Thesis

aPPRove: An HMM-based Method for Accurate Prediction of RNA-Pentatricopeptide Repeat Protein Binding Events

Submitted by

Thomas Harrison

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2015

Master's Committee:

Advisor: Christina Boucher

Asa Ben-Hur
Daniel Sloan

Abstract

aPPRove: An HMM-based Method for Accurate Prediction of RNA-Pentatricopeptide Repeat Protein Binding Events

Pentatricopeptide repeat containing proteins (PPRs) bind to RNA transcripts originating from mitochondria and plastids. There are two classes of PPR proteins. The P class contains tandem P-type motif sequences, and the PLS class contains alternating P, L and S type motif sequences. In this paper, we describe a novel tool that predicts PPR-RNA interaction; specifically, our method, which we call aPPRove, determines where and how a PLS-class PPR protein will bind to RNA when given a PPR and one or more RNA transcripts by using a combinatorial binding code for site specificity proposed by Barkan *et al.* [1].

Our results demonstrate that aPPRove successfully locates how and where a PPR protein belonging to the PLS class can bind to RNA. For each binding event it outputs the binding site, the amino-acid-nucleotide interaction, and its statistical significance. Furthermore, we show that our method can be used to predict binding events for PLS-class proteins using a known edit site and the statistical significance of aligning the PPR protein to that site. In particular we use our method to make a conjecture regarding a novel binding event between CLB19 and the second intronic region of *ycf3*.

The aPPRove web server can be found at www.cs.colostate.edu/~aPPRove and the software is available at that website for stand alone usage.

Acknowledgements

I would like to thank and express my gratitude to my advisor, Christina Boucher, for the help and guidance through the entirety of this project. Furthermore, I would like to thank Jaime Ruiz for building a front end web interface for the implemented software. I would also like to thank Asa Ben-Hur and Daniel Sloan for offering support, advice, and ideas throughout this project. I would like to thank Mark Heim and Sarah Morrison-Smith for many insightful comments. Finally, I would like to think Ian Small of The University of Western Australia for the suggestion of data sets to use as well insight on what types of analyses could be done with aPPRove.

This thesis is typset in LaTeX using a document class designed by Leif Anderson.

Table of Contents

List of Figures

CHAPTER 1

Introduction

Post-transcriptional control of RNA—which includes splicing, polyadenylation, and RNA editing—can have significant impact on the expression of a gene. One of the key factors that influence and contribute to post-transcriptional control of RNA is the availability and ability of specific proteins to bind to RNA. In short, RNA-binding proteins are those that bind to single- or double-stranded RNA and participate in forming ribonucleoprotein complexes. These complexes, in turn, exhibit a major role in post-transcriptional control of RNA [4, 5]. In this paper, we build a computational method for predicting where and how a family of RNA-binding proteins, the pentatricopeptide repeat (PPR), will bind to RNA. Our method, which we call *aPPRove*, builds upon the recent work of Barkan *et al.* [1] that determines sequence-specific binding rules for PPR proteins.

The primary structure of many RNA-binding proteins is composed of multiple repeats of a specific amino acid sequence, which recognize specific RNA sequences and/or structures [1, 6, 7, 2]. The repeated amino acid sequence is commonly referred to as a *motif sequence*. The length and the number of repetitions of a particular motif sequence varies widely across and within different classes of proteins [8]. Furthermore, these motif sequences can be highly degenerate. Nonetheless, there exists numerous computational methods that will determine the motif sequences and the number of repetitions of a motif sequence for a given RNA-binding protein, including: HMMer [9], TPRPred [10], and ScanProsite [11]. ScanProsite is used by aPPRove to determine the motif sequences.

As the name "pentatricopeptide repeat" suggests, this specific family of proteins is classified by the existence of tandem PPR motif sequences. PPR motif sequence are approximately

35 amino acids in length and are repeated any number of times [12]. These repeated motif sequences interact with a particular RNA binding site. See figure 1.1 for an illustration of the physical structure of a PPR protein and how it interacts with a binding site. PPR motif sequences are classified into three types based on motif length and composition. The most common type are the P motif sequences, which contain 35 amino acids. In comparison to the P type, the L motif sequences are slightly longer than P motif sequences, and the S motif sequences are slightly smaller than the P motif sequences. There are two classes of PPR proteins: P-class proteins only contain tandem P motif sequences, and PLS-class proteins contain alternating P, L and S motif sequences. The PLS class of proteins are predominantly involved in C-to-U RNA editing [8, 13].

The family of PPR proteins has a specific binding structure that relates to the secondary and tertiary structure of the protein. This is shown in Figure 1.1. Given the primary structure of a PPR protein, we denote the sixth amino acid of a PPR motif sequence as *position* 6 , and the first position of the next motif sequence as *position* $1'$, thus using the same notation for these sites which was used in Barkan *et al.* [1]. Hence, if there exists ℓ repeats of a motif sequence in a PPR protein then there are $\ell - 1$ adjacent positions specified by positions 6 and $1'$ in that protein. Fujii *et al.* [2] demonstrated that the amino acids at adjacent 6 and $1'$ positions show site specificity. Barkan *et al.* [1] demonstrated that these two sites work in combination to bind to a nucleotide in an RNA transcript. Therefore, the sequence-specific relationship can be cast as an alignment problem where the question is how an RNA sequence aligns to two amino acid sequences (defined by the adjacent 6 and $1'$ positions); this is with the constraint that every motif sequence must come in contact with a nucleotide in the RNA target. This is shown in Figure 1.1 (c), (d) and (e).

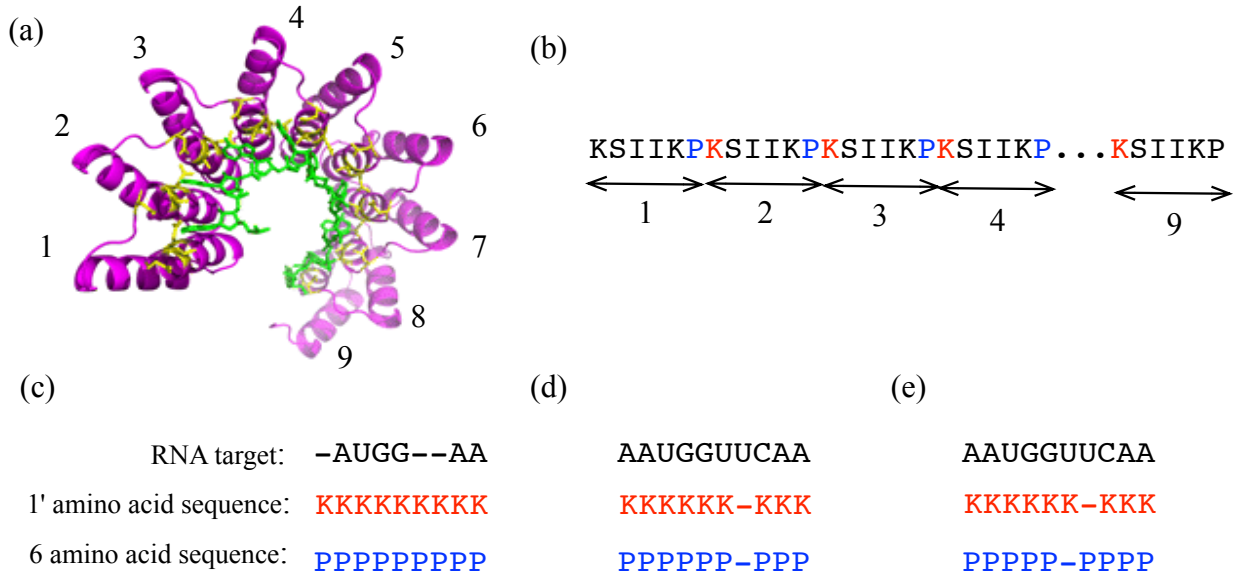

FIGURE 1.1. **Illustrating the connection between the tertiary structure of a PPR, the motif sequences, and the sequence-specific binding relationship.** (a) shows the physical structure of a PPR protein that has nine repeated motifs. This figures is from Fujii *et al.* [2]. The positions 6 and 1′ are on the internal face of each helix (shown in yellow). (b) shows the tandem motif sequences corresponding to the PPR protein in (a). Each of the repeated motifs are labelled. This is a simple fictitious example to illustrate the association between the tertiary structure of the PPR and the primary structure. The 6 and 1′ positions are highlighted in red and blue, respectively. (c) (d) and (e) show a possible binding between the amino acid sequence corresponding to the 1' positions (red), the amino acid sequence corresponding to the 6 positions (blue), and an RNA target.

PPR containing proteins are widespread throughout eukaryotes, but have undergone a large expansion in land plants. Approximately 450 different PPR encoding genes have been found in *Arabidopsis thaliana* and rice (*Oryza sativa*). These proteins have vital interactions with RNA transcripts in mitochondria and plastids [8]. Some are involved in RNA editing [13], others silence genes that encode for cytoplasmic male sterility (CMS) in flowering plants [14]. This is of particular importance since-male sterile plants are used to generate hybrid seed, which commercial agriculture heavily relies on because of its higher yield.

Our results show that aPPRove can be used to locate how and where a PPR protein belonging to the PLS class can bind to RNA and can be used to predict putative binding sites along an RNA transcript that a PPR protein is known to target. It takes a PPR protein (primary structure) and one or more RNA transcripts or RNA binding sites (which are also referred to as binding footprints) as input. It outputs the binding events that have highest statistical significance, and how the nucleotides in the RNA aligned to the amino acid pairs (defined by positions 6 and $1'$) in the PPR sequence for each binding. To the best of our knowledge, there do not exist any computational methods to predict where and how a PPR protein will bind a target RNA sequence when given a PPR sequence and one or more RNA transcripts. The statistical significance provided by aPPRove is based on the significance of the alignment of the PPR protein to its target in comparison to the alignments of the PPR to a database of transcripts. aPPRove harnesses the code provided by Barkan *et al.* [1] and implements a tailored hidden Markov algorithm to determine a score of each possible binding. Our experiments show that each of the PPR-RNA binding events presented in Barkan *et al.* [1] have high statistical significance using cross validation, which demonstrates the sensitivity and specificity of our approach; and that aPPRove is capable of detecting putative binding events. We believe our method will be a useful tool for determining novel PPR-RNA binding events; rather than solely relying on laboratory techniques, aPPRove could be used to greatly narrow the search for novel binding events.

CHAPTER 2

Related Work

The results of Barkan *et al.* [1] present a combinatorial binding code of PPR-RNA interaction that accounts for P and S motif sequences. They proposed the following combinatorial code: an amino acid pair with threonine at site 6 and aspartic acid at $1'$ will most likely bind to a guanine; an amino acid pair with threonine or serine at site 6 and aspargine at $1'$ will most likely bind to an adenine; an amino acid pair with asparagine at site 6 and aspartame acid at $1'$ will most likely bind to uracil; and an amino acid pair with asparagine at site 6 and asparagine or serine at $1'$ will most likely bind to cytosine. This binding code was expanded by the findings of Yagi *et al.* [7] and Takenaka *et al.* [6] who discovered binding preferences of L motif sequences. Both found that a proline at position 6 of an L motif sequence is likely to bind to uracil. Furthermore, the results of Takenaka *et al.* [6] showed that asparagine at position $1'$ of L motif sequences likely binds to adenine or uracil if it is paired with isoleucine, leucine, proline, threonine, or methionine at position 6. The model used for the three papers listed above involved aligning the PPR motifs of PLS proteins to the target RNA sequences such that the terminal S motif is positioned in contact with the nucleotide four base pairs upstream of an edit site on the target transcript. Okuda *et al.* [15] provides further evidence that PLS-class proteins align in this fashion. The pairing of positions 6 and $1'$ in the PPR protein reinforced the previous findings of Fujii *et al.* [2]. Lastly, the results of Kotera *et al.* [13] demonstrated that PLS-class proteins are required for RNA editing.

Prior computational work in predicting protein-RNA interaction has focused on determining the actual binding site in the primary structure of the protein or the RNA sequence

[9–11], developing protein-RNA interaction databases [16–18], and determining the likelihood that a particular protein will bind to an RNA [19–23].

The first computational method for predicting protein-mRNA interaction was proposed by Pancaldi and Bähler [19]. This method used Support Vector Machines (SVMs) and Random Forest (RF) classifiers to predict the likelihood of the interaction between an mRNA-binding protein and an mRNA. They used more than 1,000 features extracted from gene ontology terms, predicted secondary structures, mRNA properties, and genetic interactions. Two purely sequence-based approaches for predicting interaction likelihood were proposed by Muppirala *et al.* [21] and Wang *et al.* [22]. The method implemented in Muppiralla *et al.* [21] used RF and SVM classifiers to predict the probability of the interaction between an RNA-binding protein and RNA. It encoded the RNA sequences as normalized frequencies of tetrads. The protein sequences were encoded using a conjoined triad feature (CTF), and then used the amino acid composition and the nucleotide composition to predict the likelihood of one amino acid binding to a nucleotide. The method of Wang *et al.* [22] used a variation of CTF representation of protein descriptors and triads of the RNA sequence as RNA descriptors. These features were fed into both naïve Bayes and extended naïve Bayes classifiers.

A computational method specific to PPR-RNA interactions was presented in Yap *et al.* [23] where they predicted the recognition factor for an edit site on *atpF*. They aligned 6 and $1'$ for 193 known PLS-class editing factors in such a way that the terminal S motif aligned 4 base pairs upstream of the edit site and generated a score for each based on a table of log-likelihood ratios.

We note that all the methods predict the likelihood that a protein will bind to an RNA or mRNA molecule where as aPPRove predicts with how and where a PPR protein will bind to an mRNA using sequence-specific binding results.

CHAPTER 3

Algorithms and Methods

The aim of aPPRove is to build a predictive model of PPR-RNA binding using sequence-specific binding rules. This can be cast as an alignment problem. Let $S6$ and $S1'$ be the amino acid sequences defined by position 6 and position $1'$ of all adjacent motif sequences in the primary structure of a PPR protein S. If S contains ℓ adjacent motif sequences then $S6$ and $S1'$ both have length $\ell - 1$. Hence, given a PPR protein S, a RNA transcript R, and a scoring function ρ. More formally,

$$\rho(S6_i, S1'_j, R_k) : aa \times aa \times N \to \mathbb{R},$$

where $N = \{A, G, C, U, -\}$ and $aa = \{\text{all possible amino acids and } -\}$, where $-$ signifies an insertion or deletion. The goal is to find the w best alignments between R, $S6$ and $S1'$ with respect to ρ.

aPPRove can be broken down into five main steps: 1.) defining the repeat structure of the PPR by the motif sequence and number of repeats, 2.) constructing $S6$ and $S1'$, 3.) building a distribution of random alignments of $S6$ and $S1'$ to a database of RNA transcripts, 4.) aligning $S6$ and $S1'$ to one or more RNA target transcripts, 5.) and calculating the statistical significance (p-value) of the w best alignments of the PPR to target RNA transcripts.

3.1. PPR Motif Sequence Annotation

The PPR motif sequences are annotated using ScanProsite [24], which detects and annotates a protein sequence using the PROSITE database which contains signatures for various

protein families and subfamilies; each signature is defined as a regular expressions or weight matrix [25]. ScanProsite is used with the PPR signature of the PROSITE database specified. Next, the motif type is assigned for each motif by determining its length. Site specificity of the amino acid pairs varies according to the motif type. Motif sequences containing fewer than 35 amino acids are assigned as an S subtype, motif sequences containing 35 amino acids are assigned as a P subtype, and all others are assigned as an L subtype [8]. After the motif sequence annotation and type identification, $S6$ and $S1'$ were constructed from the two amino acids at position 6 and $1'$ in each motif sequence, respectively, and $S(6,\ 1')$ was formed from the pairs of amino acids from $S6$ and $S1'$ of the same motif sequence. For example, if $S1'$ consists of $DDND$, and $S6$ consists of the set $SSTS$, then $S(6, 1')$ will be $\{(DS), (DS), (NT), (DS)\}$.

3.2. Alignment of a PPR Sequence to an RNA Target

Next, we use a pair hidden Markov model to align $S(6,\ 1')$ to a target RNA sequence. It is tailored for semi-global alignment with six states: *start*, D_1, D_2, M, X, Y, and *end*. State M represents a match between an amino acid pair in $S(6,\ 1')$ and a target RNA nucleotide. States D_1, D_2 and X all represent a gap in the $S(6,\ 1')$ side of the alignment. D_1 represents a gap in $S(6,\ 1')$ before the occurrence of a single match state, and D_2 represents a gap after all match states have occurred. State X represents a gap internal to $S(6,\ 1')$, meaning a state M should occur on both sides of any state X. Using separate states for the three different types of gaps in the alignment allow for different transition probabilities leaving from D_1, D_2 and X. Having these varying probabilities is necessary for semi-global alignment using a pair hidden Markov model. State Y represents a gap in the R side of the alignment.

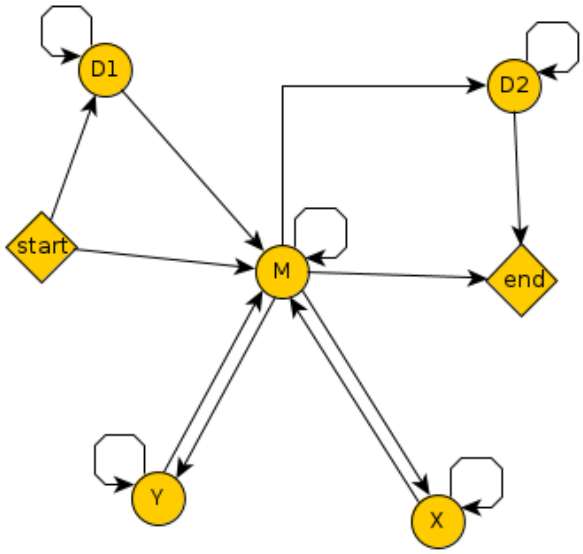

FIGURE 3.1. **Illustrating the pair hidden Markov model used.** Our pair hidden Markov model is tailored for semi-global alignment with six states: *start*, D_1, D_2, M, X, Y and *end*. State M represents a match between an amino acid pair in $S(6,\, 1')$ and a target RNA nucleotide in an RNA transcript. States D_1, D_2 and X all represent a gap on the $S6,\, 1'$ side of the alignment. State Y represents a gap in the RNA sequence. D_1 represents a gap in $S(6,\, 1')$ before the occurrence of a single match state and D_2 represents a gap after all match states have occurred. State X represents a gap internal to $S(6,\, 1')$, meaning a state M should occur on both sides of any state X.

We define a transition matrix $\mathbf{T}$ and emission matrices $\mathbf{A}$, $\mathbf{O}$, and $\mathbf{Q}$ in order to define our model. These matrices are constructed by using the binding events in figure S1 of Barkan *et al.* [1]; these binding events can be seen as alignments of the $S(6,\, 1')$ sequence of a PLS-class protein to its known RNA binding site. It is worth noting that these binding events —or alignments— are constructed by using previously identified binding rules. What results is a dataset that contains the frequency with which an amino acid pair binds to a specific nucleotide, as well as the frequency and location of insertions and deletions in the alignment. Hence, $\mathbf{T}$, $\mathbf{A}$, $\mathbf{O}$ and $\mathbf{Q}$ were defined using these data.

We now define some auxiliary variables that will be used for defining these matrices. First, we let n and m be equal to the length of R and $S(6,\, 1')$, respectively, and $F(\alpha, \beta)$

be the total number of times that state α transfers to state β, where α and β are in $\{M,\, X,\, Y,\, D_1,\, D_2,\, end,\, start,\}$. For example, $F(M, X)$ is equal to the number of times a gap follows a match in all the alignments obtained by Barkan *et al.* [1]. Let $G(i,\, j,\, k)$ be equal to the total number times the ith amino acid pair is witnessed binding to nucleotide j in motif type k. Lastly, we let γ and η be a set of pseudo-counts used for determining the probabilities for $\mathbf{T}$ and $\mathbf{A}$, respectively. We define $\gamma(i, j)$ for all possible i and j, where i and j are states in the pair hidden Markov model. The variables γ and η are similarly defined.

The 6×6 transition matrix $\mathbf{T}$ defines the probability of transitioning from any one state to any other state. More formally, we define $\mathbf{T}(\alpha,\, \beta)$ as the probability of state α transitioning to state β, where α is in $\{start,\, M,\, X,\, Y,\, D_1,\, D_2\}$ and β is in $\{M,\, X,\, Y,\, D_1,\, D_2,\, end\}$. It should be noted that our model does not allow for transitioning from the *end* state or transitioning to the *start* state. The transition probability of leaving state M or X and transitioning to any other state, *i.e.* $\mathbf{T}(M,\, \beta)$ and $\mathbf{T}(X,\, \beta)$ where β is in $\{M,\, X,\, Y,\, D_1,\, D_2,\, end\}$, are defined according to the following formula:

$$\frac{F(\alpha, \beta) + \gamma(\alpha, \beta)}{\sum_{\beta=1}^{6} \left(F(\alpha,\, \beta) + \gamma(\alpha,\, \beta)\right)}.$$

The probabilities of transitioning from *start*, D_1 and D_2 and going to any other state are dependent on n and m. Hence, $\mathbf{T}(D_1,\, M)$, $\mathbf{T}(D_2,\, end)$, and $\mathbf{T}(start,\, M)$ are defined to be equal to $1/((n-m)/2)$. Next, we define $\mathbf{T}(D_1,\, D_1)$, $\mathbf{T}(D_2,\, D_2)$, and $\mathbf{T}(start,\, D_1)$ as $1 - 1/((n-m)/2)$. We note that PLS-class proteins align in such a way that there will not be a transition to or from state X or state Y. This is because $S(6, 1')$ always aligns in a contiguous manner to its target site as shown in Barkan *et al.* [1] These two states were added for future flexibility in parametrizing the model for P-class proteins.

Since there are 20^2 possible amino acid pairs, four possible nucleotides, and three different types of PPR motif sequence, the emissions matrix $\mathbf{A}$ is of size $20^2 \times 4 \times 3$. The matrix $\mathbf{A}$ defines the emissions of state M. For example, $\mathbf{A}(IL,\, G,\, \mathsf{P})$ is the probability of witnessing the amino acid pair isoleucine (I) and leucine (L), binding to a guanine in a P motif sequence. The values for $\mathbf{A}$ were determined using the following formula:

$$\frac{G(i,\, j,\, k) + \eta(i,\, j,\, k)}{\sum_{\forall r} \sum_{\forall q} \sum_{\forall p} (G(p,\, q,\, r) + \eta(p,\, q,\, r))}.$$

The matrices $\mathbf{Q}$ and $\mathbf{O}$ have equal probability for all possible occurrences.

We use the Viterbi algorithm for pair hidden Markov models [26] to find the optimal alignment score according to probabilities assigned to our transition and emission parameters. Let $\mathbf{VD}$, $\mathbf{VM}$, $\mathbf{VY}$, and $\mathbf{VX}$ be three $n \times m$ dynamic programming matrices, where n is the number of pairs in $S(6,\, 1')$ and m is the length of R. The parameter w is provided as input by the user and causes the Viterbi algorithm to return the w optimal alignments according to the scoring scheme set by matrices $\mathbf{T}$, $\mathbf{A}$, $\mathbf{O}$, and $\mathbf{Q}$.

Upon the completion of the Viterbi algorithm $\mathbf{VD}$, $\mathbf{VX}$, $\mathbf{VY}$, and $\mathbf{VM}$ contains scores for all sub-alignments ending in state D_2, X, Y, and M, respectively. Every dynamic programming score is derived from the product of the score of the previous state, the probability of transitioning from the previous state, and the probability of the emission. The base case for this algorithm is as follows:

- Let $\mathbf{VD}(i,\, j) \wedge \mathbf{VX}(i,\, j) \wedge \mathbf{VY}(i,\, j) \wedge \mathbf{VM}(i,\, j) = -\infty$:

  $\forall (0 \leq i \leq n \wedge 0 \leq j \leq m)$

- Let $\mathbf{VD}(0,\, 0) = 1$

- Let $\mathbf{VD}(i,\,0) = \mathbf{VD}(i-1,\,0) \times \mathbf{T}(D_1,\,D_1) \times \mathbf{Q}(j)$:

  $\forall(0 < i \leq n)$

Matrices $\mathbf{VD}$, $\mathbf{VX}$, $\mathbf{VY}(i,\,j) \wedge$ $\mathbf{VM}$ are completed with the following recurrence relation for $\forall(0 < i \leq n \wedge 0 < j \leq m)$.

$$\mathbf{VM}(i,\,j) = \text{the } w \text{ max} \begin{cases} \mathbf{VD}(i-1,\,j-1) \times \mathbf{T}(D_1,\,M) \times \mathbf{A}(i,\,j) \\ \mathbf{VM}(i-1,\,j-1) \times \mathbf{T}(M,\,M) \times \mathbf{A}(i,\,j) \\ \mathbf{VX}(i-1,\,j-1) \times \mathbf{T}(X,\,M) \times \mathbf{A}(i,\,j) \\ \mathbf{VX}(i-1,\,j-1) \times \mathbf{T}(Y,\,M) \times \mathbf{A}(i,\,j) \end{cases}$$

$$\mathbf{VX}(i,\,j) = \text{the } w \text{ max} \begin{cases} \mathbf{VM}(i-1,\,j) \times \mathbf{T}(M,\,X) \times \mathbf{Q}(j) \\ \mathbf{VX}(i-1,\,j) \times \mathbf{T}(X,\,X) \times \mathbf{Q}(j) \end{cases}$$

$$\mathbf{VD}(i,\,j) = \text{the } w \text{ max} \begin{cases} \mathbf{VM}(i-1,\,j) \times \mathbf{T}(M,\,D_2) \times \mathbf{Q}(j) \\ \mathbf{VD}(i-1,\,j) \times \mathbf{T}(D_2,\,D_2) \times \mathbf{Q}(j) \end{cases}$$

$$\mathbf{VY}(i,\,j) = \text{the } w \text{ max} \begin{cases} \mathbf{VM}(i,\,j-1) \times \mathbf{T}(M,\,Y) \times \mathbf{O}(i) \\ \mathbf{VX}(i,\,j-1) \times \mathbf{T}(Y,\,Y) \times \mathbf{O}(i) \end{cases}$$

The scores of the w optimal alignments are found at $\mathbf{VD}(n,\,m)$ and $\mathbf{VM}(n,\,m)$. Traditional Viterbi decoding is used to obtain the sequence of states and hence the alignment associated with each of the w optimal scores. Each of the w optimal scores is normalized by summing up all transition and emission probabilities that correspond to transitioning to a

state M or X, subtracting $\mathbf{T}(D_1, M)$ from this total, and dividing this score by the length of the sub-alignment.

3.3. Statistical Significance of Scores

aPPRove returns a p-value for each of the w best alignments; this p-value statistic describes the probability of obtaining a normalized score that is at least as extreme as the one that was actually observed, assuming that the null hypothesis of a random alignment to a database of transcripts is true. In order to calculate p-value, we require a database of possible alignments. bey default, aPPRove considers all possible bindings to a database of plastid *Arabidopsis thaliana* transcripts. The set of *Arabidopsis thaliana* transcripts was obtained from the Phytozome website V9 [1]. By default, we align $S(6, 1')$ to each possible location in every transcript in the database and the targeted RNA transcripts for the given PPR sequence, which results in a normalized score of every position of every alignment (either to the an RNA in the database or the targeted RNA). By eye, these scores follow a normal distribution for all the PPR-RNA bindings we consider. For example, Figure 3.2 illustrates the distribution of aligning the protein, MEF26 to it's target database. The p-value of the is calculated using the null hypothesis that the normalized score is equal to the mean of the distribution.

aPPRove uses the *Arabidopsis thaliana* plastid transcripts by default, however, any user-defined database of RNA transcripts can be specified. If run with a custom database, aPPRove will provide the p-values of the w optimal normalized scores by using the normal distribution of normalized scores of aligning S to the database. In addition, it is possible to run aPPRove without using any database (default or otherwise). In this case, aPPRove

[1] http://phytozome.jgi.doe.gov/pz/portal.html

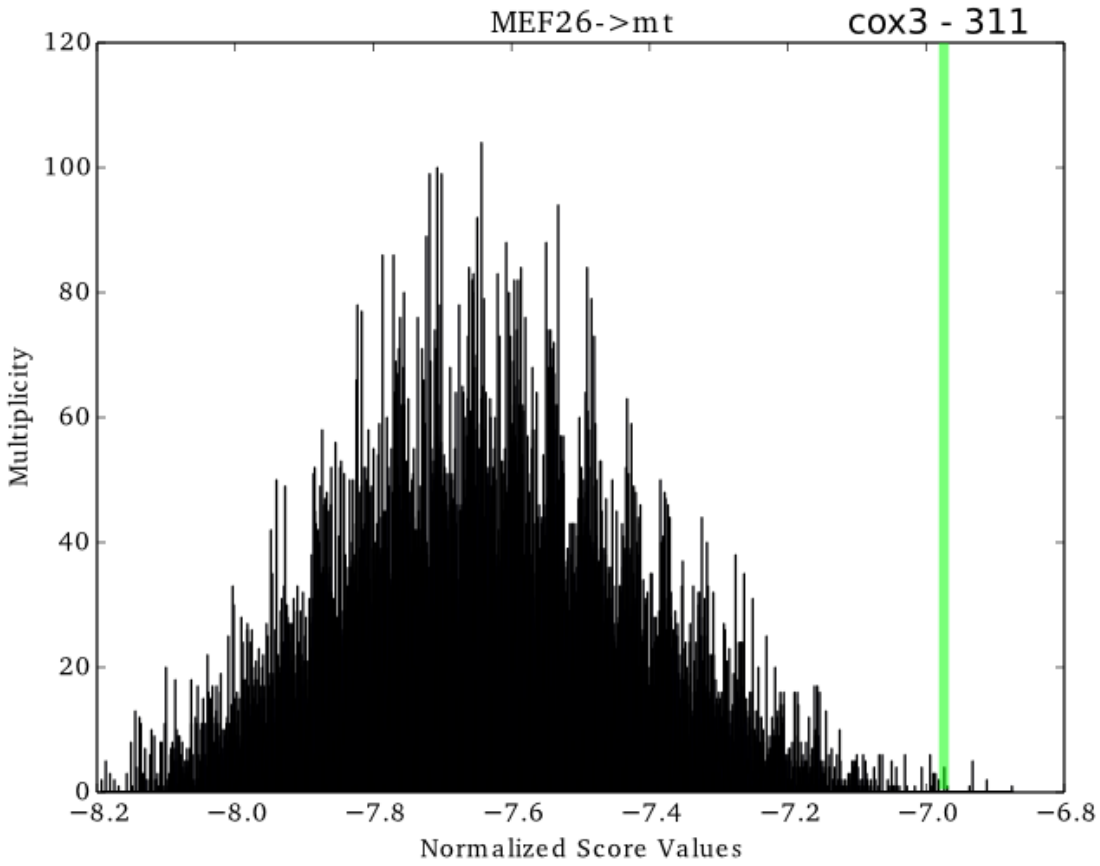

FIGURE 3.2. **Demonstrating the distribution of normalized scores built by taking all possible alignments of a PLS protein to its respective target database.** The normalized score of aligning the protein to its binding footprint is expected be located far on the extreme right end of the distribution. The green line indicates where the score of aligning MEF26 to its known binding site on *cox*3 is located on the distribution generated by aligning the $S(6, 1')$ sequence of MEF26 to the target database.[3]

outputs the normalized scores of the w optimal alignments and the details of the alignments but no p-values.

CHAPTER 4

Results

4.1. Data

We used the dataset from figure S1 of Barkan *et al.* [1] to parameterize and evaluate the performance of our model. This dataset is composed of 30 PLS-class proteins and their known binding footprint RNA sequences. Because some proteins bind to multiple targets, there is a total of 55 instances of a PPR protein paired with a known binding footprint. All of these proteins target transcripts originating from either mitochondria or plastids. Of the 30 proteins, 27 are from *Arabidopsis thaliana*, two are from moss (*Physcomitrella patens*), and one is from rice (*Oryza sativa*). Protein sequences from this dataset were extracted from either GenbBank [27] or Uniprot [28]. However, PpPPR56, PpPPR71, PpPPR78 and PpPPR79 were not used for the evaluation since they were only available as sequence fragments. PRR2263 was not used since it is only available as a hypothetical sequence, and MEF14 was not used because we were not able to find PPR motif sequences using PrositeScan.

4.2. Statistical Analyses of Aligning Proteins to Their Target Footprints

It is expected that the normalized score of aligning $S(6, 1')$ of each protein to its known binding site should generally be larger than a random alignment of that protein to the transcripts originating from the organelle it targets. In order to find the statistical significance of a PPR protein binding to its known binding footprint, we compared the score of aligning $S(6, 1')$ of each PPR protein to their own binding site against every possible contiguous alignment of $S(6, 1')$ to a database of transcripts. Two databases were used for this investigation. One database consisted of all transcripts from the *Arabidopsis thaliana* plastid and the other

consisted of all transcripts from the *Arabidopsis thaliana* mitochondrion. We selected the database to use for each run based on what type of organelle transcripts that particular protein targets. We evaluated our method by using Leave One Out cross validation (LOO) for each PPR and RNA binding site pair. Thus, for each pair, we parametrized the hidden Markov model using all other pairs except the one being evaluated, ran the trained model on the pair that was removed, and determined the normalized score for the pair of interest. Using the transcript database, a p-value for each PPR and RNA binding site pair was found using its normalized score and then adjusted using the Benjamini Hochberg method [29]. As shown in figure 4.1 the median for all 55 p-values is approximately 0.013, and there are two adjusted p-values above 0.1 which belong to MEF1 targeting *nad2* and CRR28 targeting *ndhD*. These two p-values are unsurprising because the alignments in Barkan *et al.* [1] show that the these two proteins align to their target sites in such a way that amino acids pairs with high site specificity are not paired with their preferred nucleotide. This includes two out of the six amino acid pairs with high site specificity in the alignment of CRR28 to *ndhD* and three out of the six amino acid pairs with high site specificity in the alignment of MEF1 to *nad2*. Thus, this validates our approach and demonstrates that the known PPR and RNA binding pairs can be identified by considering the extreme values of the distribution.

To find the false positive rate (FPR) for each of the 55 PPR and RNA binding footprint pairs, we compared the score of aligning $S(6, 1')$ of each PPR protein to its known binding site against every possible contiguous alignment of $S(6, 1')$ to a database of decoy transcripts. The two decoy databases were created by generating a random permutation for each transcript from the target database. Like the analyses using the target databases, we evaluated our method by using LOO for each PPR and binding site pair. For each pair the

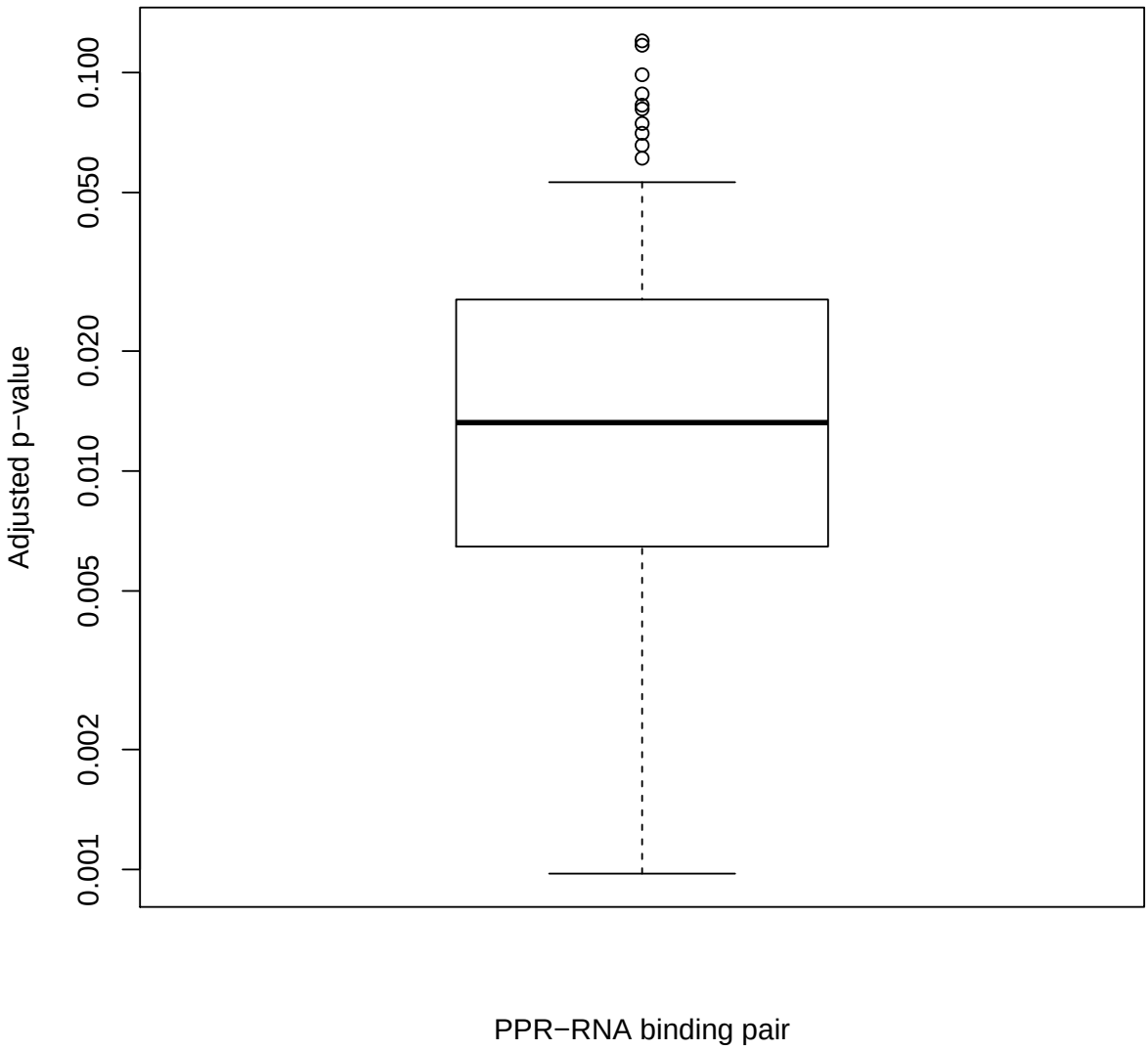

FIGURE 4.1. **Demonstrating a boxplot of the 55 Benjamini Hochberg adjusted p-values of the normalized score of aligning the proteins to their known binding sites against the target database.** The median of all p-values is 0.013.

FPR was calculated by the ratio of the number of alignments to the decoy database that had a normalized score greater than or equal to the score of aligning the PPR to its binding footprint over the total number of alignments to the decoy database. As shown in figure 4.2, the median for these ratios was 0.0076 with a range of 0.12.

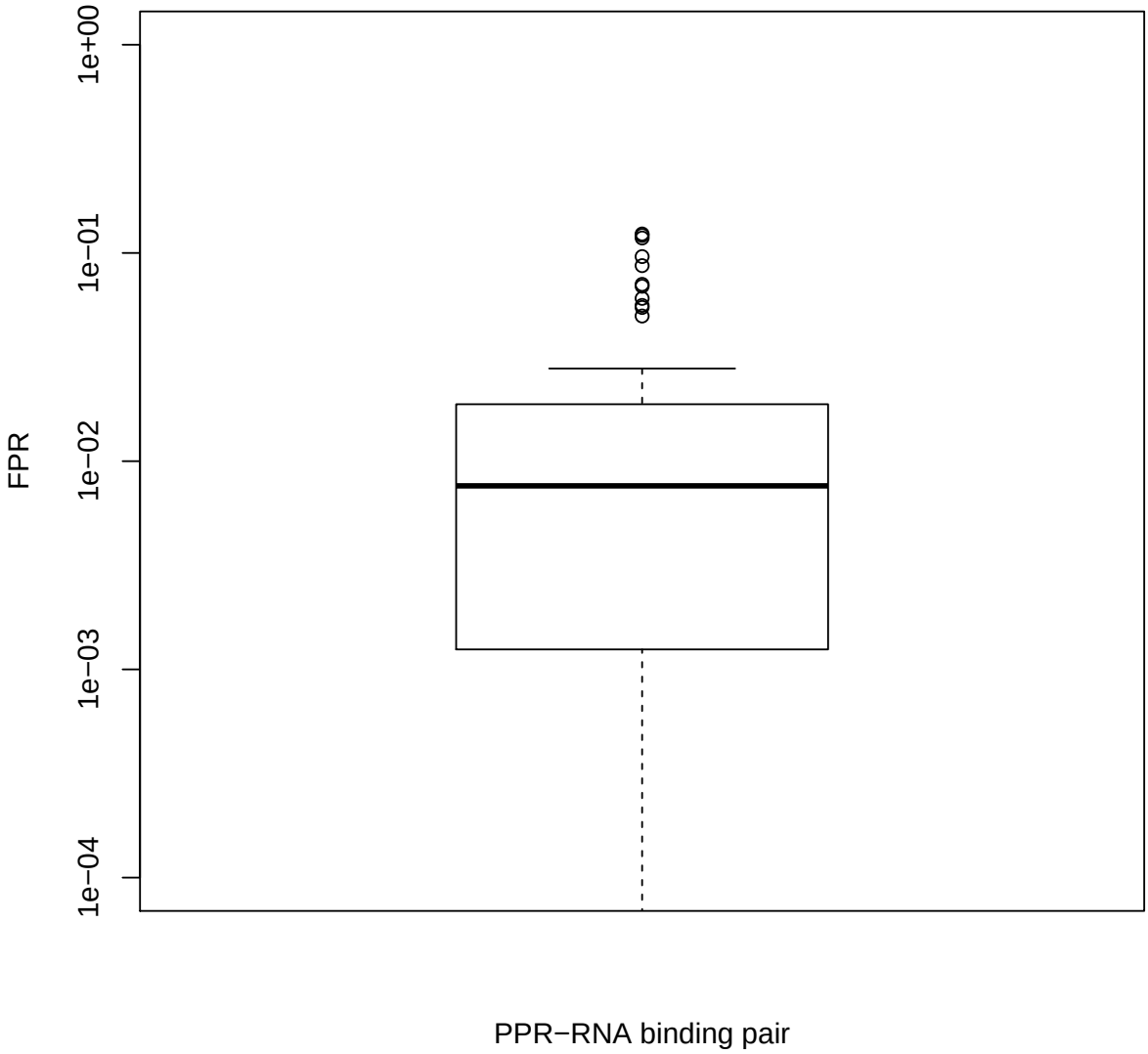

FIGURE 4.2. **Demonstrating the FPR of all 55 PPR-RNA pairs.** We compared the score of aligning $S(6, 1')$ of each PPR protein to their own binding site against every possible alignment to a database of decoy transcripts. The median FPR for all pairs is 0.0076, and the range is 0.12.

Site-specific RNA editing factors continue to be discovered at a rapid rate, including many that have been identified since the dataset that we used to train our model was compiled [1]. For example, Arenas-M. *et al.* [3] demonstrated in the absence of MEF26, *cox3*-311 editing is completely abolished and *nad4*-166 is only partially edited. Using aPPRove, we confirmed

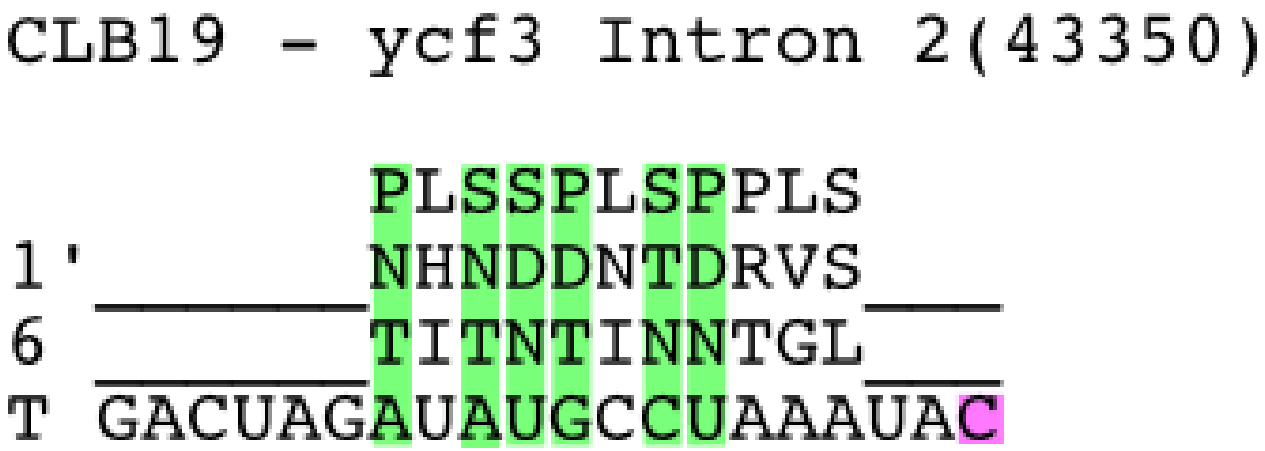

FIGURE 4.3. **Demonstrating the putative binding event of CLB19 and *ycf3*.** This figure shows the alignment of the putative binding event of CLB19 and the binding site located upstream of the edit site at position 43,350 of the *Arabidopsis thaliana* plastid genome. Pairs highlighted in green are considered to be statistically correlated amino acid-nucleotide pairs as specified by Barkan *et al.* [1]. The C highlighted in magenta is the edit site of the binding footprint.

that the two predicted binding sites with the alignment ending four base pairs upstream of the two edit sites were both among the top 41 hits out of 66,500 possible alignments in the mitochondrial target database with p-values less than 0.0005.

4.3. Binding Event Prediction Using Previously Discovered Edit Sites

aPProve can be used to predict binding events when an edit site is discovered but its editing factor is unknown. To detect putative binding events, we aligned the 12 PPR proteins known to target the *Arabidopsis thaliana* plastid from our data set to one of the nine minor binding sites [30] found at genomic position 43,350 located in the second intronic region *ycf3*. We sampled the sequence 30 base pairs upstream of the edit site and aligned all 12 PPR proteins to it. Of these proteins, CLB19 had the lowest p-value at 0.00005 and aligned to this target site in such a way that all 6 amino acid pairs with high site specificity aligned to its preferred nucleotide (see figure 4.3). Given the low p-value as well as the distance from the edit site, we predict that CLB19 is the editing factor for this edit site.

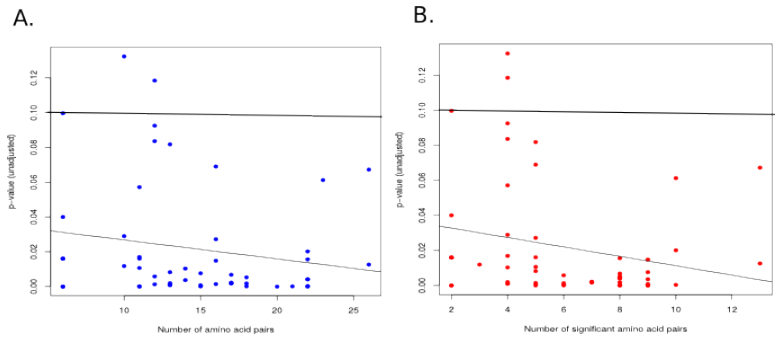

FIGURE 4.4. **Demonstrating the adjusted p-values with respect to: A.) the total numbers of motif binding pairs in the protein, and B.) the total number of motif binding pairs that have statistically significant site preference according to Barkan *et al.* [1].** The regression lines on both plots demonstrate that there is an anti-correlation between the number of amino acid binding pairs and p-value. Graph A has a Pearsons Correlation sample estimate of -0.335024 with a p-value of 0.01241. Graph B has a Pearsons Correlation sample estimate of -0.3978517 with a p-value of 0.00263

4.4. Factors influencing aPPRove's predictability

We note that aPPRove is more successful in predicting the binding of PPR proteins with a larger number of motif sequences than proteins with a smaller number of motif sequences. A smaller number of motifs will likely result in more false positives due to the existence of fewer amino acid pairs that show preference in the $S(6, 1')$ sequence. Figure 4.4 demonstrates the adjusted p-values with respect to the total numbers of motif binding pairs in the protein, as well as the total number of motif binding pairs that have statistically significant site preference according to Barkan *et al.* [1]. The regression line on both plots demonstrate that there is a negative correlation between the number of amino acid binding pairs and p-values. Hence, the experiments indicate that aPPRove can successfully locate how and where binding PLS-class of PPR proteins will bind to it target transcript and also provide the statistical significance of the binding.

CHAPTER 5

Conclusion

In this paper, we presented a method that used the primary binding code of PPR proteins to predict how a protein will bind to a target transcript or binding footprint. Our method is unique in that it can be used to detect where and how a PPR protein binds to an RNA as opposed to assessing the likelihood of interaction. Again, we note that the hidden Markov model was parametrized with a dataset involving protein-RNA interactions of only PLS-class proteins, thus aPPRove captures the intricacies of how the PLS class of PPR proteins bind to their target, but it may not accurately portray how a P-class PPR protein will bind to its target. The lack of data regarding P-class PPR protein interactions prevent us from parametrizing the model specifically for this subfamily of proteins. It is possible that the onset of high throughput methods of quantifying protein-RNA interaction [31] may allow for future progress in modelling the interaction of P-class proteins and their target transcripts. Finally, if there is a known edit site, aPPRove can be used to detect putative binding events. Detecting these events is one of the most beneficial and powerful uses of aPPRove.

Lastly, we note that the data used for our investigation was compiled from a number of experimental techniques that are not high-throughput. One commonly used technique is a gel mobility shift assay. This involves mixing the RNA-binding protein with a short RNA sequence and running the sample through a gel. If the RNA is bound to the protein it will run slower because of the larger size. If not, it will quickly run through the gel. Using this technique allows for the separation of bound and unbound RNA molecules. Performing this experiment on many different RNAs can narrow down the necessary window for binding. Prikryla *et al.* [32] demonstrated other methods that are specific to PPR proteins. Although

these techniques are not high-throughput, there is evidence that such methods are on the horizon. In 2014, Tome *et al.* [31] developed a high-throughput sequencing-RNA affinity profiling assay by adapting a high-throughput genome sequencer to quantify the binding of a protein to millions of RNAs. As high-throughput methods become more commonplace, larger and more datasets will become available. aPPRove is one method that can be easily adapted with forthcoming data and thus be used to predict the binding of other families and subfamilies of proteins.

Bibliography

[1] A. Barkan, M. Rojas, S. Fujii, A. Yap, Y. Chong, C. Bond, and I. Small, "A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins," *PLoS Genetics*, vol. 8, pp. 1509–1512, 2012.

[2] S. Fujii, C. Bond, and I. Small, "Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution.," *Proceedings of the National Academy of Sciences*, vol. 108, no. 4, pp. 1723–1728, 2011.

[3] A. Arenas-M, A. Zehrmann, S. Moreno, M. Takenaka, and X. Jordana, "The pentatricopeptide repeat protein MEF26 participates in RNA editing in mitochondrial cox3 and nad4 transcripts.," *Mitochondrion*, vol. 19, no. B, pp. 126–134, 2014.

[4] D. Hogan, D. Riordan, A. Gerber, D. Herschlag, and P. O Brown, "Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system," *PLoS Biology*, vol. 582, pp. 1997–1986, 2008.

[5] T. Glisovic, J. Bachorik, J. Yong, and G. Dreyfuss, "RNA-binding proteins and post-transcriptional gene regulation," *FEBS Letters*, vol. 6, p. e255, 2008.

[6] M. Takenaka, A. Zehrmann, A. Brennicke, and K. Graichen, "Improved computational target site prediction for pentatricopeptide repeat RNA editing factors," *PloS ONE*, vol. 8, p. e65343, 2013.

[7] Y. Yagi, S. Hayashi, K. Kobayashi, T. Hirayama, and T. Nakamura, "Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants," *PloS ONE*, vol. 8, no. 3, p. 1, 2013.

[8] C. Lurin *et al.*, "Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis," *Plant Cell*, vol. 16, no. 8, pp. 2089–2103, 2004.

[9] R. Finn, J. Clements, and S. Eddy, "HMMER web server: interactive sequence similarity searching," *Nucleic Acids Research*, vol. Web Server Issue, no. 39, pp. W29–W37, 2011.

[10] M. Karpenahalli, A. Lupas, and J. Söding, "TPRpred: a tool for prediction of TPR-, PPR- and SEL1-like repeats from protein sequences," *BMC Bioinformatics*, vol. 8, no. 2, 2007.

[11] E. de Castro *et al.*, "ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins," *Nucleic Acids Research*, vol. 34, no. Web Server issue, pp. W362–W365, 2006.

[12] I. Small and N. Peeters, "The PPR motif - a TPR-related motif prevalent in plant organellar proteins," *Trends in Biochemical Sciences*, vol. 25, no. 2, pp. 46–47, 2000.

[13] E. Kotera, M. Tasaka, and T. Shikanai, "A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts," *Nature*, vol. 433, pp. 326–330, 2005.

[14] Z. Wang *et al.*, "Cytoplasmic male sterility of rice with Boro II cytoplasm is caused by a cytotoxic peptide and is restored by two related PPR motif genes via distinct modes of mRNA silencing," *Plant Cell*, vol. 18, no. 3, pp. 676–687, 2006.

[15] K. Okuda, H. Shoki, M. Arai, T. Shikanai, I. Small, and T. Nakamura, "Quantitative analysis of motifs contributing to the interaction between PLS-subfamily members and their target RNA sequences in plastid RNA editing," *The Plant Journal*, vol. 80, pp. 870–882, 2014.

[16] B. Lewis, R. Walia, M. Terribilini, J. Ferguson, C. Zheng, V. Honavar, and D. Dobbs, "PRIDB: a protein-RNA interface database," *Nucleic Acids Research*, vol. 39, no. suppl. 1, pp. D277–D282, 2011.

[17] S. Fujimori, K. Hino, A. Saito, S. Miyano, and E. Miyamoto-Sato, "PRD: A protein-RNA interaction database," *Bioinformation*, vol. 8, pp. 729–730, 2012.

[18] T. Wu *et al.*, "NPInter: the noncoding RNAs and protein related biomacromolecules interaction database," *Nucleic Acids Research*, vol. 34, no. Database issue, pp. D150–D152, 2006.

[19] V. Pancaldi and J. Bähler, "In silico characterization and prediction of global protein-mRNA interactions in yeast," *Nucleic Acids Research*, vol. 39, no. 14, pp. 5826–5836, 2011.

[20] M. Bellucci, F. Agostini, M. Masin, and G. Tartaglia, "Predicting protein associations with long noncoding RNAs," *Nature Methods*, vol. 8, pp. 444–445, 2011.

[21] U. Muppirala, V. Honava, and D. Dobbs, "Predicting RNA-protein interactions using only sequence information," *BMC Bioinformatics*, p. 489, 2011.

[22] Y. Wang *et al.*, "De novo prediction of RNA-protein interactions from sequence information," *Molecular BioSystems*, vol. 9, no. 1, pp. 133–142, 2013.

[23] A. Yap, P. Kindgren, C. Colas des Francs-Small, T. Kazama, S. Tanz, K. Toriyama, and I. Small, "AEF1/MPR25 is implicated in RNA editing of plastid atpF and mitochondrial nad5 and also promotes atpF splicing in *arabidopsis* and rice," *The Plant Journal*, vol. 81, no. 5, pp. 661–669, 2014.

[24] A. Gattiker, E. de Castro, and E. Gasteiger, "ScanProsite: a reference implementation of a PROSITE scanning tool," *Applied Bioinformatics*, vol. 1, no. 2, pp. 107–108, 2002.

[25] A. Bairoch, "PROSITE: a dictionary of sites and patterns in proteins.," *Nucleic Acids Research*, vol. 20, no. Supplement, pp. 2013–2018, 1991.

[26] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, "Biological Sequence Analysis," *Cambridge University Press; 1st edition*, 1998.

[27] D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, and D. Wheeler, "GenBank," *Nucleic Acids Research*, vol. 31, no. 1, pp. 23–27, 2003.

[28] The UniProt Consortium, "Activities at the Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 42, pp. D191–D198, 2014.

[29] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, vol. B 57, pp. 289–300, 1995.

[30] H. Ruwe, B. Castandet, C. Schmitz-Linneweber, and D. Stern, "*Arabidopsis* chloroplast quantitative editotype," *FEBS Letters*, vol. 587, no. 9, pp. 1429–1433, 2013.

[31] J. Tome, A. Ozer, J. Pagano, D. Gheba, G. Schroth, and J. Lis, "Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling," *Nature Methods*, vol. 11, pp. 683–688, 2014.

[32] J. Prikryla, M. Rojasa, G. Schusterb, and A. Barkan, "Mechanism of RNA stabilization and translational activation by a pentatricopeptide repeat protein," *Proceedings of the National Academy of Sciences*, vol. 108, no. 1, pp. 415–420, 2010.