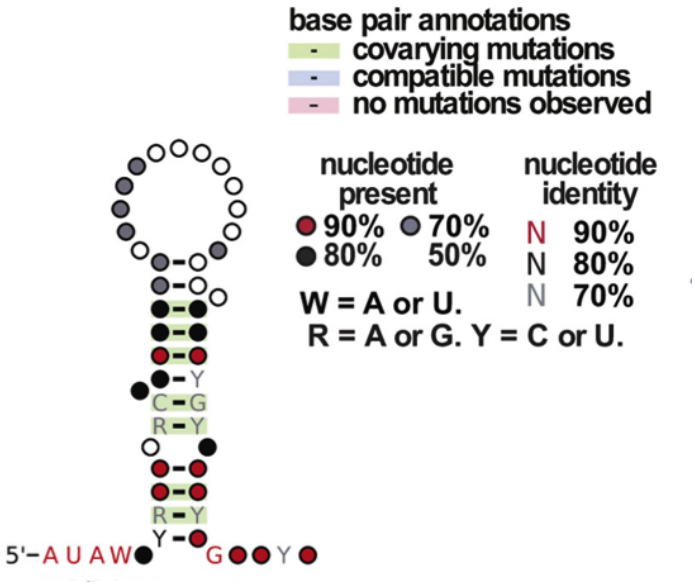

Figure 1.1: Hypothesised optimal VapC target. Image adapted from [1]

matis. Also reported by [1] is a hypothesised model for sites efficiently cleaved by VapC. The key features of this model, shown in Figure 1.1, are a four-base motif: AUAA or AUAU followed by a stem-loop structure—a structure where two complementary areas of the single-stranded RNA molecule pair together forming the double-stranded stem. While it would appear that this model is consistent with efficiently cleaved sites, not all sites conforming to the model are efficiently cleaved by VapC. Subsequent laboratory testing (currently unpublished) showed that sequences specifically fabricated to conform to this model were not cleaved by VapC.

This study aims to expand or modify the hypothesised model to produce an accurate description of the features of sites efficiently cleaved by VapC and explores the use of computational models to do so. Computational methods have uses both in the processing of data and the construction of models. While empirical, biochemical methods exist that could be used to experimentally identify exact locations and structures of VapC cleavage sites, these methods are resource intensive and the cost of a genome wide search for such sites would be immense. Compared to this, a computational model which predicts

sites cleaved efficiently by VapC would be potentially inexpensive—processing power is vastly cheaper than the time, equipment and expertise required to conduct genome-wide determinations with conventional biochemical methods.

It should be emphasised that the aim of this study is not to replace existing biochemical methods or experts in the field of molecular biology but to provide a new tool which supplements existing methods. A successful model could assist in reducing the amount of further experimentation required to characterise VapC targets. Furthermore, the process for producing a model which accurately predicts VapC activity in *M. smegmatis* could potentially be generalized and applied to other species and other members of the VapC family of proteins.

To build predictive models, this study employs structured learning techniques and a combination of sequential and structural information. Sequential information describes the sequence of nucleotides than compose the RNA while structural information describes the spatial arrangement of the RNA molecule. This first set of experiments attempt to construct two-class models using sites from the most heavily downregulated genes as one class and sites from other genes as the other class. These experiments are designed to establish that the most significantly affected genes have some distinctive shared characteristics. This approach does not produce a useful model and the division of the data into two classes is discontinued.

Further experiments use all available data and aim to build models that predict the efficiency with which a site is cut by VapC. At first sites resembling the model shown in Figure 1.1 are considered: the four-base AUAW motif, a stem-loop downstream of the motif and the adjacent subsequences. Again, this proves insufficient for training an accurate model suggesting the need to include a higher level of detail in the representation of possible cleavage sites. In an effort to develop a representation that includes enough detail, more structural features are considered.

1.2 Outcomes

Despite the range of sequential and structural features considered, no successful or promising models are produced by the experiments presented in this thesis. This outcome suggests it may not be possible to build such a model with the data available or that a substantially more detailed representation of RNA structure should be devised.

1.3 Aim

This study aims to produce a computational model capable of accurate prediction of RNA sites efficiently cleaved by VapC in *M. smegmatis*. In an effort to achieve this aim, a series of models are trained with sequential and structural data for collections of candidate cut sites. A promising model—one that is highly accurate—can be evaluated through the fabrication of novel RNA sequences that are consistent with the model and then testing the novel RNA with empirical laboratory methods to verify if they are cleaved by VapC.

1.4 Summary of Chapters

Chapter 2 gives an overview of important previous work, relevant areas of biological and bioinformatic theory, and the data processing, classification and prediction algorithms used in experiments throughout this study.

Chapters 3 to 5 outline the experimental process and results of this study. Chapter 3 presents preliminary experiments designed to establish understanding of the data. The datasets used for these experiments simplify the prediction of how efficiently a site is cut by VapC. Rather than predicting a continuous variable, the genes most significantly affected by VapC are collected as one class of instances, with all other sites grouped into the second. These experiments do not result in a successful model, which suggests a more detailed approach is required.

Chapter 4 moves from binary classification to numeric prediction of the extent to which a gene is downregulated based on the features at a candidate cut-site. No successful or promising model is produced which suggests that a more detailed representation of instances is needed for accurate prediction.

Chapter 5 first presents experiments which attempt to build predictive models similar to those tested previously in Chapter 4 but with additional structural features of the RNA. Secondly, the chapter discusses possibilities and usefulness for further increases to the level of detail used for characterising the structural features of the regions around candidate cut-sites.

Chapter 6 first summarises the experimental process and results of the study. The chapter then proceeds to discuss other possibilities for future work and presents final conclusions.

Chapter 2

Background

This chapter introduces the core concepts and previous work used throughout this thesis as a basis for both the design of experiments and the interpretation of results. First, the role, composition and structure of RNA are presented and computational methods for predicting RNA structures are outlined. Secondly, the role and properties of toxin-antitoxin systems are explained. Finally, an overview of classification and prediction algorithms is presented with further elaboration on the properties of specific techniques used in this investigation.

2.1 RNA

Ribonucleic acids (RNA) are large molecules involved in many fundamental biological processes within a cell; primarily the decoding and construction of proteins from genes coded in DNA, and the control of this process. This investigation is concerned with messenger RNA (mRNA)—RNA which contains sequences transcribed from DNA describing the composition of a particular protein or proteins.

2.1.1 RNA Composition and Structure

RNA molecules consist of long sequences of four bases: adenine, cytosine, guanine and uracil (typically denoted by their initials: A, C, G, U). These bases

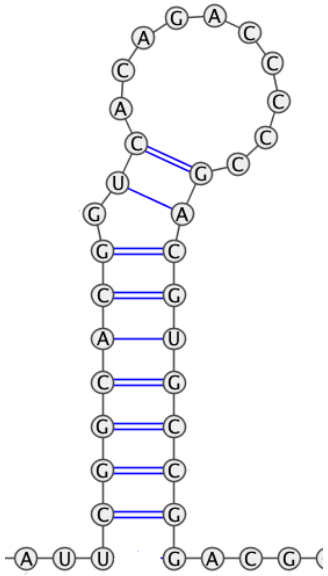

Figure 2.1: An example of a stem-loop. Image produced using the VARNA RNA visualization tool [6].

are the same as the four bases of DNA with the exception of the use of uracil in RNA rather than thymine (T). DNA molecules are well known to consist of two complementary strands of bases forming a double helix structure where A complements T and C complements G. Unlike DNA, RNA molecules are single-stranded but, like DNA, the physical properties of each base are such that the molecule is more stable when bases are paired. Bases most commonly form pairs according to the canonical Watson-Crick base pairings: A-U and C-G but other base pairings exist. This leads to complementary regions of single RNA strands forming double helix structures with one another, thereby folding the molecule and forming secondary structures called stem-loops, also referred to as 'hairpins' or 'hairpin stem-loops' [4, 5]. Figure 2.1 illustrates a straightforward example of such a structure; the stem is the region of predominantly matched pairs, possible with small bulges of unmatched nucleotides—in this case with a single unmatched 'G'; paired subsequences of RNA are not necessarily perfect complements but closer matching regions form more stable structures. The loop is the single stranded region above the stem.

A more complex class of secondary structures, known as pseudoknots, are

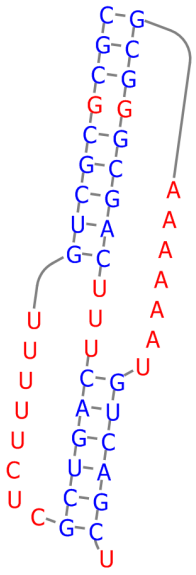

Figure 2.2: An example of a pseudoknot. Image produced using PseudoViewer 3 [7].

formed where two or more stem-loops partially overlap each other. Figure 2.2 illustrates an example pseudoknot. Here two double-stranded stem regions are observed where each includes part of the loop region corresponding to the other stem.

2.1.2 Modelling and Prediction of RNA Structures

Sequence data alone is insufficient to establish the function of a given RNA; molecular structure also plays an important role. Accurate determination of the structure of a RNA molecule is difficult. Biochemical methods exist for performing such determinations under laboratory conditions [5]. However, due to the cost of such methods, computational prediction techniques have been developed to infer the structure of RNA from its sequence. The problem of finding an optimal RNA folding is computationally complex but the relative low cost of computing power and memory makes computational prediction of structure a more practical approach than biochemical determinations provided it is sufficiently accurate.

A multitude of RNA structure prediction methods exist although most of

these represent incremental improvements upon some pre-existing approach [8, 9, 10, 11, 12, 13]. Furthermore, no comprehensive comparison of all or the most prominent of these methods exists. This section outlines the broad principles which form the basis of RNA structure prediction.

The basis of the oldest RNA structure prediction techniques is minimization of free energy in the molecule; the configuration with the least free energy will be the most stable [8]. The energy used by bonds between pairs of bases is measurable and is used in optimizing the RNA structure [14, 15]. Although soundly based on the electrochemical properties of the molecules involved, this prediction process is made impractical by the enormous number of possible configurations for all but short sequences; many slightly different configurations might be close to optimal. As a result, a large amount of the variation in prediction techniques lies in strategies for focusing on likely configurations or identifying and ignoring configurations that can be quickly identified as far from optimal. Such improvements do not inherently improve the accuracy of predictions but improvements in the processing and memory costs of prediction allow more data to be processed. In turn, the ability to process more data improves prediction outcomes by allowing for the construction of more detailed models encapsulating more information.

One difficulty encountered by the thermodynamic energy based approach is that RNA will not always fold to match the strict minimum energy configuration found by exhaustive testing. This is at least in part due to imperfect information—RNA structure prediction is typically performed on relatively short, local sequences, which are in reality part of a longer sequence. While the sequence under inspection could be extended, computation costs increase steeply as data is increased and it is often not practical or possible to model structures of entire RNA molecules. As a result, modelling techniques do not predict a single optimal structure but include any near-optimal structures that fall within a certain energy threshold from the observed minimum. This threshold can be set arbitrarily according the requirements of the application.

Interestingly, while often similar, these structures can be vastly different while having similar energy levels [9].

Some structure prediction techniques use sequences with known structure as a basis for prediction of the structure of some unknown but similar piece of RNA. Although genetic sequences differ between organisms, many genes of an organism function almost identically to similar genes found in related species through conserved structural similarities [10, 11]. Such approaches are not appropriate for use in the methods employed by this study due to their reliance on the availability of an existing body of reliable structural data—such data is not available.

2.1.3 Pseudoknot Prediction

Pseudoknots are a significant difficulty for computational structure prediction because the presence of a pseudoknot is highly context dependent. The general task of pseudoknot prediction has been shown to be NP-hard [16] although many approaches are capable of predicting select types of pseudoknots. Such techniques use heuristics to restrict and explore possibilities and while it may not be provable that such predictions are optimal, these predictions have been shown to be good approximations of reality. Unfortunately, methods which may produce promising results are rendered somewhat impractical by computational complexity. While it is tenable to apply such methods to small sets of testing data, the processing time and memory required to process large amounts of data are enormous. In the absence of definitive, practical tools for producing data about pseudoknots, this investigation does not consider them for the most part, but Chapter 5 discusses evidence as to whether pseudoknot characterization could contribute to a successful model.

2.1.4 RNALfold

Where prediction of secondary structure is required, this study uses the RNALfold algorithm, part of the ViennaRNA software package [12, 13]. This algo-

rithm is designed to provide a practical method for scanning sequences for small local structures. Rather than testing all possible subsequences of its input RNALfold is restricted to a predefined maximum structure size and identifies locally stable structures of that fixed size or smaller. RNALfold is a suitable algorithm because this study is primarily concerned with local structure around a site—such as the stem-loop present in the model presented in Section 2.2—which RNALfold is designed to predict, making it a practical option.

2.2 Toxin-antitoxin Systems and Virulence Associated Proteins B and C

Toxin-antitoxin systems are groups of two or more genes. One protein encoded by these genes, the toxin, inhibits certain cell functions by damaging RNA molecules. The complementary antitoxin protein inhibits or neutralises the toxin. Certain conditions within the cell result in higher expression of the toxin allowing these proteins to attack RNA.

Virulence Associated Proteins B and C (VapBC) is a large family of toxin-antitoxin proteins, VapC being the toxin and VapB the antitoxin. VapC proteins are ribonucleases (RNases)—catalysts which cause RNA molecules to degrade into smaller molecules. This prevents the translation of the RNA into protein [4].

This thesis is concerned with VapC in *M. smegmatis* as an example VapC toxin. Characterization of the function of this example could lead to increased understanding of related proteins.

Previous work has presented observations which strongly suggest VapC targets RNA transcripts with specific features in order to carry out its physiological function—regulation of gene expression [1]. While VapC can cleave any mRNA, only optimal or near optimal targets are cleaved efficiently resulting in proportionately greater down-regulation of the corresponding genes. Findings

from these earlier studies include a hypothesised model of the optimal motif, shown in Figure 1.1. The model shown in Figure 1.1 encapsulates the current theory that VapC targets some short motif, in this case AUAW ('W' representing either A or U) and some local secondary structure—the stem-loop following the motif. While evidence suggests that VapC targets are consistent with this model, this model is not sufficient for the differentiation of sites cleaved efficiently by VapC and sites which are not. It is clear that other factors need to be considered because not all transcripts containing sites resembling this motif are cut as expected. Currently unpublished experimentation at the University of Waikato has shown that RNA sequences constructed specifically to conform to this model are not cleaved by VapC.

2.3 Classification Techniques

Some experiments described in this thesis adapt RNA sequence data for use in classification. This section outlines algorithms commonly used for classifying data. This set of common techniques is selected as a range of techniques which function differently to provide multiple opportunities for producing a successful model or observations which contribute to producing such a model. Although the core aim of this investigation is to produce a model capable of accurately predicting sites that are efficiently cut by VapC, ideally, a successful model would also provide an explanation for its decision making. Some classifications methods, such as decision trees, are readily interpretable; human observers can easily interpret and analyse how the model makes decisions. The implementations of the algorithms used in this study are components of WEKA [17, 18].

2.3.1 OneR

The OneR algorithm generates a single rule for classification. That is, a single attribute is tested when classifying an instance. The attribute which produces

the lowest error in training is the attribute chosen for the single rule. While simple, OneR has often been shown to perform well when compared with other methods [19]. In the experiments presented here, OneR classifiers are primarily used for comparison against more complex classifiers. Circumstances where a OneR classifier outperforms a more sophisticated classifier—one that is able to use more than one attribute for prediction—suggest noisy or near random data, that the second classifier has been poorly trained, or that some weakness of the algorithm in question makes it unsuitable for the data.

2.3.2 J48

J48 is a Java implementation of the C4.5 algorithm [20]. The algorithm derives decision trees using information gain as a measure of the most useful attributes for classification. J48 decision trees are fast to train and evaluate. Decision trees are also straightforward for a human observer to visualize and interpret.

2.3.3 Random Forest

Random forest is a classification technique based on bagging. A random forest classifier is an ensemble of randomly generated decision trees. Each random tree is a decision tree constructed using a randomly selected subset of attributes. Each tree makes an independent classification decision and the majority classification of the set of trees is used as the final output of the forest. While very small random forests or random trees in isolation are of no practical value as a classifier, sufficiently large random forests have been shown to perform well under many circumstances [21]. Although a random forest classifier offers no ability to directly derive explanations of how combinations of attributes influence classification outcomes, the potential for good performance justifies its inclusion here.

2.3.4 *k*-Nearest Neighbours

k-Nearest Neighbours (k-NN) is a type of instance-based learning [22]. Rather than using specific combinations of attributes to predict the class, k-NN compares entire instances to find a small number (k) of instances most similar to the instance being classified. The similarity of two instances is determined by the number of equal or (for numeric values) close attributes shared by those instances. An instance will be classified as the majority class of the selected similar instances [23].

2.3.5 Naive Bayes

Naive Bayes classifiers are probabilistic models built based on Bayes' Theorem. Training of a Naive Bayes classifier assumes attributes are independent [24]. Naive Bayes classifiers can be trained quickly and are often capable of good performance with relatively small training sets.

2.4 Numeric Prediction Techniques

Not all experiments presented in this thesis use classification techniques, in some cases numeric prediction is appropriate. Numeric prediction is necessary when the class is a continuous variable rather than a discrete variable.

The implementations of the algorithms used in this study are components of WEKA [17, 18].

2.4.1 Linear Regression

Linear regression is a simple method of numeric prediction: each attribute is a assigned a weight and the class is then expressed as the attributes. Although this method has the major weakness of assuming linearity, it can be useful for gaining insight into the data as guidance for subsequent application of more sophisticated methods [18].

2.4.2 *k*-Nearest Neighbours

k-NN is the one algorithm described here used for both classification and numeric prediction. As for classification, the k instances most similar to the current instance are identified. The value predicted is the mean class value of the selected instances [22, 23].

2.4.3 Model Trees

Model trees combine decision trees and linear regression. A tree is constructed by the division of data into different branches by testing an attribute at each internal node of the tree. Each leaf of the tree is a linear regression function. Model trees are capable of greater accuracy than linear regression because separate linear functions are used for subsets of the data. This can be useful when a clear, but non-linear, pattern is present [25, 26]. The WEKA implementation of model trees is referred to as M5P.

Chapter 3

Two Class Approach

A preliminary approach to identifying characteristics of VapC cut sites is to build a two class model—predicting whether a site will either be cut or not. This chapter details a series of relatively simple experiments devised to explore the problem in small steps. These experiments build familiarity with the data and show the predictive usefulness of several attributes in isolation. The overarching aim here is not to build a final model, but to use simple approaches to guide the way to doing so.

3.1 Data Preparation

The set of positive instances is produced using a set of the most downregulated genes as reported in [1]. Negative instances are sites that match the initial consensus model from the entire genome sequence excluding regions containing the positive instances—a pool of approximately 3000 instances. In order to have a balanced dataset, a random negative instance is selected for each positive instance; this results in 98 instances. To avoid the introduction of errors through sampling only a small number of the available negative instances, each experiment presented in this chapter is repeated with multiple different random subsets of negative instances and the average results presented. Each experiment in this chapter utilises the same suite of classification techniques: OneR, J48, Random Forest, Naive Bayes and k-NN using

the WEKA framework [17, 18].

3.2 AUAW sites

An initial experiment is conducted using a dataset constructed using the following attributes in addition to the class: 10 bases upstream and 10 downstream of the four-base motif, and the fourth base in the motif (W, a variable A or U). Ten bases upstream and downstream are used because VapC is a relatively small molecule and therefore only able to target short sequences [5]. This experiment deliberately omits the stem-loop found in the original consensus model in order to determine the effectiveness of a model which considers the cut-site and surrounding sequence data only. Models are evaluated using 10-fold cross-validation.

3.2.1 Results

Table 3.1 outlines the performance of different classification algorithms on the dataset described above. All are relatively close to 50% accuracy and all produce a similarly high amount of error. In a two class domain, an average of approximately 50% accuracy would be achieved by simply randomly classifying instances; a successful or promising model must exceed this threshold by a significant margin.

OneR, the least sophisticated method used, is the most accurate at 55.1% average accuracy but barely better than random class assignments. The poor performance of all algorithms on this data suggests important features of a cut-site are not represented in the data. This is not an unexpected result, the dataset used to train these models deliberately uses only sequence information despite some expectation that structure is a key part of cut-site characterization. These results suggest that a sequence-only approach is insufficient and more information is required. To this end, Section 3.3 outlines similar experiments which include structural information in addition to the sequence

Classifier	Accuracy	Root mean squared error
OneR	55.1%	0.6701
J48	42.86%	0.6497
Random Forest	43.88%	0.5328
Naive Bayes	42.86%	0.5915
K-NN	51.02%	0.6701

Table 3.1: Classification accuracy for AUAW sites

information used here.

3.3 AUAW sites with Stem-loop present

The second iteration of this experiment considers the structure of the RNA. The original consensus model introduced in Section 2.2 includes a stem-loop directly downstream of the four-base motif. To integrate this information into the model, a subsequence of 70 rather than 10 bases is retrieved for the downstream region of the instance. The downstream subsequence is then processed using the RNALfold algorithm [12, 13]. RNALfold predicts possible structural configurations of the subsequence and ranks the predictions using the minimum free thermodynamic energy principle [8]. The highest ranked of these predictions for each instance is included in the dataset. Table 3.2 lists and describes the attributes used: '-1' to '-5' are the nearest upstream nucleotides; similarly, positions '+1' to '+5' are the nearest downstream nucleotides not part of a stem-loop; positions 's1' to 's5' are nucleotides at the base of the stem-loop, the portion of the structure spatially nearest the cut site; 'W' is the fourth nucleotide in the AUAW motif (either A or U) and 'D' is the distance from the end of the AUAW motif to the downstream stem-loop. Figure 3.1 illustrates how the attributes are arranged around the candidate cut-site. This set of attributes includes regions spatially near the candidate cut-site because VapC can only directly interact with a small region of RNA near the cut-site [27]. As

Attribute	Description	Value
-5	Nucleotide five bases upstream from motif	A, U, C, G
-4	Nucleotide four bases upstream from motif	A, U, C, G
-3	Nucleotide three bases upstream from motif	A, U, C, G
-2	Nucleotide two bases upstream from motif	A, U, C, G
-1	Nucleotide one base upstream from motif	A, U, C, G
w	Fourth base of the motif	A, U
D	Distance from motif to stem-loop	integer
s1	First nucleotide of the stem-loop	A, U, C, G
s2	Second nucleotide of the stem-loop	A, U, C, G
s3	Third nucleotide of the stem-loop	A, U, C, G
s4	Fourth nucleotide of the stem-loop	A, U, C, G
s5	Fifth nucleotide of the stem-loop	A, U, C, G
+1	Nucleotide one base downstream from motif	A, U, C, G
+2	Nucleotide two base downstream from motif	A, U, C, G
+3	Nucleotide three base downstream from motif	A, U, C, G
+4	Nucleotide four base downstream from motif	A, U, C, G
+5	Nucleotide five base downstream from motif	A, U, C, G
cuts?	Class attribute	boolean

Table 3.2: Attributes of the dataset used in Section 3.3.

in Section 3.2, models are trained using each of: OneR, J48, Random Forest, Naive Bayes, and k-NN and evaluated with 10-fold cross-validation.

3.3.1 Results

Table 3.3 shows the average performance of several classification techniques for the dataset described above. These results are similar and in some cases identical to those produced in Section 3.2. The lack of differences between the two sets of results suggests the new attributes characterizing the down-

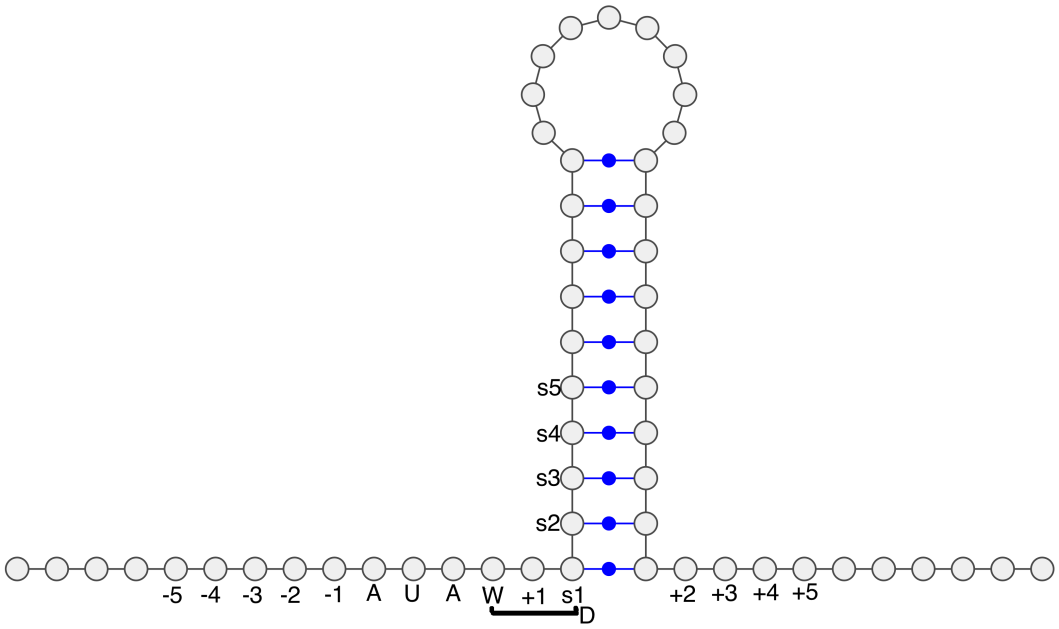

Figure 3.1: Generalised diagram of an instance from the dataset used in Section 3.3 showing the features used to construct the instance and how these are positioned relative to the central AUAW motif.

stream stem-loop are less useful to classification than those already included in Section 3.2.

Table 3.4 shows attributes ranked by information gain (attributes scoring below 0.01 are omitted). Information gain indicates the strength of an attribute as a predictor. It is notable that all these information gain values are low—all are below 0.1—suggesting these attributes—nucleotides at specific locations around a candidate cut-site—are of little value to classification. Section 3.4 presents a refinement based on these findings which generalizes attributes to have less dependence on specific locations.

3.4 Generalisation of attributes

Sections 3.2 and 3.3 derive individual attributes for bases near a candidate cut-site in isolation. Such an approach could reduce predictive effectiveness if the presence of certain bases within a region is a stronger determining factor in VAPC activity than the exact positions of specific individual bases. There is

Classifier	Accuracy	Root mean squared error
OneR	55.10%	0.6701
J48	42.86%	0.6497
Random Forest	48.98%	0.5246
Naive Bayes	42.86%	0.5915
K-NN	48.98%	0.71

Table 3.3: Average Classification accuracy and error for AUAW sites with downstream stem-loop

also the possibility that a better model would consider some specific locations alongside the presence or absence of some feature within a subsequence. The experiments in this section mirror those previously presented in this chapter but divide each instance into small regions and use the amount of each base present in each region as attributes.

The data used is the same as that used in Section 3.3 but it is represented differently. This dataset uses the attributes listed in Table 3.5: the number of A, C, G and U bases observed in each of three regions—upstream, downstream and the base of the predicted stem loop; the fourth base in the AUAW motif; the distance between the AUAW motif and the stem-loop; and the class.

3.4.1 Results

Table 3.6 presents average accuracy and error rates for simple classifiers using the generalized dataset. These results are similar to those produced previously in Sections 3.2 and 3.3. Most approaches do not achieve even 50% accuracy. Such poor performing classifiers suggest that the small set of instances used does not contain sufficient information for training classifiers. Furthermore, there is no apparent confirmation in the results that the attributes derived are features of a RNA site relevant to the ability of VapC to cut that site.

Information Gain	Attribute
0.063282	p+1
0.052446	p+3
0.041682	p+5
0.040789	s2
0.036978	s1
0.029565	p+2
0.024255	p-1
0.020522	s3
0.019968	p-3
0.019758	p-4
0.013863	p-2

Table 3.4: Attributes ranked by information gain. Attributes with calculated information gain below 0.01 are omitted. Attributes p+n signify positions after the predicted stem-loop, attributes p-n signify positions upstream of the AUAW motif, attributes sn signify positions on the stem.

3.5 Discussion

None of the approaches detailed in this chapter perform at even near useful levels. This suggests two-class classification may be an oversimplification of the problem; an insufficient representation has been used; or the datasets used are too small to train useful models. It was understood initially that VapC activity is not limited to the small set of genes used as positive instances in these experiments, these are genes most affected by the enzyme. It remains conceivable that simple models could give some insight into what features contribute to encouraging VapC activity and thus provide direction for creating more refined models but the results presented here suggest the simplification of the data to a small two class dataset does not leave enough information for producing a useful model. As a result, the experimentation outlined in

Attribute	Description	Value
-A	Number of A nucleotides present in upstream region	integer
-U	Number of U nucleotides present in upstream region	integer
-C	Number of C nucleotides present in upstream region	integer
-G	Number of G nucleotides present in upstream region	integer
w	Fourth base of the motif	A, U
D	Distance from motif to downstream stem-loop	integer
sA	Number of A nucleotides present in stem-loop	integer
sU	Number of U nucleotides present in stem-loop	integer
sC	Number of C nucleotides present in stem-loop	integer
sG	Number of G nucleotides present in stem-loop	integer
+A	Number of A nucleotides present in downstream region	integer
+U	Number of U nucleotides present in downstream region	integer
+C	Number of C nucleotides present in downstream region	integer
+G	Number of G nucleotides present in downstream region	integer
cuts?	Class attribute	boolean

Table 3.5: Attributes of the dataset used in Section 3.4.

Classifier	Accuracy	Root mean squared error
OneR	51.02%	0.6999
J48	48.98%	0.6319
Random Forest	47.96%	0.5213
Naive Bayes	41.84%	0.5969
KNN	51.02%	0.5762

Table 3.6: Average performance for generalized AUAW sites with stem-loop

Chapters 4 and 5 does not make such simplifications.

Chapter 4

Numeric Prediction

The series of experiments described in Chapter 3 uses simplified data in a series of attempts to gain insight into both the problem domain and data. However, these datasets appear to contain insufficient information for making classifications, leading to unsuccessful models. This chapter presents a similar series of experiments using both more data and a more sophisticated understanding and representation of the underlying biological processes described by these data.

Despite Chapter 3 representing the activity of VapC upon a given sequence of RNA as either cleaving or not, this property is more correctly represented as a continuous measure of how efficiently that RNA is cleaved. VapC will cleave any piece of RNA, with the nature of the RNA determining how efficient this process is. In the experiments described in this chapter, the extent of the downregulation of a gene is used to represent how efficiently VapC cleaves that gene's transcript. Under this approach there is now a continuous variable being predicted: how efficiently a transcript is cleaved.

The experiments in this chapter aim to model the extent to which a site is affected by VapC through predicting gene expression ratios. Rather than using a subset of the available data, all data recorded via microarray analysis in [1] is considered. In addition to direct numeric prediction, several discretization schemes are also tested (grouping instances into ranges of values rather than

prediction specific values).

4.1 Regression Methods

The first experiments closely resemble the two-class experiments outlined in Chapter 3; sites resembling the consensus model (discussed in Section 2.2) are collected as instances and attributes are derived from the surrounding RNA sequence. The same set of attributes outlined in Section 3.3 is used: 10 bases upstream of the 4 base motif, the variable fourth base in the motif, distance (number of bases) between predicted stem-loop and motif, 5 bases at the bottom of the stem, and 10 bases downstream of the four-base motif. Unlike the previous procedure, instances are selected on a gene by gene basis. The data reported by [1] includes the gene expression ratio for each gene and a measure of significance for each observation. For this experiment only statistically significant observations are considered. For each of these genes, the gene sequence along with 150 upstream and downstream bases are extracted from the sequenced reference *M. smegmatis* genome. Instances are generated from each of these sequences by locating AUAW subsequences with a predicted downstream stem-loop up to 10 bases away. This results in a total of 745 instances. Each instance is then labelled with the expression ratio recorded for that gene.

This dataset is now used to train models using a set of regression techniques: linear regression, model trees and k-NN. As previously, the WEKA implementations of these algorithms are used [17].

4.1.1 Results

Table 4.1 records the performance of three numeric prediction techniques on the previously described dataset. Linear Regression and M5P perform very similarly with correlation coefficients of 0.13 and 0.12 respectively. On further inspection of the model trees trained by the M5P algorithm, it can be seen

Model	Correlation coefficient	Root mean squared error
Linear Regression	0.13	0.40
M5P	0.12	0.41
k-NN	0.04	0.60

Table 4.1: Average correlation coefficient and error for numeric prediction techniques

that the trees are pruned to a single leaf node—the algorithm is unable to find any significant features by which the data could be divided and more accurate predictions produced. Such a model is equivalent to a linear regression model hence the closeness in the performance of these two approaches. The third numeric predictor, trained using k-NN, is less accurate than linear regression and M5P with a correlation coefficient of 0.04. Although the ideal result is clearly a perfect correlation, correlation coefficients in excess of 0.6 could suggest a promising model that can be improved. With performances well below this margin these models are far from accurate which suggests the need for additional information to be included in the representation.

4.2 Discretization

Discretization converts a numeric variable (in this case expression ratio) into discrete ranges meaning a model is only required to put each instance into the correct range rather than correctly predict an exact numeric value. If groups with strong similarity between their members can be identified, a reliable model can be produced. Conversely, it is possible to produce situations where there is little meaning in the classes produced by discretization; the classes imposed may be purely arbitrary with no common features among instances grouped together or meaningful differences between instances not grouped together.

The two class approach explored in Chapter 3 represents a supervised discretization scheme, with one class consisting of sites from the most downreg-

ulated genes and the other class consisting of sites from anywhere else. In contrast, this section explores the effectiveness of multiple unsupervised discretization schemes, grouping different combinations of ranges of expression ratios into sets of classes.

Two discretization methods are used, both based on the discretization tools available in WEKA: equal width and equal frequency. Both approaches divide data into a predefined number of discrete classes. Equal width discretization takes the full range of possible values and divides it into smaller ranges of equal magnitude. Equal frequency discretization aims to produce groups of uniform size although in practice classes are typically not perfectly uniform—instances with equal original values are required to be be assigned to the same class resulting in minor variations in the size of each class. Both discretization approaches are tested multiple times varying the number of classes used.

4.2.1 Results

Tables 4.2, 4.3 and 4.4 record average performance of three classifiers for datasets discretized into three, five and seven equal width classes respectively. While these results may look promising at the outset, they are misleading and serve as an illustration of the drawbacks of discretization. Table 4.2 shows three classification techniques all achieving an average of 95.31% classification accuracy. This is due to the discretization scheme grouping over 95% of instances together. With such an unbalanced dataset, simply classifying any instance as a member of this majority group results in correspondingly high accuracy with relatively low error.

Using seven equal width classes has the same problem but slightly less pronounced—one class consists of approximately 75% of instances. Table 4.4 shows the average classification accuracy for each of three classifiers is close to 75% which suggests all instances are assigned the majority class. Closer inspection of the trained classifiers confirms this.

Table 4.3 shows that discretization with 5 equal width classes results in the

lowest average classification accuracy and the highest error of the three equal width discretized datasets. Unlike the three and seven class datasets, this dataset has two large classes rather than one. This results in classifiers which do not simply assign all instances to the majority class. However, the models produced are not effective at distinguishing which of the two large classes an instance belongs to resulting in high error.

The difficulty with equal width discretization is in handling outlying values. In this case a few high expression ratio values lead to one or two very large classes and a number of almost empty classes consisting of the outliers themselves. Equal frequency discretization removes this problem by forming classes each consisting of approximately the same number of instances. Tables 4.5, 4.6 and 4.7 record average performance of three classifiers for datasets discretized into three, five and seven equal frequency classes respectively. For all three classifiers considered, these results follow a trend: as the number of classes increases, both classification accuracy and error decrease. Furthermore, random forests consistently produce the highest accuracy with the lowest error. However, for each classifier, the accuracy achieved is only slightly greater than would be achieved through random class assignment. This suggests there is little information within the discretized data which classifiers can use to determine which class an instance should be assigned.

4.3 Discussion

In this chapter two groups of experiments have been described. The first use numeric prediction techniques to predict the expression ratio of a gene based on candidate VapC cleavage sites on or near the gene. Models produced do not achieve a high level of correlation. These results could be due to an insufficient quantity of data. It is possible that a set of approximately 750 instances is insufficient to identify patterns within a noisy dataset, however, more data is not available at present; there is no immediate solution to this

Classifier	Accuracy	Root mean squared error
OneR	95.31%	0.18
J48	95.31%	0.17
Random Forest	95.31%	0.18

Table 4.2: Results for equal width discretization—3 classes

Classifier	Accuracy	Root mean squared error
OneR	60.06%	0.40
J48	64.26%	0.33
Random Forest	64.10%	0.30

Table 4.3: Results for equal width discretization—5 classes

Classifier	Accuracy	Root mean squared error
OneR	74.47%	0.27
J48	74.90%	0.24
Random Forest	76.19%	0.24

Table 4.4: Results for equal width discretization—7 classes

Classifier	Accuracy	Root mean squared error
OneR	37.18%	0.65
J48	36.31%	0.56
Random Forest	41.20%	0.47

Table 4.5: Results for equal frequency discretization—3 classes

Classifier	Accuracy	Root mean squared error
OneR	21.76%	0.56
J48	22.61%	0.49
Random Forest	25.76%	0.40

Table 4.6: Results for equal frequency discretization—5 classes

Classifier	Accuracy	Root mean squared error
OneR	16.31%	0.49
J48	16.39%	0.43
Random Forest	19.19%	0.35

Table 4.7: Results for equal frequency discretization—7 classes

problem. Furthermore, RNA features meaningful to the activity of VapC may not be represented within the dataset; these experiments consider a variety of sequence and structural features but he inclusion of more features may be necessary to distinguish the efficiently cleaved sites from the inefficiently cleaved sites. Experimentation in subsequent Chapter 5 adds more structural features to each instance in an effort to identify all relevant information needed to make accurate predictions of VapC cleavage sites.

Additional experiments in this chapter used a number of discretizations applied to the data in order to observe any significant groupings of similar instances that might exist. These experiments proved unfruitful. Equal width discretization schemes do not handle outlying values well, leading to unbalanced datasets and classifiers which assign all test instances to the majority class. Such classifiers are of no use. Equal width discretization proved equally unfruitful, with classifiers only yielding slightly higher performance than random classification.

Chapter 5

Stem-loop Interactions

The preceding chapters describe a series of experiments which attempt to characterize sites that are efficiently cleaved by VapC. Chapter 3 uses classification techniques with a two class dataset. Chapter 4 uses numeric prediction to produce models which predict the expression ratio of genes. The sole structural feature included in the datasets constructed for previous experiments is the stem-loop downstream of the four-base AUAW motif shown in the hypothesised model introduced in Section 2.2. However, this hypothesised model has been shown to be insufficient for the identification of sites efficiently cleaved by VapC through physical biochemical techniques; further information is required for an accurate model of VapC targets. Furthermore, these previous experiments do not show any predictive value in the sequential and structural characteristics tested. While there is strong evidence that sites that are cut efficiently conform to the hypothesised model introduced in Chapter 1, the issue of distinguishing these sites from similarly composed sites that are not cut efficiently remains outstanding. This suggests that more characteristics need to be considered in order to produce a successful model.

This chapter considers the possibility that a more detailed structural representation of a candidate cut-site is needed for making accurate prediction. First, the predictive power of the inclusion of additional secondary structures near candidate cut-sites is explored through experimentation. The inclusion

of an upstream stem-loop is based on the understanding that enzymes such as VapC target small regions of RNA (in this case the four bases AUAW). Local structure around such regions influences both how efficiently the enzyme can cleave the RNA and whether the enzyme is able to access its target; the structural configuration of an area of RNA could make certain areas physically difficult or impossible for the VapC molecule to reach. Secondly, further possible structural considerations are discussed along with how experiments using such data could be approached.

5.1 Introducing upstream stem-loops

For this set of experiments, a dataset is constructed similar to the dataset described in Section 3.3 with the addition of an upstream stem-loop. The upstream portion of the sequence is input into the RNALfold algorithm (as previously done with the downstream subsequence) and a characterisation of the strongest nearby result is included in the dataset. Table 5.1 lists the features included when constructing an instance. Positions '-1' to '-5' in the figure are the nearest upstream nucleotides not part of a stem-loop. Similarly, positions '+1' to '+5' are the nearest downstream nucleotides not part of a stem-loop. Positions 'd1' to 'd5' are nucleotides at the base of the downstream stem-loop—the portion of the structure spatially nearest the cut site. Similarly 'u1' to 'u5' are nucleotides at the base of the upstream stem-loop. 'W' is the fourth nucleotide in the AUAW motif (either A or U). 'Dd' is the distance from the end of the AUAW motif to the downstream stem-loop and 'Du' is the distance from the end of the AUAW motif to the upstream stem-loop. Figure 5.1 illustrates illustrates how the attributes are arranged around the candidate cut-site. As previously, the expression ratio for each gene and the statistical significance of each observation is taken from the findings of [1].

As in Chapter 4, instances are constructed for genes whose observed expression ratios are considered statistically significant. Each site on or near (i.e.

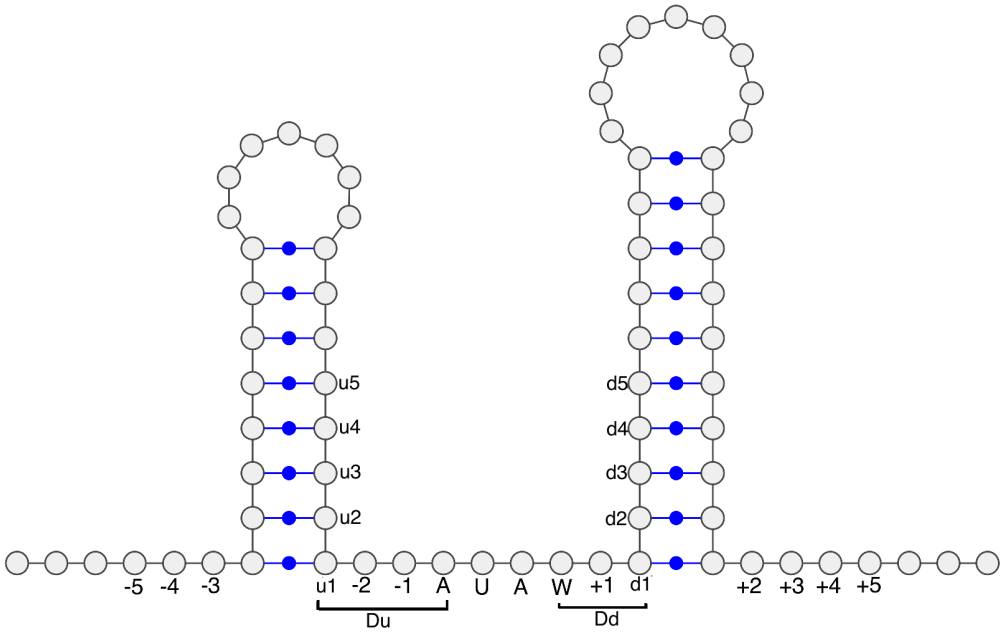

Figure 5.1: Generalised diagram of an instance from the dataset used in Section 5.1 showing the features used to construct the instance and how these are positioned relative to the central AUAW motif.

within 150 bases of either end) a gene has an instance constructed. This results in 745 instances. This dataset is used to build numeric prediction models using three techniques: linear regression, M5P and k-NN. Each approach is tested using 10-fold cross-validation.

5.1.1 Results

Table 5.2 outlines the performance of the three regression techniques for this new double stem-loop dataset. The highest correlation coefficient achieved is 0.1—these models have no predictive power. In this respect, these results are all similar to the results produced without the inclusion of an upstream stem-loop (see Table 4.1). The lack of significant performance improvement resulting from more detailed data suggests that it is possible that the structure upstream of a site does not have meaningful influence over VapC activity. Alternately, these observations may reflect a lack of sufficient detail within

Attribute	Description	Value
-5	Nucleotide five bases upstream from motif	A, U, C, G
-4	Nucleotide four bases upstream from motif	A, U, C, G
-3	Nucleotide three bases upstream from motif	A, U, C, G
-2	Nucleotide two bases upstream from motif	A, U, C, G
-1	Nucleotide one base upstream from motif	A, U, C, G
u1	First nucleotide of the upstream stem-loop	A, U, C, G
u2	Second nucleotide of the upstream stem-loop	A, U, C, G
u3	Third nucleotide of the upstream stem-loop	A, U, C, G
u4	Fourth nucleotide of the upstream stem-loop	A, U, C, G
u5	Fifth nucleotide of the upstream stem-loop	A, U, C, G
Du	Distance from motif to upstream stem-loop	integer
w	Fourth base of the motif	A, U
Dd	Distance from motif to downstream stem-loop	integer
d1	First nucleotide of the downstream stem-loop	A, U, C, G
d2	Second nucleotide of the downstream stem-loop	A, U, C, G
d3	Third nucleotide of the downstream stem-loop	A, U, C, G
d4	Fourth nucleotide of the downstream stem-loop	A, U, C, G
d5	Fifth nucleotide of the downstream stem-loop	A, U, C, G
+1	Nucleotide one base downstream from motif	A, U, C, G
+2	Nucleotide two base downstream from motif	A, U, C, G
+3	Nucleotide three base downstream from motif	A, U, C, G
+4	Nucleotide four base downstream from motif	A, U, C, G
+5	Nucleotide five base downstream from motif	A, U, C, G
gene start	Distance from motif to the start of the gene	integer
expression ratio	Class attribute	real number

Table 5.1: Attributes of the dataset used in Section 5.1.

Model	Correlation	Root mean squared error
Linear Regression	0.10	0.41
M5P	0.10	0.42
k-NN	0.07	0.42

Table 5.2: Average correlation coefficient and error for numeric prediction using the upstream stem-loop dataset

the dataset; some essential feature could be absent or otherwise represented poorly. Section 5.2 discusses possibilities for improving the representation of RNA structure.

5.2 Improving Structural Representation

It is possible that the mere presence of nearby stem-loop structures is not a useful predictor of VapC activity. Interaction between these structures could be more important than each structure in isolation. Two or more stem-loops can form a pseudoknot; a compound structure where portions of the unpaired loops pair with each other [4]. To investigate this, the unmatched loop portions of the two predicted stem-loops are considered—these are the regions that would be part of a pseudoknot if a pseudoknot were present.

Using the RNAPKplex algorithm [12, 28] to search for pseudoknots near the verified cut-sites reported by [1] does not produce any evidence of the presence of a pseudoknot near VapC cut-sites that could be a feature important to VapC activity. However, it should be noted that this search is limited to the small number of cut sites documented by [1]; additional time and, more importantly, data would provide more reliable and compelling conclusions regarding the possible importance of local pseudoknots. It should also be reiterated that algorithmic pseudoknot prediction techniques are imperfect which could lead to important structural features not being identified.

It may be possible to find meaningful interactions between stem-loops near

a cut-site rather than explicitly searching for pseudoknots. If some higher-order structure is present, there will be regularities within the single-stranded loop regions of the stem-loops. If these regions have predominantly complementary sequences, the pairing of these single stranded regions may be a part of the overall molecular structure.

Unfortunately representing such structures in a sufficiently detailed yet uniform way is a difficult task and not something this thesis can claim to present a solution to. In many instances local secondary structures are simple, such as the example illustrated by Figure 5.2, which shows an AUAA site directly between two small stem-loop structures. If every stem-loop relevant to this investigation was similar to those shown in Figure 5.2, it would be conceivable that the structural features of each candidate cut-site could be characterised into a succinct set of attributes much like has been done to prepare datasets for previously presented experiments. However, there is substantial variation in the structures identified around different candidate cut-sites. Figure 5.3 illustrates a pair of complex structures known as helical junctions. Helical junctions are essentially a compound of multiple stem-loops. Characterizing such structures in a way that retains the general properties necessary for accurate prediction for regions whose structures are complex and substantially different to each other is not a problem whose answer immediately suggests itself. Furthermore, it should be emphasised that, unlike the illustrations presented here, RNA molecules are three-dimensional. The relative position and orientation in three-dimensional space of local structures is important in determining what interactions are possible, but difficult to effectively represent within an abstract dataset. It is not clear how one might represent such three-dimensional structures in a dataset that can be used for prediction.

It is also important to consider that, when considering large complex structures within RNA and the interaction of those structures, more than the sequentially nearby area is important. While it can be useful to look for small local structures within a small region, when considering the RNA macromolecule

as a whole, regions that are sequentially distant may in reality be spatially adjacent in three dimensions due to the overall folding of the molecule. This is possibly the reason why prediction of VapC activity based solely on local information yielded limited success. Without experimentally validated data as a basis, the development of a model which includes such detailed structural information would be predominantly guesswork. To further pursue this line of investigation, the researcher would require a substantially larger amount of sites VapC is known to cleave efficiently, in addition to the expression ratios for each gene. With this data it would be possible to test and refine methods of using the structure of entire RNA molecules to build a detailed model of the structure of efficiently cleaved sites. However, even if this information were readily available, there is no immediately clear way to then apply the information in the construction of predictive models. The sheer amount of information needed to describe the entire structure of a RNA molecule in a general enough format to allow meaningful comparison with other molecules suggests that construction of a predictive model with this information would be difficult if not impossible.

While the potential for the use of macromolecular structure for prediction is predominantly conjecture, the fact remains that complex structures could reasonably be important features that contribute to VapC activity. While it is not clear how this information could be used for prediction, it is clear that it is reasonable to remain open to the possibility that complex structure may be greatly important in determining how efficiently VapC cleaves a site.

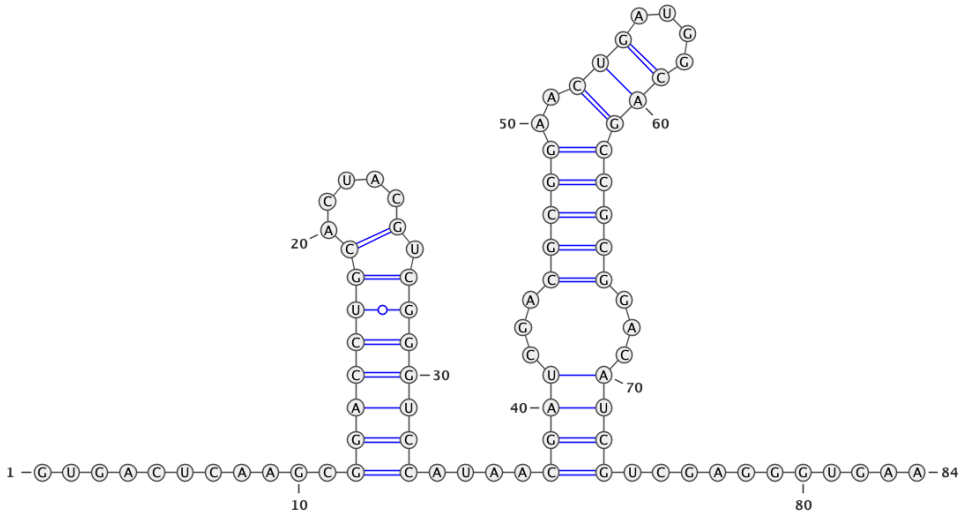

Figure 5.2: An example of a simple stem-loop structure predicted by RNALFold. Image produced using the VARNA RNA visualization tool [6].

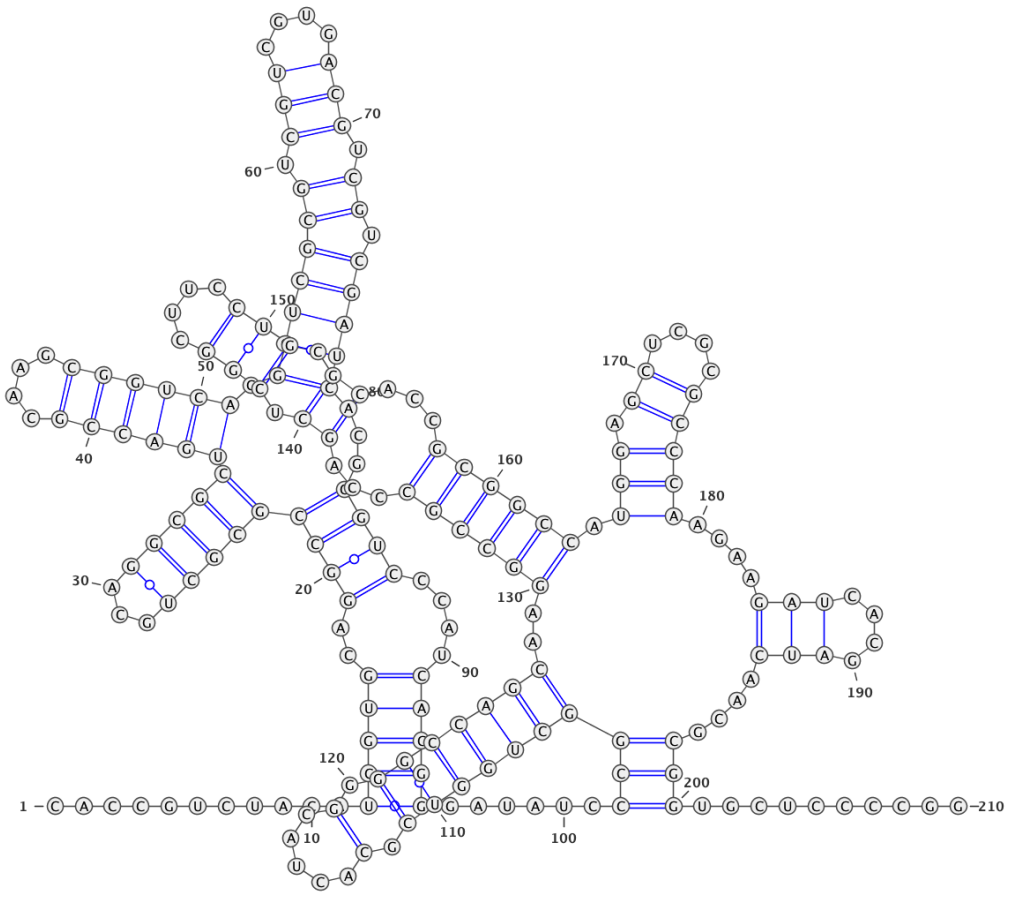

Figure 5.3: An example of a complex stem-loop structure predicted by RNALFold. Image produced using the VARNA RNA visualization tool [6].

Chapter 6

Summary, Future Work and Conclusions

6.1 Summary

This thesis has explored machine learning methods to characterise the sequential and structural features that influence where VapC cleaves RNA transcripts in *M. smegmatis*. Chapters 3 to 5 present the experimental portion of this study. Chapter 3 uses a simplified view of the data; adopting a two-class classification approach in order to establish any strong, yet simple, patterns within the data. Although the models produced might be expected to heavily over-fit the data, due to the small size of the datasets used, the models produced perform poorly and are unable to significantly outperform random classification.

Chapter 4 explores numeric prediction and uses all available data, rather than a subset as used in Chapters 3. Models are trained to predict the expression ratios observed for each gene in the *M. smegmatis* genome reported by [1]. The expression ratio of a gene is directly related to the efficiency with which VapC can cut a site or sites within the RNA transcript of that gene. No successful or promising model is produced. The data is further adapted, first by discretization, that is to say, attempting to divide the data into groups that

can be used by a classification algorithm. Secondly, each instance is modified to use a more generalised representation: rather than using the nucleotides at specific locations as attributes, the composition (proportions of each base) of regions near a site is used. Neither of the discretized or generalised datasets produce models that significantly outperform earlier experiments.

Chapter 5 adds greater structural information to the datasets used in Chapter 4. Despite the higher level of detail, models trained on this dataset present no improvement on previous results. Further possibilities for using complex RNA structures in prediction are discussed, but both time and computational constraints preclude thorough experimentation.

6.2 Future Work

This section outlines some possible avenues down which the lines of experimentation presented in this thesis could be continued. Some possible continuations are relatively straightforward but insufficient time or data has resulted in the inability to fully realise these experiments in the course of this investigation. Experimental exploration of other areas of interest, such as the use of the three-dimensional structure or entire RNA molecules, is highly speculative. The experiments presented throughout Chapters 3 to 5 have not resulted in a successful model for predicting VapC cleavage sites. However, they have not shown that it is impossible to make a successful model; further investigation and experimentation is required to produce such a model or prove the impossibility or infeasibility of the general approach.

6.2.1 Additional data

It may be the case that more experimentally verified data is required to better guide the creation and testing of a predictive model. The primary source of data is the microarray analysis presented in [1] which associates each gene in the *M. Smegmatis* genome with its level of expression after the introduction

of VapC. Although it is clear that VapC cleaves many RNA transcripts within the *M. Smegmatis* cell, only a small number of precise locations were available when designing the experiments presented in this thesis. The availability of the exact locations of all (or a large number of) sites efficiently cleaved by VapC would provide a promising opportunity for the construction of an accurate model. While collecting complete data for *M. Smegmatis* would alleviate the need for a predictive model, the insight gained through the process development of an accurate predictor could potentially be generalised and applied to other species. Part of the initial motivation for this study was the potential to use computational methods to predict VapC targets without the need the experimentally verify the efficiency with which VapC cleaves a large number of sites. The results of this study have shown this to be difficult. It may not be possible to accurately predict VapC activity with the data available to this study; additional data may be necessary to make the construction of a successful model possible.

6.2.2 Use of large RNA structures in prediction

Chapter 5 introduced and discussed some of the possibilities for the use of the structure of larger areas of RNA for prediction. The results reported throughout this thesis suggest that datasets limited to small areas contain insufficient information for the training of a successful model. It is reasonable to hypothesise that a more detailed representation of RNA structure is required to characterise candidate cut-sites such that sufficient information is included for the training of a successful model. Rather than characterising the structural features of a small local region, as in the experiments presented throughout previous chapters, it may be possible to use the structure of entire RNA molecules for prediction. The practicality of training a model with the three-dimensional structures of entire RNA macro-molecules is questionable. The general process of predicting the structure of RNA from a given sequence is imperfect but, through allowing sufficient time and computational resources,

accurate structures can be produced. However, even ignoring all difficulties in the acquisition of accurate structural information, the processing of large detailed structures introduces further difficulties which may impede the construction of predictive models. Representation of the data alone is a major difficulty. Each instance is itself a large amount of complex data which needs to be stored in a manner general enough that instances can be meaningfully compared and characteristics important to predicting VapC activity are retained. It is difficult to speculate how the three-dimensional structure of RNA molecules can be represented as a finite vector of attributes for use with prediction or classification techniques. It may be the case that a structured learning approach is not appropriate for this domain; other approaches should instead be explored. Techniques exist for the determination of the similarity of RNA structures, for example [29]. A comprehensive review of existing techniques and their suitability for use comparing candidate VapC cut-sites is required to pursue this line of research.

6.3 Conclusion

The overarching aim of this study is to construct a computational model capable of accurate prediction of how efficiently a site on an RNA molecule is cleaved by VapC in *M. smegmatis*. The experimental process undertaken to achieve this aim has not produced a successful model. The models produced throughout the experiments presented in this thesis have universally possessed no meaningful predictive capabilities; these models are not capable of significantly outperforming models which make decisions entirely at random.

Part of the initial plan for this study included the use of a successful or promising model to design novel RNA sequences that should be efficiently targeted by VapC. These novel sequences could then be fabricated and tested with VapC under laboratory conditions as an indicator of the model's performance. This aspect of the study remains incomplete because no model could

be produced that is promising enough to justify the fabrication and testing of novel RNA. Without observing a model that demonstrates some meaningful predictive capability, it is only possible to adjust the features considered by the model until the features that affect VapC activity are identified. As Chapters 5 and 6 have already discussed, there are many possible additions that could be made to the representations of RNA sites used in this study. It may be the case that a more detailed representation with features not considered in this study could be used to produce an accurate model. It may also be the case that a greater body of information is required; as it does not appear to achieve meaningful prediction with the data available to this study.

References

[1] J. L. McKenzie, J. Robson, M. Berney, T. C. Smith, A. Ruthe, P. P. Gardner, V. L. Arcus, and G. M. Cook, "A VapBC toxin-antitoxin module is a posttranscriptional regulator of metabolic flux in mycobacteria," *Journal of Bacteriology*, vol. 194, no. 9, 2012.

[2] R. D. Magnuson, "Hypothetical functions of toxin-antitoxin systems," *Journal of Bacteriology*, vol. 189, no. 7, pp. 6089–6092, 2007.

[3] J. Robson, J. McKenzie, R. Cursons, G. Cook, and V. Arcus, "The VapBC operon from mycobacterium smegmatis is an autoregulated toxin-antitoxin module that controls growth via inhibition of translation.," *Journal of Molecular Biology*, vol. 390, no. 3, pp. 353–367, 2009.

[4] D. Elliott and M. Ladomery, *Molecular Biology of RNA*. OUP Oxford, first ed., 2011.

[5] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*. Garland Publishing, Garland Science, 6th ed., 2015.

[6] K. Darty, A. Denise, and Y. Ponty, "VARNA: Interactive drawing and editing of the RNA secondary structure," *Bioinformatics*, vol. 25, no. 15, pp. 1974–1975, 2009.

[7] Y. Byun and K. Han, "PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots," *Bioinformatics*, vol. 25, no. 11, pp. 1435–1437, 2009.

[8] J. S. McCaskill, "The equilibrium partition function and base pair binding probabilities for RNA secondary structure," *Biopolymers*, vol. 29, pp. 1105–1119, 1990.

[9] M. Zuker, "On finding all suboptimal foldings of an rna molecule," *Science*, vol. 244, no. 4900, pp. 48–52, 1989.

[10] V. Juan and C. Wilson, "RNA secondary structure prediction based on free energy and phylogenetic analysis," *Journal of Molecular Biology*, vol. 289, no. 4, pp. 935–97, 1999.

[11] B. Knudsen and J. Hein, "Pfold: RNA secondary structure prediction using stochastic context-free grammars," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3423–3428, 2003.

[12] R. Lorenz, S. H. Bernhart, C. Hner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, "ViennaRNA package 2.0," *Algorithms for Molecular Biology*, vol. 6, no. 1, 2011.

[13] I. L. Hofacker, B. Priwitzer, and P. F. Stadler, "Prediction of locally stable RNA secondary structures for genome-wide surveys," *Bioinformatics*, vol. 20, no. 2, pp. 186–190, 2004.

[14] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, D. H. Turner, and I. Tinoco, "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 19, pp. 7287–7292, 2004.

[15] D. H. T. David H. Mathews, "NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure," *Nucleic Acids Research*, vol. 38, 2009.

[16] R. B. Lyngsø, *Automata, Languages and Programming: 31st International Colloquium, ICALP 2004, Turku, Finland, July 12-16, 2004. Proceedings*, ch. Complexity of Pseudoknot Prediction in Simple Models, pp. 919–931. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[18] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 3rd ed., 2011.

[19] R. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, pp. 63–91, 1993.

[20] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] D. Aha and D. Kibler, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991.

[23] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theor.*, vol. 13, no. 1, pp. 21–27, 1967.

[24] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Eleventh Conference on Uncertainty in Artificial Intelligence*, (San Mateo), pp. 338–345, Morgan Kaufmann, 1995.

[25] R. J. Quinlan, "Learning with continuous classes," in *5th Australian Joint Conference on Artificial Intelligence*, (Singapore), pp. 343–348, World Scientific, 1992.

[26] Y. Wang and I. H. Witten, "Induction of model trees for predicting continuous classes," in *Poster papers of the 9th European Conference on Machine Learning*, Springer, 1997.

[27] J. L. McKenzie, J. M. Duyvestyn, T. Smith, K. Bendak, J. MacKay, R. Cursons, G. M. Cook, and V. L. Arcus, "Determination of ribonuclease sequence-specificity using pentaprobes and mass spectrometry," *RNA*, vol. 18, no. 6, pp. 1267–1278, 2012.

[28] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, "Fast folding and comparison of RNA secondary structures," *Monatshefte für Chemie / Chemical Monthly*, vol. 125, no. 2, pp. 167–188, 1994.

[29] K. Zhang, L. Wang, and B. Ma, *Combinatorial Pattern Matching: 10th Annual Symposium, CPM 99 Warwick University, UK, July 22–24, 1999 Proceedings*, ch. Computing Similarity between RNA Structures, pp. 281–293. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999.

[30] J. McKenzie, J. Duyvestyn, T. Smith, K. Bendak, J. Mackay, R. Cursons, G. Cook, and V. Arcus, "Determination of ribonuclease sequence-specificity using pentaprobes and mass spectrometry.," *RNA*, vol. 18, no. 6, pp. 1267–1278, 2012.

[31] G. Blin, A. Denise, S. Dulucq, C. Herrbach, and H. Touzet, "Alignments of RNA structures," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 2, pp. 309–322, 2010.