UNIVERSITY OF CAMBRIDGE

DEPARTMENT OF ENGINEERING

A Cross-Genome Study of the Pentatricopeptide Repeat Protein

by

Haydn King (JE)

Fourth-year undergraduate project in

Group F, 2012/2013

I hereby declare that, except where specifically indicated, the work submitted herein is my own original work.

Signed: _____ Date: _____

[This page is intentionally left blank]

A Cross-Genome Study of the Pentatricopeptide Repeat Protein

by

Haydn King (JE)

Fourth-year undergraduate project in Group F, 2012/2013

Technical Abstract

Synthetic biology is a new and developing engineering field which attempts to apply proven engineering concepts and techniques such as standardisation, characterisation and encapsulation to biology in order to design new and useful forms of life.

The field often makes use of analogies with electrical engineering – in synthetic biology, well characterised components are combined and interact together in biological circuits. Circuits have been designed to perform a wide variety of tasks, such as the production of useful chemicals, detecting the presence of harmful toxins or producing bioluminescent light.

Since designing novel proteins from scratch is beyond our technological reach at present, progress in synthetic biology usually results from the characterisation of natural systems found in one organism which are then applied in different situations.

This project investigates such a system commonly found in plant cells, where the cell nucleus is thought to control gene expression in the chloroplasts and other organelles using signals encoded by members of the Pentatricopeptide Repeat (PPR) protein family. This system is particularly interesting for synthetic biology as there is a deficit of well characterised parts for accurately controlling the production of an arbitrary protein. PPRs have been shown to be able to do this naturally – by binding with targeted mRNA molecules in the chloroplast, they are able to either increase of decrease expression levels of particular proteins. It is hoped that characterisation of this class of proteins will allow the design of PPRs which can control expression of an arbitrary protein.

In addition to their development as a tool for synthetic biology, an understanding of the interactions between PPRs and mRNA will be vital to understanding the control mechanisms at work in the chloroplast, which is a necessary step if we wish to exploit the large potential for biosynthesis within the chloroplast.

The first part of the project is concerned with automating the detection and annotation of naturally occurring PPRs in existing genomes. While PPRs have been discovered in a wide variety of plants, a specific, clearly defined and repeatable method for their detection and localisation has not. An algorithm for reliable and repeatable PPR discovery was developed in the first part of the project.

In the second part, this algorithm was applied to a large variety of plant genomes in order to assess the abundance of PPR proteins. It was found that the majority of the plants surveyed had a similar number of PPRs (between 400 and 600). It is hypothesised that this is due to a considerable conservation of function of PPRs across genomes.

Each member of the PPR family binds to a particular RNA sequence, but attempts to directly predict the RNA binding sequence of a given PPR were thwarted by the limited number of well characterised PPRs. However, the binding sites of a set of characterised PPRs was identified in the model organism *A. thaliana*. Likely homologs of the genes containing these binding sites were then found in the chloroplast genomes of numerous other plants and regions of very high sequence similarity to the original binding sites were found, supporting the hypothesis that PPR function is highly conserved across different species.

The final part of the project studied the usefulness of PPRs in the context of synthetic biology. If, given a target binding sequence, we could design a PPR protein which would bind to this sequence then it was established that it should be possible to construct logic gates using known PPR-RNA interactions.

The fact that each each PPR would act as an isolated signalling modality has the potential to revolutionise the field as in most projects problems due to cross-talk of reporter molecules fundamentally limit the complexity of the circuit.

Contents

[This page is intentionally left blank]

Chapter 1

Introduction

This project investigates the newly discovered family of pentatricopeptide repeat (PPR) proteins, which are vital to plant biology and could become an exciting new tool in the field of synthetic biology. The project spans the space between engineering and the life-sciences and this section provides an introduction to the relevant fields and the motivation behind the project.

1.1 Molecular Biology

Molecular biology is the study of the molecular basis of biology. It is mostly concerned with the understanding of the systems and processes that occur within a living cell. Naturally, the field overlaps considerably with other areas, such as genetics (the study of genes and heredity) and biochemistry (the study of the chemical processes of life).

While the field itself is rather broad, much of it is underpinned by what is referred to as the central dogma of molecular biology – DNA makes RNA makes proteins. This central dogma describes the flow of information within a cell and the mechanisms which regulate this flow. Many of these processes are highly complicated and remain poorly understood, but much progress has been made since the discovery of DNA in the 1950s to understand these mechanisms. Figure 1.1 shows the most important of these and how they convert between the three most important classes of molecules in the cell.

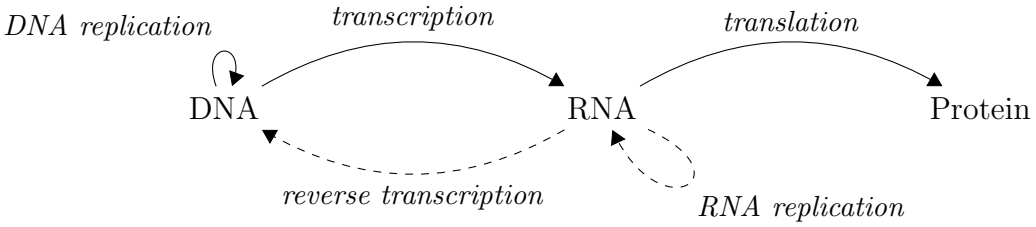

Figure 1.1: **The main processes in molecular biology.** The three most common are shown using solid lines while two important but less common processes are shown in dotted lines.

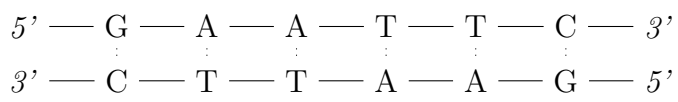

Figure 1.2: **An example of a DNA fragment.** Covalent bonding along the strands is shown with solid lines and hydrogen bonding between the two strands is shown using dotted lines. In this case the forward and reverse strand happen to be palindromic, although this is by no means necessary. Note that the decision as to which strand is considered the forward and which is the reverse is arbitrary – in fact other naming schemes (such as Watson/Crick) also exist.

Molecules of DNA are the cell's long term storage mechanism – recent research estimates the half-life of DNA to be 521 years (Allentoft *et al.*, 2012). DNA molecules are long sequences of simple molecules called nucleotides, the sequence of which encodes all the genetic information of the cell. Each nucleotide contains a nucleobase which is either Adenine, Cytosine, Guanine or Thymine (A, C, G or T) and it is the sequence of bases which determines the information content of the molecule. The nucleotides are linked together in a strand which is only read in one direction, known as the 5' to 3' direction.

These strands do not normally exist independently, but instead form hydrogen bonds with each nucleobase in a complementary strand, forming double stranded DNA. The sequence of the complementary strand is determined by the hydrogen bonds formed between bases – A binds with T, G binds with C. The strands bind such that they are read in opposite directions – the 5' to 3' direction on the reverse strand is in the reverse direction to that of the forward strand – hence converting a sequence from the forward to reverse strand is referred to as a *reverse complement*. The two strands are coiled around each other into DNA's characteristic double-helix structure.

The data stored in DNA is read by a molecule called RNA polymerase which produces a single stranded RNA copy of a section of the DNA in a process called *transcription*. RNA is similar to DNA, but is short-lived (lasting minutes to hours) and so the RNA copy is referred to as a messenger-RNA (mRNA) molecule. This message is then read by a ribosome, a molecule which translates the mRNA into a protein in a process referred to as *translation*. Proteins are linear chains of amino acids linked by flexible joints which fold into a very specific shape and perform many important functions within the cell. The region of DNA which codes for a particular protein is called a *gene*.

The processes of transcription and translation through which genes are expressed (produce proteins) are typically very tightly controlled by the cell, as this is the main way of influencing the levels of various proteins within the cell and thus the cell's overall activity.

1.1.1 Transcription

Both DNA and RNA have an alphabet of four symbols and so during transcription DNA's alphabet, $\{A, C, G, T\}$, is mapped directly to that of RNA, $\{A, C, G, U\}$, where thymine is replaced with uracil. Transcription is a one-to-on mapping and is thus bijective, and indeed a less common process called reverse-transcription performs the inverse mapping from RNA to DNA.

Transcription does not act on an entire DNA strand at once but instead transcribes a subsequence of the DNA called a transcription unit, which contains one or many genes. These units are marked by promoters which are regions of DNA upstream of the transcription unit that initiate transcription by promoting RNA polymerase binding. Modulating promoter activity in response to the concentration of another molecule is a common control motif. Transcription units are terminated by terminator regions, which cause the RNA polymerase to cease transcription and release the mRNA.

Transcripts don't just contain genes, they usually include non-coding regions at either end called the 5' and 3' untranslated regions (UTRs) respectively. They can also contain other non-coding regions called *introns* which are often found within gene sequences and are removed from the RNA message in a process called RNA splicing before translation. Introns do not contain any useful sequence and tend to complicate matters significantly as efforts to predict their location accurately and reliably have thus far failed – they are best detected using reverse transcription.

1.1.2 Translation

Translation is the process by which an RNA message is converted into a protein. In higher cells (eukaryotes), mRNA undergoes further processing and is exported from the nucleus before translation while in lower organisms (prokaryotes) translation begins immediately after transcription.

Proteins are a sequence of amino acids, where each acid comes from an alphabet of 20 amino acids. Each acid is coded for by 3 bases of RNA, which are referred to collectively as a codon. Since there are 4^3 possible codons and only 20 amino acids, the code is overcomplete – several different codons map to the same amino acid – and thus a definitive reverse mapping from protein to DNA sequence is impossible. As well as coding for amino acids, special codons mark the start and end of a protein. A start codon (AUG) marks the beginning of a protein and one of the three stop codons (UAG, UAA and UGA) terminate the translation of the protein.

In translation, molecules called ribosomes bind to the mRNA, reading the sequence 3 bases (one codon) at a time and constructing the appropriate protein until a stop codon is found, when the ribosome detaches and releases the protein. The point where the ribosome binds is called a ribosome binding sequence (RBS), the consensus is termed the Shine-Dalgarno sequence, which is found a few bases upstream of the start codon.

Because of the triplet nature of translation, proteins are sensitive to frame shifts – since bases are read three at a time, if the reading frame is shifted by one or two bases, the resulting amino acid sequence is effectively unrelated to the encoded one. This can be the case if an unexpected intron is present as introns need not be multiples of three codons long.

mRNA is more fragile than DNA but is also targeted by exonucleases, a class of enzyme which degrade RNA molecules, preventing the production of more protein. Similar processes exist which degrade proteins over time, recycling their amino acids to form new proteins. These degradation processes mean that a gene must continue to be *expressed* (transcribed and translated) at a constant rate for the concentration of its protein to remain constant.

1.1.3 Controlling Expression

Control of protein production is typically achieved using several layers of control at different stages. For example, the lac operon controls the production of enzymes which allows the cell to metabolise lactose, a carbon source. It is energetically favourable for the cell to directly metabolise glucose if it is available as lactose is harder to process, and so the cell can save energy by only turning on its lactose processing machinery when

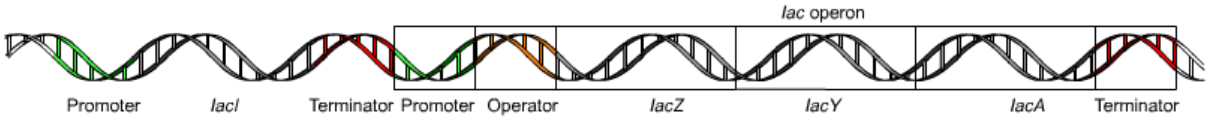

Figure 1.3: **Annotated impression of the lac operon.** (not to scale) It contains two transcription units and a total of four genes. The first (leftmost) unit contains the *lacI* gene and is expressed constitutively (continuously). The protein which is produced is known as the lac repressor, and in the absence of lactose it binds tightly to the operator region, preventing transcription of the second transcriptional unit. However, when lactose is present outside the cell, a small amount will diffuse across the cell wall and into the cell, where it binds with the lac repressor, preventing it from binding to the operator and thus allowing transcription of the second unit to begin.

Of the three genes that are then expressed, two are directly relevant. *lacY* encodes a membrane protein which actively pumps more lactose into the cell, causing positive feedback, and *lacZ* which produces an enzyme which breaks down lactose into glucose and galactose which can be metabolised more easily.

Glucose in the cell interacts with the membrane protein, reducing the rate at which it imports lactose and introducing a second control loop. As the concentration of glucose increases, less lactose is pumped into the cell and so the lac operon becomes less active, reducing transcription of the second unit. [Image Credit: Wikimedia Commons]

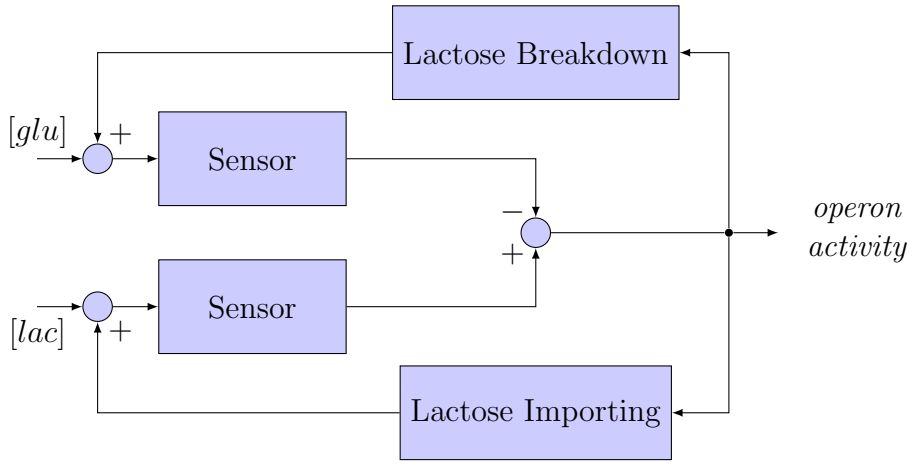

Figure 1.4: **Simplified block diagram of the lac operon**, showing only the most important interconnections. In the presence of lactose, transcription is turned on and more extracellular lactose is pumped into the cell, causing a positive feedback loop. Simultaneously, lactose is broken down into glucose (and galactose) which inhibits transcription, causing a negative feedback loop.

only lactose is available. This is achieved by the lac operon as described in figure 1.3 and shown schematically in figure 1.4.

Although the lac operon was one of the first such control structures to be discovered (and remains among the best understood), many other ingenious ways of tightly controlling protein production have been discovered, some of which act on transcription, some on translation and others on a combination of the two.

1.2 Synthetic Biology

Synthetic biology is a relatively new engineering discipline with the goal of applying proven engineering techniques such as standardisation, characterisation and encapsulation to biology. Synthetic Biology aims to use these design principles to combine existing phenomena to build new, artificial forms of life. The field is often confused with its spiritual predecessor, genetic engineering, which although similar in some respects does not explicitly embrace engineering principals such as standardisation with the attendant benefits of modularity of abstraction.

Synthetic biology can be thought of as programming, but with DNA instead of machine code. An example project which nicely captures this idea is Tabor's bacterial edge detector (Tabor *et al.*, 2009). Bacteria were programmed to produce a colourless chemical messenger in the absence of light and to produce a dark pigment in the presence of both light and the chemical messenger. When a film of these bacteria was exposed to a pattern of light and dark, the messenger produced in the dark regions diffuses into the light, where it stimulated the production of the pigment, resulting in an edge detection effect.

While this and other such simple demonstrations show some of the potential of synthetic biology, they lack wider applications and are of somewhat limited scale. A major problem in expanding this work is the lack of targeted reporter molecules, which carry signals through the biological circuit. In the edge detector example, two molecular signals are produced when light is not present – AHL, a cell-to-cell signalling molecule and cI, a transcriptional repressor molecule. Both AHL and cI are known to affect the promoter $P_{lux-\lambda}$; while AHL stimulates expression, cI strongly represses it. With expression of the dark pigment being driven by $P_{lux-\lambda}$, both light and AHL are required to cause the

pigment to be produced.

The effect of the molecules AHL and cI on $P_{lux-\lambda}$ is one of a small but growing number of well understood control motifs. But since reusing the same promoter/signal combination in the same cell is impossible due to cross-talk, there are simply not enough signalling modalities or characterised promoters available to perform more complex logic within the cell. Indeed, it is often the case that signalling molecules have multiple functions within the cell such that changing the concentration of one molecule to suit our goals may cause a seemingly unrelated area of the cell's metabolism to malfunction with undesirable consequences.

A separate and more practical project was the effort to produce artemisinin (the most effective known anti-malarial) in a cheaper and more scalable way. Malaria is a treatable disease which caused roughly 2,000 deaths *per day* in 2010, mainly because it mostly affects the developing world where access to anti-malarials is poor (WHO, 2011). Artemisinin is found naturally in sweet wormwood, but it is slow and expensive to extract directly from the plant and chemical synthesis is also an expensive and laborious process. Synthetic biologists were able to extract the metabolic pathway responsible for the biosynthesis of artemisinic acid (a natural precursor) and insert it into yeast (Ro *et al.*, 2006). Artemisinin produced in this manner was approved for sale by the World Health Organisation very recently (8 May, 2013). Shortly before the approval, long-term malaria campaigner Bill Gates said of the project -

> "By lowering the price and stabilising the supply, semi-synthetic artemisinin can be a breakthrough in the fight to control, eradicate, and eventually eliminate this disease".

The major limiting factor this project becoming viable was yield. In order to produce a useful amount of the drug, the metabolic pathway involved had to be up-regulated – i.e. more metabolic flux had to be directed though it. This led to a difficult balance – too little flux and very little artemisinic acid would be produced, too high and too much of the cell's energy would be used, causing the cells to grow slowly if at all. As well as this, growing yeast on an industrial scale is relatively expensive, and requires specialist equipment. Therefore it is desirable to search for host platforms which are better suited

to biosynthesis than yeast, in order to maximise the yield to cost ratio.

1.3 The Chloroplast

Chloroplasts are a major centre for biosynthesis in plants as they perform photosynthesis to provide energy for the plant. The result of an ancient symbiosis, up to 1000 of these primitive relic cells can be found within each plant cell, where they make an excellent target for synthetic biology. Their primitive nature makes them similar to bacteria, but with access to the more sophisticated plant cell machinery and good potential for biosynthesis. For example, the native enzyme RuBisCO is so abundant in the chloroplasts that it can represent up to 50% of overall soluble leaf protein.

Chloroplasts contain limited genetic information and expression machinery separate from that of the plant cell, which produces the molecules required for expression and several proteins vital to the photosystem. However, many proteins found in the chloroplast are in fact produced by the nucleus and then imported into the chloroplast.

The style of operon-based control common in prokaryotic cells has bot been identified in the chloroplast – instead most genes are transcribed constitutively (Sugita and Sugiura, 1996), leading to constant mRNA levels. It is also known that the mRNA transcripts in chloroplasts often do not contain a ribosome binding site (such as a Shine-Dalgarno sequence) at all or that such a sequence is not in the expected location (Sugiura *et al.*, 1998; Zerges, 2000).

This does not reflect the clear variation in protein levels found in the chloroplast, whose activity varies considerably during the chloroplast's circadian cycle – energy is stored during the day and consumed at night. Since mRNA levels are constant, this modulation must occur at a post-transcriptional level, and there is evidence to suggest that the nucleus is involved in controlling the cycle (Matsuo *et al.*, 2006).

Chloroplast mRNAs also undergo significant post-transcriptional processing such as C-U editing (where a genome-encoded C is converted to a U) and less commonly, U-C editing (Castandet and Araya, 2011). The underlying purpose of this RNA editing remains an open question. One theory is that it corrects for unfavourable mutations which have accumulated in the chloroplast genome and that removing these changes artificially would increase the efficiency of the plant (Fujii and Small, 2011). However,

it is also possible that editing is a vital method allowing to nucleus to tightly control expression in the chloroplast and that removing the mutations would result in plants which were unable to control their chloroplasts.

Understanding regulation in chloroplasts is a vital step before they can be effectively used for synthetic biology. For example, attempting to synthesise useful chemicals at the wrong point in the circadian cycle (*i.e.* during the day) would likely result in very poor growth and so we must be able to tap into the natural regulation system already present in the system.

1.4 Nucleotide Sequence Binding Proteins

Accurately predicting the structure of a protein from its amino acid sequence is very difficult and in most cases impossible. The main difficulty is that the bonds between amino acids in a protein have high rotational flexibility, giving amino acids a very high number of degrees of freedom, which often thwarts attempts to find the structure with minimum free energy. Even when this can be found, protein folding is often a complex process involving several interactions with other chaperone proteins which guide protein folding, meaning that the global minimum energy solution may not represent the actual shape of the protein.

As a result, modelling interactions between proteins and nucleotides such as DNA and RNA is impossible in the general case and these interactions are instead found and characterised using empirical techniques.

One of the first such interactions to be characterised were zinc fingers, a protein motif whose folding structure is stabilised by the incorporation of a zinc ion into the structure. Zinc fingers are common in many organisms and each one recognises and binds to a specific triplet of nucleotides. Zinc fingers have since been reverse-engineered such that a chain of them can be designed to bind to effectively arbitrary sequences (reviewed in Gaj *et al.*, 2013).

However, zinc fingers have limited modularity as each motif binds to three bases with varying specificity, meaning that there are often other sequences to which the protein will bind which can be hard to predict. Transcription activator-like effectors (TALEs) are modular repeat regions in which each repeat recognises a single DNA base via a simple

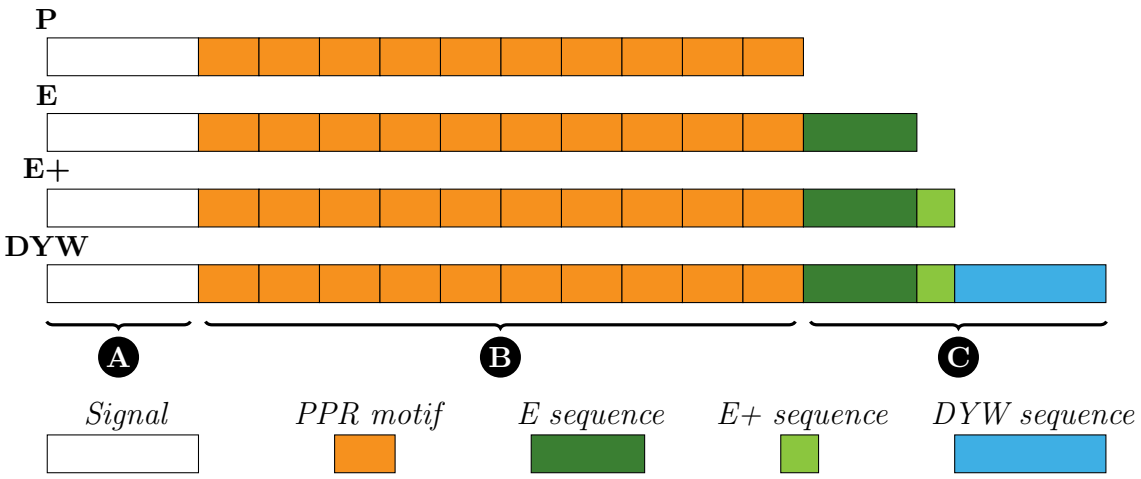

Figure 1.5: **Anatomy of the four classes of PPR protein.** (**A**) The signal peptide acts like an address header, essentially encoding the target location of the PPR within the cell (**B**) The binding region contains between 2-30 PPR motifs which specify the binding preferences of the protein (**C**) The tail region is optional contains either E, E and E+ or E, E+ and DYW conserved sequences.

code which is dependent on two hyper-variable amino acids within the repeat motif. The superior modularity of TALEs makes them inherently easier to design than zinc fingers. This has led to numerous novel applications, such as designing new transcription activators, although their main use is in selectively editing the genome to allow for novel studies of protein function (reviewed in Sun and Zhao, 2013).

These two classes of proteins have opened up new and exciting methods for basic research, gene therapy and synthetic biology.

1.5 The PPR Family

1.5.1 Discovery and Classification of the PPR Family

The Pentatricopeptide Repeat (PPR) family is a group of proteins commonly found in plants which show some similarities with TALEs, but bind to specific RNA sequences rather than DNA. They contain tandem degenerate repeating motifs, most commonly 35aa in length, which are referred to as PPR motifs, so called due to their similarity to the tetratricopeptide repeat (TPR) protein family which are known to be involved in protein-protein binding (Small and Peeters, 2000).

PPRs are found exclusively in the nuclear genome, with the mature proteins commonly being transported to organelles such as the chloroplast or mitochondria, where they are known to affect translation in numerous ways.

The typical PPR protein contains three regions, shown in figure 1.5. The first is a signal peptide which targets the protein to a particular organelle. The mechanism behind this is common to many proteins which are sent to particular locations within the cell (such as the chloroplast of mitochondria) and not a particularity of the PPR family.

The second region is an array of PPR motifs containing between 2 and 30 motifs. This is the region which binds with RNA, and thus the region which specifies the sequence which is bound. The motifs are degenerate – although they contain many similarities, they are not identical and in fact have quite considerable differences in some cases. The standard PPR motif is the P motif which is 35 amino acids long, but long (L) and short (S) variants are common in some proteins. The PPR motifs cause the protein to bind tightly to a specific mRNA sequence, and it is believed that pairs of contiguous motifs confer the particular protein's binding preference (Kobayashi *et al.*, 2012).

The third region is a tail sequence which is only present in some PPRs. The tail regions are known to contain a number of other conserved sequences whose exact function is unknown. Three main classes of tail sequence have been identified, the E subgroup which contains only 'E' motifs, the E+ subgroup which contains 'E' and 'E+' motifs and the DYW subgroup which contains 'E', 'E+' motifs and is terminated by a 'DYW' motif (Lurin *et al.*, 2004). Although the precise function of these tail motifs remains unknown, evidence suggests that the tail as a whole is involved in editing the RNA sequence near to the binding region (Yagi *et al.*, 2013b).

1.5.2 Known interactions with mRNA

PPRs can modulate gene expression and are involved in a variety of post-transcriptional RNA processing steps such as RNA editing, splicing and stability (Nakamura *et al.*, 2012; Schmitz-Linneweber and Small, 2008).

RNA editing

Several PPR proteins with tail motifs have been associated with RNA editing and in these cases the PPR binding site has been located a short distance from the edit site (Okuda

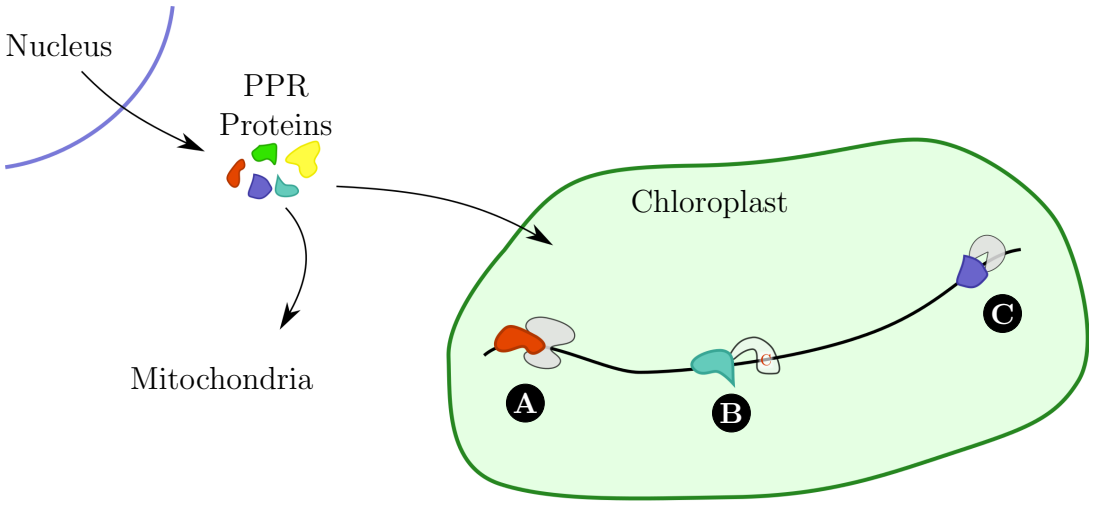

Figure 1.6: **Summary of PPR activity in the chloroplast.** (**A**) Increased translation – PPRs can promote ribosome recruitment and increase the translation rate of the transcript (**B**) RNA editing – PPRs edit the RNA transcript, as well as playing a vital role in the removal of introns (**C**) RNA stability – PPRs can increase the stability of the transcript by protecting against degradation by endonucleases, or decrease the stability, possibly by recruiting endonucleases. Similar processes are known to occur in the mitochondria.

et al., 2007; Yagi *et al.*, 2013b). A particular protein can be responsible for many edits within the chloroplast by having multiple binding sites within the chloroplast genome (Okuda and Shikanai, 2012), allowing a single protein to edit multiple genes.

C-U RNA editing has vital consequences for protein translation. Proteins begin with a start codon (AUG), which marks the position where translation should start. If instead the genome codes an ACG codon then the protein will not be expressed unless a C-U edit event occurs. Conversely, if a CAA codon is present in the gene, then an RNA editing event could convert this to UAA – a stop codon, causing translation to be terminated.

Increasing Translation

It has been shown that PPR binding to the 5' and 3' UTRs can stabilise mRNA transcripts and reduce degradation by ribonucleases (Pfalz *et al.*, 2009; Prikryl *et al.*, 2011). This increases protein yield as more mRNA will be present at any time, increasing the rate at which protein is created. In addition to stabilisation, PPR binding can facilitate ribosome recruitment and can thus be responsible for the initiation of translation.

Decreasing Translation

PPRs have been shown to be responsible for restoring fertility to plants affected by Cytoplasmic Male Sterility (CMS) (Bentolila *et al.*, 2002), which is of commercial importance in breeding. PPRs prevent sterility by preventing the production of specific proteins which cause the condition (Kazama *et al.*, 2008). While the specific interaction preventing translation is unknown, it is thought to be due to cleavage of the mRNA transcript or degradation (Wang *et al.*, 2006). It is also possible that the PPR out competes the ribosome when binding to the ribosome binding site.

Binding Rules

The mechanism behind the PPR-RNA binding interaction remains unknown, and it is not yet possible to accurately predict PPR binding domains from the amino acid sequence of the protein. One problem is that the exact structure of the tandem PPR motifs is not known, although other proteins which also contain PPR motifs appear to show a helical structure (Howard *et al.*, 2012; Ringel *et al.*, 2011), suggesting that PPR binding might act in a fashion similar to that of TALE repeats (Rubinson and Eichman, 2012).

As of writing, two major theories on the PPR binding rules exist, (Barkan *et al.*, 2012; Yagi *et al.*, 2013a). Both are based on statistical inference on the small number of characterised PPR-RNA interactions and are discussed in more length in section 2.3.

1.6 Hidden Markov Models

A Hidden Markov Model (HMM) is a model of a Markov process where the state is unobserved. Each state emits a symbol from an alphabet with probability dependent on the current state and it is the sequence of symbols which is observed rather than the states themselves.

HMMs are commonly used in bioinformatics in order to describe and predict repeating patterns in the sequence (Durbin *et al.*, 1998). The Pfam database, maintained by the Wellcome Trust Sanger Institute, is an open library of HMMs describing a large range of protein families which are freely available to all.

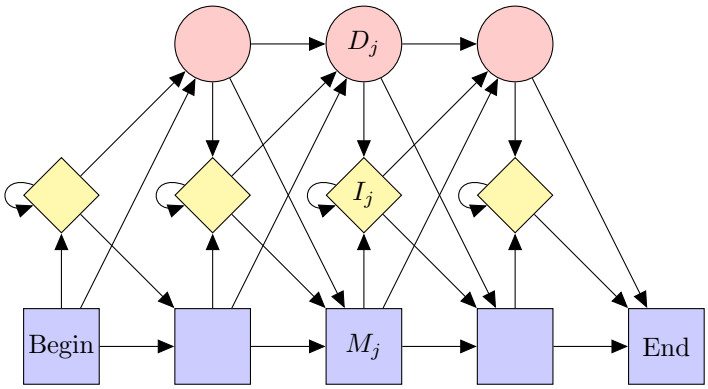

Figure 1.7: **A profile HMM.** Squares are *match* states, which emit a symbol according to the consensus sequence, diamonds show *insert* states, which insert one or more symbols into the sequence and circles are *delete* states which emit no symbols and effectively skip one or more symbols in the model.

The states M_j, I_j and D_j are collectively referred to as a node, and an HMM will contain as many nodes as there are symbols in the consensus sequence.

1.6.1 HMMER

HMMER is a collection of command line tools implementing all the basic algorithms required for using HMMs in a biological context (Eddy, 2010). HMMER is capable of constructing models from aligned sequences and of searching a target sequence for instances of the model.

HMMER does not use general HMMs (which can have any topology), but instead uses a profile HMM (pHMM). pHMMs have a fixed topology described in figure 1.7 and only the transition, emission and insertion probabilities have to be learnt. This restriction places minimal constraints on the type of sequence which can be modelled and allows learning to be done using an expectation maximisation algorithm. A more in depth discussion of pHMMs and their use in bioinformatics is given in Durbin *et al.* (1998).

A sequence can be searched for likely matches with a pHHM model using the *hmmsearch* program which makes heavy use of heuristics in order to search efficiently for possible match sequences. The parameters for the particular heuristics used can be tuned using command line arguments so that the similarity required between a matching sequence and the model can be set.

The *jackhmmer* program was also used during the project. This program iteratively

searches a protein sequence with a consensus protein sequence. First, the consensus sequence is aligned with the target, and an initial pHMM is created. The target is searched using the pHMM, and the resulting matches are used to create a better pHMM until convergence is reached and the final matches are reached.

1.7 Project Outline

Since their initial discovery and characterisation, TAL effectors have enabled a great many exciting projects in both synthetic biology and basic research. However, PPRs offer several potential advantages over TALEs, particularly in naturalistic control within the chloroplast and programmable translation-based control of expression. PPRs are naturally able to influence expression upwards *or downwards*, giving them more flexibility than TALEs.

There are still many problems which must be solved in order to effectively use PPRs, and the goal of this project is to explore these issues and investigate the potential of the PPR protein for synthetic biology. The particular goals of the project are to :

1. Implement an algorithm to accurately detect and annotate unlabelled PPR proteins in DNA

2. Investigate the presence of PPR proteins in genomes other than *A. thaliana*, a model plant

3. Attempt to predict binding locations of naturally occurring proteins, and find potential ways of improving the estimates

4. Discover whether known PPR binding domains are conserved across different chloroplast genomes

5. Model the activity of theoretical networks of PPR proteins in order to demonstrate their potential for creating biological circuits

The first item is discussed in chapter 2, items 2 to 4 in chapter 3 and the final item is discussed in chapter 4.

Chapter 2

Automating PPR discovery

Most sequenced genomes are also annotated with information about the genes they encode a specific locations. These annotations are usually helpful when navigating the genome as many are the result of empirical study. The majority of PPRs have not been studied in detail, however, and thus their annotation is unreliable and doesn't contain information about the PPR motifs found within the proteins.

For this reason it was necessary to develop routines for discovering and extracting PPR proteins from unannotated genomes, using a hidden Markov model of the characteristic repeat motifs to find likely targets. Having identified PPRs, we would like to be able to predict their binding locations within organelles.

The development of algorithms to achieve these things is discussed in this chapter.

2.1 pyHMMER

The HMMER suite provides all of the basic algorithms required in order to perform an HMM search on a target amino acid sequence, but it does have some limitations.

The first is that it is a command line program and does not have bindings to any programming language. HMMER reads inputs from files, and writes out tabular output data to file which would be very time consuming to parse by hand.

HMMER cannot translate sequences on the fly – is can only compare a protein model with a protein target, and so the genome must be translated before HMMER is invoked. There are a total of six possible reading frames (3 forwards and 3 backwards, due to the 3:1 nature of translation) and the genome must be searched in each of these six frames.

HMMER is not commercial software and is developed by a group of scientists at the Howard Hughes Medical Institute (HHMI) under an open licence for research purposes. As such, it contains a number of minor bugs which are not fatal to the software's functionality, but can sometimes cause problems under particular circumstances. The

most problematic of these causes enormous memory usage (over 20GB[1]) and prevents the program from completing.

In order to overcome these issues, a python wrapper for HMMER called pyHMMER was designed and written as part of this project. Python was chosen as the main language mainly because of its excellent library support – for example the biopython library solved many of the difficulties when working with biological sequences without extra effort.

pyHMMER does not implement all the features available in HMMER, but rather it implements those which were most vital to this project. Its main features are -

- Read and write *.hmm* files, HMMER's custom file format for storing HMMs
- Execute searches using *hmmsearch* and *jackhmmer*, accepting all valid command line arguments and returning their output as biopython objects, handling the creation and removal of all the necessary temporary files automatically
- Seamlessly perform six-frame translations on the fly (implemented in C for best performance) and correctly map the location of each match to the original target alphabet
- Automatically terminate HMMER processes which attempt to allocate more memory than the system can sensibly be expected to provide[2] and then call HMMER sequentially with subsections of the target, automatically mapping all matches back to the original target
- Fully unit-tested with python's *unittest* framework

pyHMMER has been developed under an open-source licence and is freely available from *https://github.com/haydnKing/pyHMMER*, although all code used in this project was written by myself.

2.2 Automated PPR Detection and Extraction

Several algorithms were developed and compared in order to extract PPRs from unannotated genomes. The results of each algorithm were compared with experimentally

[1] One particular instance of this error is due to an unsigned integer wrap-around which causes significantly more memory to be requested from the operating system than could possibly be needed

[2] Linux only

validated PPRs and the best chosen.

Before development could begin, a HMM of the PPR repeat motif was required. There are four such models available in Pfam[3], and each one was tested on known PPRs in order to discover which model worked best when searching for motifs. It was found the PPR_3 model is most sensitive to the motifs, yet still returns few false positives.

Armed with this model, the final algorithm for discovering PPRs proceeds as follows for each chromosome within the genome

1. Perform a HMM search on the whole sequence. This will discover the most obvious motifs only

2. Cluster the motifs into groups such that members of the same group are on the same strand and are within a certain distance of each other

3. For each group, extract and 'envelope' region containing each motif in the group along with large margins either side.

4. Search each envelope region for PPR motifs. This search is more focussed than the previous one and will reveal more motifs than previously. Discard any envelopes which contain only one motif.

5. Starting from the first position of the first motif, search backwards one codon at a time until a start codon ('ATG') is reached

6. Starting from the last position in the final motif, search forwards one codon at a time until a stop codon is reached ('TGA', 'TAG' or 'TAA'). Extract the putative PPR from between the start and stop codon

7. Check for PPRs which overlap. Each set of overlapping proteins should be removed and a new, larger envelope extracted. The algorithm then continues again from step 4 with the new envelopes

8. Check for PPRs where the motifs have filled the envelopes – i.e. ones which are missing a start or a stop codon due to not having searched far enough. Extract larger envelopes for these proteins and continue from step 4

[3]http://pfam.sanger.ac.uk/search/keyword?query=PPR

9. Search each protein for gaps between motifs which are the correct size to fit a PPR motif. Search these regions specifically, increasing HMMER's sensitivity to look for reluctant PPR motifs. Also search the beginning and the end of the protein in this way

10. Search each protein for small (2/3 codon) gaps between the motifs and move the end position of the previous motif in order to fill these gaps. This allows the motifs to be classified as P, L or S

11. Classify the proteins depending on which types of motifs they contain. Extract the protein sequence of each tail sequence and classify it using *jackhmmer* to search for the known consensus sequences for E, E+ and DYW motifs

12. Predict each protein's sub-cellular location using the *targetP* program

A brief discussion of the rationale and implementation of the most important steps follows.

The reason why the motifs found in step 1 cannot simply be accepted is due to the degenerate nature of the motifs. It would be possible to decrease HMMER's reporting threshold as to return all possible motifs but since the search space is large and the model accepts a wide range of sequences, there would be a large number of false positives. By using default values for these thresholds there is unlikely to be a problem as HMMER is designed to show the most probable matches and only a few of the possible false positives. The presence of a few false positives at this stage is not an issue because the chance of finding several false positives immediately adjacent to each other (as would be required to pass the later stages of the algorithm) is highly unlikely.

Having found the most obvious motifs, step 2 groups motifs which are believed to belong to the same protein. Initially this was restricted to motifs which were in the same reading frame (i.e. the gaps between start of each motif were multiples of three), but this was later expanded to all motifs on the same strand, as introns (see sections 1.1.1 and 1.1.2) are known to be present in some PPR motifs. Experiment showed that grouping motifs which were within $1500bp$ of each other gave good results. Grouping was implemented by first sorting the motifs into ascending order and then searching through linearly giving a cost of $O(n \log n)$ rather than the cost of $O(n^2)$ required for exhaustively

comparing each motif.

Envelopes are then extracted from these groups in step 3. The term 'envelope' is borrowed from HMMER's output and refers to the fact that we expect there to be a PPR somewhere within this region, but we are not sure where exactly. For envelopes on the reverse strand, the sequence is extracted such that it reads in the 5' to 3' direction on that strand. It is important to maintain a record of where in the target sequence the envelope came from, as this information may be required later in the algorithm. A margin of $1000bp$ either side of the group was found to give good results.

The search space in step 4 is several orders of magnitude smaller than the first pass and so the chances of finding multiple high-scoring matches by chance are negligible. Shortening the target in this way effectively moves HMMER's baseline for scoring matches such that lower scoring matches which would previously have been written off as noise are now treated as legitimate matches.

Steps 7 and 8 effectively correct for situations where the parameter values chosen for grouping and envelope extraction do not perform well. For example, if step 1 detects the first and last motifs from a particularly long PPR then these will be treated as belonging to separate proteins up until this point. Similarly, if only one motif was detected then the size of the actual protein may be larger than the envelope which is extracted.

These two steps introduce loops into the algorithm and thus introduce the worrying possibility of an infinite loop preventing the algorithm from completing. In the case of step 7 this cannot happen as for each iteration of the loop the number of putative proteins is half that of the previous loop, meaning that no infinite loop is possible. An infinite loop is also impossible in 8, as the growth of the envelope is limited by the size of the search query. However, since each loop iteration is expensive and adds only a constant length to the envelope this could take quite some time in the worst case. To protect from this, a large upper bound was placed on the maximum length of a protein.

Proteins which are input to step 9 often contain gaps of around 35 amino acids – the correct size for a repeat motif – and comparison with known proteins shows that a motif should indeed be placed in this region. These motifs can be found by searching these regions with a lower reporting threshold than the default. A plausible explanation of these poorly conserved motifs is that the presence of relatively well conserved (and thus

well folded) regions on either side of the degenerate motif increases its tendency to fold correctly. However it could also be the case that these regions simply represent a gap in the recognition chain (where any base would be accepted) or an intron; more empirical results are needed in order to determine this in every case. Setting each of the parameters *F1*, *F2* and *F3* to 0.5 gave a reasonable trade-off between finding likely reclusive motifs and rejecting random sequences.

Studies such as Lurin *et al.* (2004) have shown that tandem PPR repeat motifs tend not to have small gaps between them. Since pyHMMER returns the location of the HMMER model, each match is the same length as the model. This is corrected for in step 10, such that the motifs can be classified as type P (length = 35aa), L (length > 35aa) or S (length < 35aa).

The final two stages classify the extracted proteins depending on their type and subcellular targeting. Step 11 makes use of the *jackhmmer* program which iteratively constructs HMM models of a consensus sequence based on a target sequence and is supported by pyHMMER. The final step uses *targetP*, a well respected prediction algorithm for subcellular localisation(Emanuelsson *et al.*, 2000).

2.3 Predicting PPR Binding regions

Given an identified and well annotated PPR protein, predicting the RNA footprint to which is binds is not straightforward. In the case of TALEs (section 1.4), a well known mapping exists between the amino acids at specific locations within the repeat and the preferred DNA base of that repeat. Unfortunately, such a mapping has yet to be confirmed for the PPR family, although two main suggestions have been made, by Barkan *et al.* (2012) and by Yagi *et al.* (2013a).

Since neither the structure of of the PPR motif or of a PPR-RNA complex has been solved, the only method to elucidate the rules governing binding preferences is by looking for statistical dependencies between the amino-acid sequence and RNA footprint of known PPR-RNA pairs. This is the strategy used by both papers mentioned above and so they lead to similar results. The papers each provide methodologies to convert each PPR motif into a distribution over each the symbols in RNA, $\{A, C, G, U\}$.

The next two sections outline methods for discovering likely binding sites in a partic-

ular target given a sequence of such distributions.

2.3.1 Profile Hidden Markov Models

The first method which was tested was using pHMMs, as this would allow the use of HMMER's advanced searching algorithms. This seems a simple task – the probabilities given at each motif give the emission probabilities at each node, insert probabilities can be uniform and the transition probabilities can be determined empirically.

The first issue which arises is that pHMM models which are used for searching with HMMER must be normalised in order to make score calculations. This involves estimating the parameters of the distribution of model scores for random sequences. A program called *hmmsim* exists in the HMMER package for just this purpose, however the program is not a mainstream part of the HMMER package and is present more as a tool for testing HMMER's internals rather than use by the end user. As a result, it is not as stable as other HMMER tools and has not been tested with all possible use cases. Unfortunately, one particular unanticipated use case is the normalisation of a DNA model – *hmmsim* is hard-coded to accept only protein models as this is HMMER's primary use case.

This problem was circumvented by writing the model as a protein model by expanding the probabilities of each of the four bases to those of the 20 amino acids (i.e. the first 5 amino acids corresponding to an 'A' etc. . .). However models which had been normalised in this way did not prove to be effective when searching large sequences even when a known high scoring sequence was inserted into an otherwise random target.

For this reason, pHMMs were abandoned as a method of predicting binding footprints.

2.3.2 Position-Specific Scoring Matrices (PSSMs)

PSSMs are another common technique for discovering particular sequences in a target. They are similar in nature to pHMMs (which can be considered as a generalisation of the PSSM), but are generally significantly simpler. Each column of the matrix corresponds to a particular position in the sequence and the rows specify the probability of each possible symbol appearing in that location.

The probabilities are generally stored as logarithms, such that the probability of any particular sequence of length N is simply the summation of N values from the sequence.

PSSMs can incorporate the background distribution by storing log-odds scores such that

$$m_{i,j} = \log \frac{p_{i,j}}{b_i}$$

where $p_{i,j}$ is the probability of observing symbol i at location j and b_i is the probability of observing symbol i in the background sequence.

PSSMs are easy to construct given the distribution of symbols at each location, but require more work when searching for highly scoring sequences, particularly given that PPR binding footprints often contain bases which are not actually bound to a motif – an insert in pHMM terminology.

A simple algorithm for finding maxima proceeds as:

1. Score the sequence at possible model position in the sequence
2. Discard the positions which aren't a local maximum
3. For each maximum, try inserting gaps at each location in the model to attempt to increase the score of the maximum, then return the highest scoring matches

This algorithm is somewhat inefficient, but it returns the highest scoring alignments in a reasonable time. Step 1 may seem the most inefficient as every possible alignment is tested, but this actually interacts rather well with the memory cache on modern computers as it searches through the data linearly. Step 3 is in fact the rate limiting step in most cases, as the number of possible combinations of gap locations grows rapidly with both the length of the model and the number of gaps.

Chapter 3

Results

The algorithm developed in section 2.2 was applied to the genomes of several target organisms. PPRs were found in similar numbers in almost all of the plants tested, implying a high level of similarity in the method and degree of PPR interaction between the nuclear genome and the organelles in most plants.

Two recently published methods of predicting binding domains were compared and found to be inaccurate when searching in sequences as long as the chloroplast genome. Instead, a large set of chloroplast genomes were searched for putative homologs of the binding domain of a set of 12 characterised PPRs from Arabidopsis. It was found that these binding domains in the chloroplast appear to be highly conserved, suggesting that the relevant PPRs in the nuclear genome may also be well conserved.

3.1 Survey of PPRs in Plants

3.1.1 Selected Genomes

Despite recent advances, whole genome sequencing is still expensive, time consuming and error prone and as a result only a small subset of plant genomes have been fully sequenced. Those genomes selected as part of the survey are shown in table 3.1 and consist mostly of plants, since this is where most PPRs are known to reside.

One interesting exception is *P. falciparum*, which is the parasite which causes the most dangerous form of malaria in humans. It is included here because *P. falciparum* contain an apicoplast – an organelle similar to the chloroplasts found in plants and thought to be the result of a secondary endo-symbiosis.

Complete sequences for the selected genomes were obtained from the National Centre for Biotechnology Information (NCBI) genbank genomes repository.

Genome	Abbrv.	Description
Arabidopsis thaliana	At	Thale cress, winter annual
Brachypodium distachyon	Bd	Purple False Brome, grass species
Citrus sinensis	Cs	Orange
Eutrema parvulum	Ep	Small herb
Eutrema salsugineum	Es	Halophyte (tolerates high salt)
Glycine max	Gm	Soya Bean, legume
Gossypium raimondii	Gr	Cotton
Malus x domestica	Mx	Apple
Medicago truncatula	Mt	Barrel Clover, legume
Oryza brachyantha	Ob	Grass species, distant rice relative
Oryza sativa	Os	Rice
Ostreococcus tauri	Ot	Unicellular green algae
Plasmodium falciparum	Pf	Malarial parasite, contains apicoplasts
Solanum lycopersicum	Sl	Tomato
Sorghum bicolor	Sb	Grass species
Zea mays	Zm	Maize

Table 3.1: **Target genomes searched for PPR proteins**

3.1.2 Extraction Results

The genomes listed in table 3.1 were searched using the algorithm described in section 2.2. Figure 3.1 shows the number and type of PPRs found in each genome of interest, ordered by number found.

It is clear that *P. falciparum* contains no PPRs at all and so although the apicoplast is superficially similar to the chloroplast, an entirely different control mechanism is at work in them. *O. tauri* contains very few PPRs, which is unsurprising as it is a relatively simple unicellular organism.

The majority of the plants surveyed contain between 400 and 600 PPR proteins, however, the clear exception to this is *G. max*, the soya bean, which contains considerably more putative PPR proteins than any other of the surveyed plants – 940 in total. The reasons behind this are uncertain, although the genome is known to contain several repeats of some proteins (Schmutz *et al.*, 2010).

Figure 3.2 shows a histogram of the number of motifs found in each PPR stacked by family. PPRs most commonly have between 10 and 15 motifs, the most common being 13, although some PPRs are around 30 repeat regions in length. Figure 3.2 also shows the probability of a binding domain appearing in a random sequence of 150,000bp – the

approximate length of a chloroplast genome. This distribution is derived in appendix A.

It is interesting that the majority of PPRs appear to the right of the sharp drop in probability, which shows that most PPRs are potentially specific enough to bind to one region only within the chloroplast.

3.2 Binding Locations

A key problem which must be overcome in order to develop PPR-based technology is understanding how the amino acid sequence of the proteins determines the RNA sequence to which it binds. This knowledge could be used in two ways, firstly to predict the binding footprints of PPRs which are known to exist in order to discover their role in the chloroplast and secondly in order to make designing PPRs with pre-specified binding preferences a reality. This section focusses on the former problem.

We wish to discover a method to map from the amino acid sequence of a repeat motif in a PPR to the RNA base to which it binds. Two such mappings have been proposed recently, the first in Barkan *et al.* (2012) and the second in Yagi *et al.* (2013a). These mappings are similar in many senses, they both predict a distribution over the four bases based on the amino acids found at particular locations within the motif. Barkan coding makes use of the amino acids found at positions 6 and 1' (where 1' is the first amino acid in the next motif), while Yagi coding uses the amino acids found at 3, 6 and 1' (referred to as positions 1, 4 and ii in Yagi *et al.* (2013a)).

Both the Barkan and Yagi coding schemes were implemented in Python and tested using the PSSM prediction algorithm described in section 2.3.2. Characterised PPRs (given in Yagi *et al.* (2013a)) were searched against the *A. thaliana* chromosome and the scores were compared to the average output for an entirely random sequence of equal length. It was found that neither scheme produced any matches with scores high enough to be statistically significant. In fact, in most cases the known PPR binding site was not even in the top 10% of matches. While there is statistical evidence to support that the identified amino acids are involved in binding, there is clearly there is some way to go in developing an understanding of this phenomenon.

There are two main problems which came to light while studying these papers.

1. PPR motifs are often not very well defined – the two papers even use differing

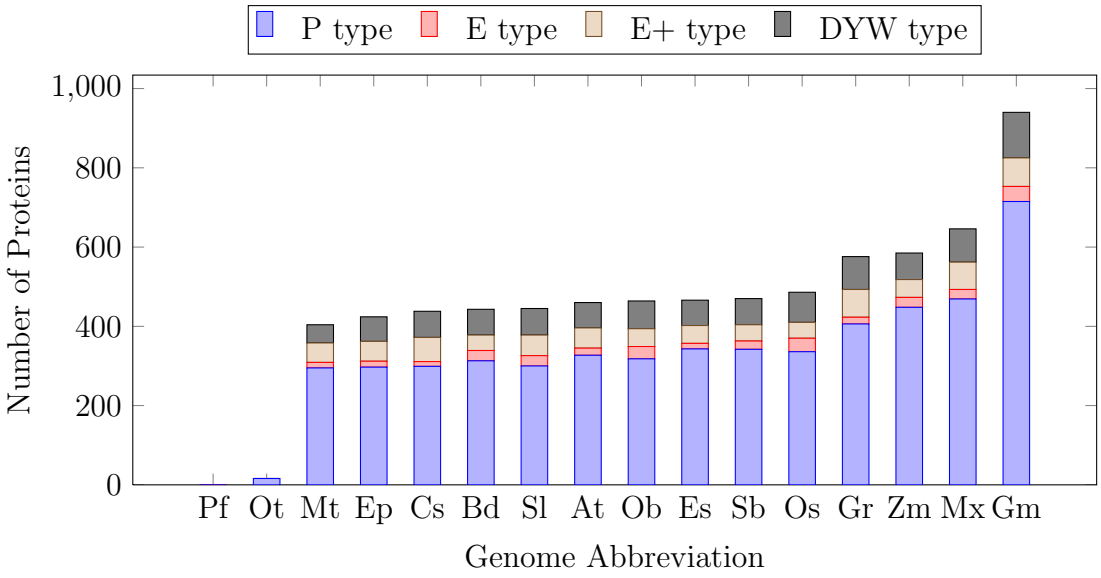

Figure 3.1: **The number of PPR proteins found in each genome**, stacked by type (as defined in figure 1.5). The two non-plants, *O. tauri* and *P. falciparum* contain none or very few, whereas most of the plants surveyed contain a similar number of PPRs, with the exception of *G. max* which contains an unusually large number.

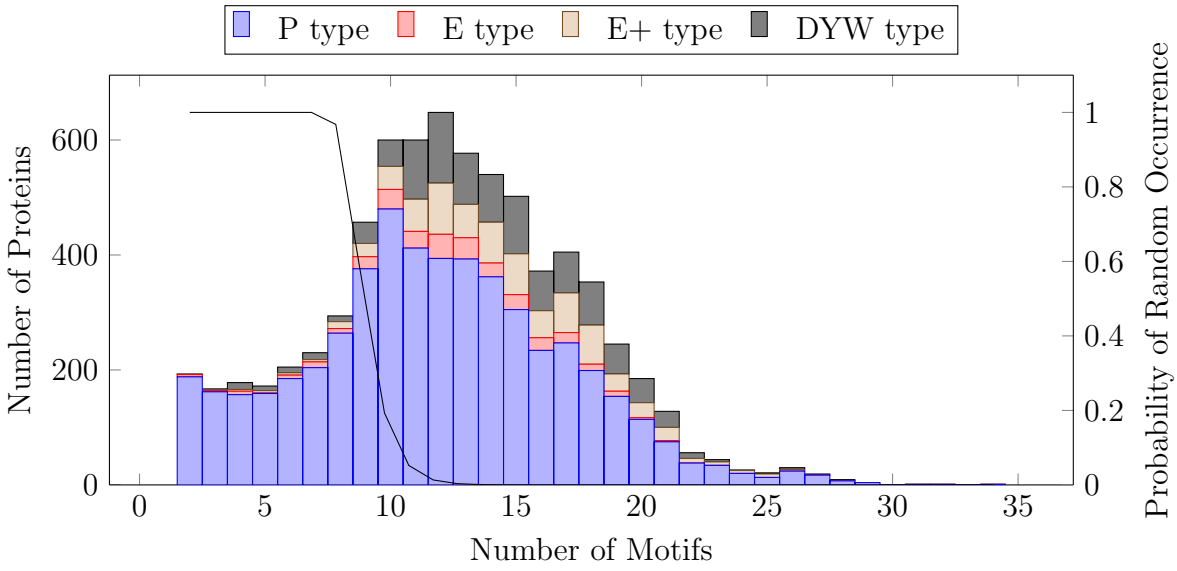

Figure 3.2: **The distribution of the number of motifs found in proteins**, summed over each of the organisms surveyed and stacked by type, shown plotted alongside the probability of a sequence of that length appearing by chance in a random DNA sequence of length 150000bp. PPRs are most commonly between 10 and 15 motifs long, which gives enough specificity to make finding a particular sequence by chance unlikely assuming only a single possible binding sequence per protein.

conventions on where exactly the motif starts, and although this discrepancy is noted in the latter paper it is not justified

2. There are few well known PPR-RNA interactions

The first of these problems will improve with the second – we can only improve our prediction of exact motifs when we understand more about them.

Discovering More PPR-RNA Pairs

While the number of fully sequenced nuclear genomes remains low, a large number of chloroplast genomes have been sequenced. This is due to their relatively small size (120-170kb) which makes them considerably easier and cheaper to sequence than the nuclear genome.

Chloroplast genomes are generally well conserved between different organisms – although genes are sometimes rearranged or duplicated, most chloroplast genomes contain roughly the same gene content. This can be used to our advantage by looking for potentially homologous binding sites in other chloroplasts to predict the presence of a similar protein in the nuclear genome. Homologs could then be confirmed experimentally with relative ease, as we would know enough about the sequence of the homologs to be able to physically extract them from the nuclear genome (*e.g.* by PCR) and we would also already have a good idea of what the binding footprint is. Having a large group of proteins whose binding footprint changes only slightly would be a great aid in elucidating the binding scheme.

This theory was tested using the 340 chloroplast (and closely related organelle) genomes available from the NCBI's Organelle Genome Resource. Directly searching other genomes for sequences similar to known binding domains would not produce results of much significance as the size of the genome means that similar sequences are likely to exist there by chance. Instead, the protein sequence of the gene closest to the known binding site was used to search the other genomes for potential homologs using HMMER's *phmmer* program. Since DNA to amino acid coding is a overcomplete, even if a protein is found with an identical amino acid sequence the DNA sequence could be very different.

The overwhelming majority of the other 339 genomes were found to have proteins with very strong sequence similarity to those in Arabidopsis, as expected.

The DNA sequence of these homologous proteins was then searched for regions similar to the original binding sites using a PSSM approach. Many of them contained sequences with very high homology with the original binding domain in Arabidopsis ($> 90\%$ in some cases). In addition, where a genome had multiple potential homologs, the same sequence differences between the binding sites was observed, suggesting that there might indeed be a similar PPR protein acting on the region.

In one example, the Arabidopsis PPR protein PDE247 is known to have two binding domains, '*acacgtgcaa*' and '*agaagcccaa*'. These exact sequences are found in the chloroplast genes *psbK*, *trnH*, and *ycf2*. A total of 10 potential homologs for these genes were found in the chloroplast genome of the tree fern *Alsophila spinulosa*, one for *psbK* and 9 for *ycf2*. Eight of these potential homologs contained the exact sequence '*agaagcctaa*', which differs from the known binding domain in Arabidopsis by only one base. Sequence similarities such as this were found to be common in many of the available chloroplasts.

Clearly experimental work is required to verify these potential homologs and to attempt to characterise the PPRs which may be present, but it seems hopeful that there are a large number of PPRs similar to those already characterised present in the genomes of other plants.

Chapter 4

Simulating PPR Activity

SYNTHETIC biologists often draw on analogies with electrical engineering when designing new systems. Individual components can be assembled to form biological circuits, where the transcription rates can be seen as analogous to the current flowing through the device.

Once we solve the issues discussed in the previous chapter, PPRs could be designed to bind to arbitrary RNA sequences. What kind of components could we then build for our biological circuits? In this chapter, a generalised ODE model of a network of interacting PPRs is developed and used to show that PPRs could potentially be used to perform basic logic operations at the translational level.

4.1 Simulation Model

In order to capture PPR activity, we need to model the interaction between the cell's expression machinery and RNA-PPR binding. This is most simply done using an ODE model, where each process is modelled using a rate equation.

The first two processes to model are transcription and translation. Assuming that the substrates necessary for RNA and protein production are readily available and inexhaustible, these processes simple to model as neither DNA nor RNA are consumed in either reaction. This is not always the case in a cell, but for low enough rates of transcription and translation this assumption holds. Thus, the rate of production is simply proportional to the amount of DNA or RNA present respectively.

We also need to model the production of the $\mathrm{RNA_i}{-}\mathrm{P_j}$ complex, as shown in 4.1, where $\mathrm{P_j}$ represents the j^{th} protein in the model. Note that this reaction is reversible, and so two rates are required – one representing the forwards reaction and one representing the backwards direction.

$$\mathrm{RNA_i} + \mathrm{P_j} \rightleftharpoons \mathrm{RNA_i}{-}\mathrm{P_j} \tag{4.1}$$

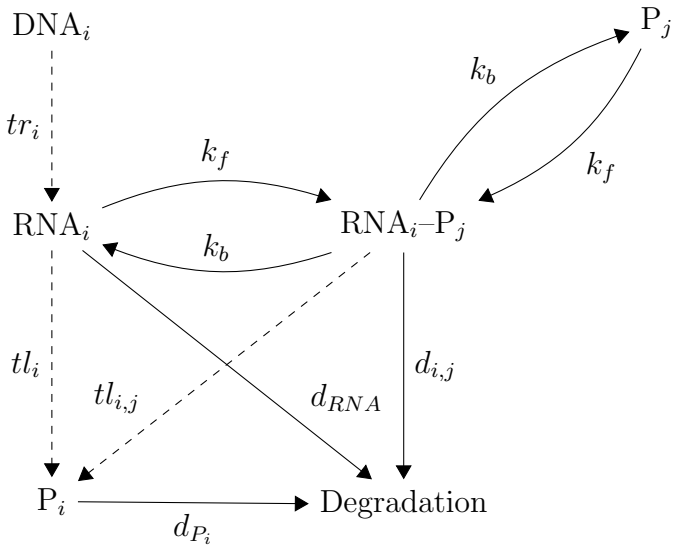

Figure 4.1: **Diagram of the general reaction network** for the i^{th} protein in the network. Dotted lines show production, where the molecule at the start of the arrow are not consumed; full lines show where molecules are converted to other types.

In addition to these processes, we need to model translation of the $\text{RNA}_\text{i}{-}\text{P}_\text{j}$ complex and also degradation of RNA, P and RNA−P. Each of these reactions is simple to model in isolation as the rate is simply proportional to the concentration of the molecule, as we again assume that all the necessary substrates are available. Figure 4.1 summarises these reactions for the i^{th} protein.

The differential equations governing the concentration of each molecule are shown in 4.2.

$$\begin{aligned}
\frac{d[\text{RNA}_\text{i}]}{dt} &= tr_i - [\text{RNA}_\text{i}] \cdot d_{RNA_i} - \sum_j \Big([\text{RNA}_\text{i}][\text{P}_\text{j}] \cdot k_{f_{i,j}} - [\text{RNA}_\text{i}{-}\text{P}_\text{j}] \cdot k_{b_{i,j}} \Big)\, i \\
\frac{d[\text{P}_\text{i}]}{dt} &= [\text{RNA}_\text{i}] \cdot tl_i + \sum_j [\text{RNA}_\text{j}{-}\text{P}_\text{i}] \cdot tl_{C_{i,j}} - [\text{P}_\text{i}] \cdot d_{P_i} \\
&\quad - \sum_j \Big([\text{RNA}_\text{i}][\text{P}_\text{j}] \cdot k_{f_{i,j}} - [\text{RNA}_\text{i}{-}\text{P}_\text{j}] \cdot k_{b_{i,j}} \Big) \\
\frac{d[\text{RNA}_\text{i}{-}\text{P}_\text{j}]}{dt} &= \Big([\text{RNA}_\text{i}][\text{P}_\text{j}] \cdot k_{f_{i,j}} - [\text{RNA}_\text{i}{-}\text{P}_\text{j}] \cdot k_{b_{i,j}} \Big) - [\text{RNA}_\text{i}{-}\text{P}_\text{j}] \cdot d_{C_{i,j}}
\end{aligned} \tag{4.2}$$

Where:

$$
\begin{aligned}
tr_i &= \text{is the translation rate of the } i^{\text{th}} \text{ protein} \\
d_{RNA_i} &= \text{degradation rate of RNA}_\text{i} \\
d_{C_{i,j}} &= \text{degradation rate of RNA}_\text{i}\text{−P}_\text{j} \text{ complex} \\
k_{f_{i,j}} &= \text{rate of association of RNA}_\text{i} \text{ and P}_\text{j} \\
k_{b_{i,j}} &= \text{rate of dissassociation of RNA}_\text{i} \text{ and P}_\text{j}
\end{aligned}
$$

This system can be easily converted into a state-space formulation by defining the state space to be the concatenation of the concentrations of each molecule in the system.

These equations contain a large number of parameters for which numerical values are needed in order to simulate a network of PPRs. We can improve the situation by making some simplifying assumptions, namely

1. Transcription rates are either high or low, depending on input conditions
2. Free RNA transcripts are either highly translated or very slowly translated
3. RNA−P complexes either have high translation rates and long half-lives (excitatory PPR binding) or low translation rates and short half-lives (repressive PPR binding)

These three assumptions reduce the number of parameters required and suggest a compact graphical representation of a network. Each protein is represented by a circle which contains a symbol indicating whether the protein is naturally translated or not. Proteins which are produced in response to inputs to the system contain instead a letter indicating the input to which they respond. PPR interactions are represented by lines between the proteins – an arrow between A and B represents that protein A binds to RNA B causing excitatory binding, while a perpendicular bar represents repressive binding.

4.2 Simulation Results for Logic Gates

Figures 4.2, 4.3 and 4.4 show simulations of NOT, OR and NAND logic gates respectively using figures derived from the literature (see Andersen *et al.*, 1998; So *et al.*, 2011), although the plots shown varied little over a large range of parameter values.

Description	Value
High transcription rate	30 transcripts minute^{-1}
Low transcription rate	0.03 transcripts minute^{-1}
On translation rate	0.693 proteins(transcript $\times$ minute)$^{-1}$
RNA half-life	2 minutes
Protein half-life	40 minutes
Long Complex half-life	10 minutes
Short Complex half-life	1 minute
Binding Rate	0.5 minute^{-1}
Unbinding Rate	4×10^{-9} minute^{-1}

Table 4.1: Rates used for the simulations shown in sections 4.2 and 4.3. Each was taken from Andersen *et al.* (1998); So *et al.* (2011), although some of the values are known to vary greatly in the cell. Although only the ration of binding to unbinding rate has been measured (via the disassociation constant. The plots shown were stable for a large range of binding rate.

It is impossible to conclude with certainty before these simulations are validated experimentally, but they do suggest that interesting and useful logic operations are indeed achievable using known PPR-mRNA interactions.

A major limitation of this model is that only one PPR may affect a transcript at a time and there is no attempt to model competition between different PPRs over a binding site. Both of these things can happen in reality since there is nothing to prevent a second PPR from binding at a different location on the same transcript. We could attempt to model these interactions (possibly using stochastic simulations rather than ODEs), but since so very little is known of the molecular processes driving PPR-RNA interactions such a model would be largely speculative.

4.3 Implementing Tabor's Edge Detector

We can use this model to simulate an implementation of Tabor's bacterial edge detector (introduced in section 1.2). The system described in Tabor *et al.* (2009) produces pigment in response to the presence of AHL and light. Our inputs to the PPR network are the AHL signal (S) and dark ($\overline{L}$), and output a pigment signal (P) according to

$$P = S \cdot L$$

$$P = \overline{\overline{S} + \overline{L}}$$

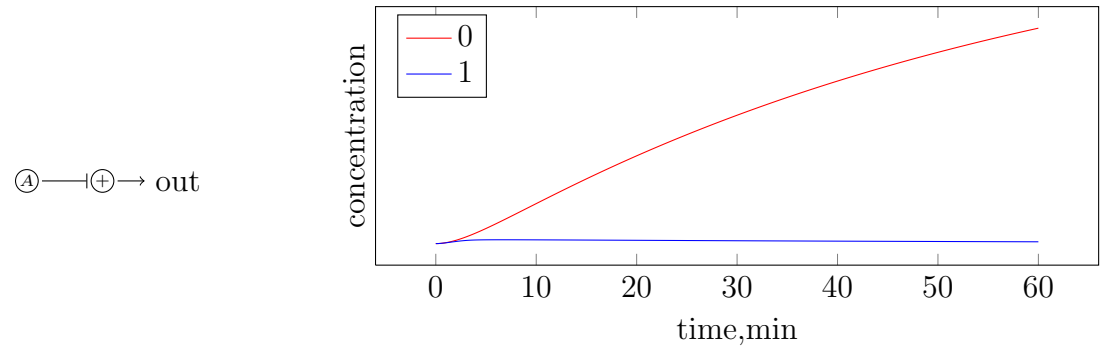

Figure 4.2: **An implementation of a NOT gate.** Expression of A causes the output to be repressed.

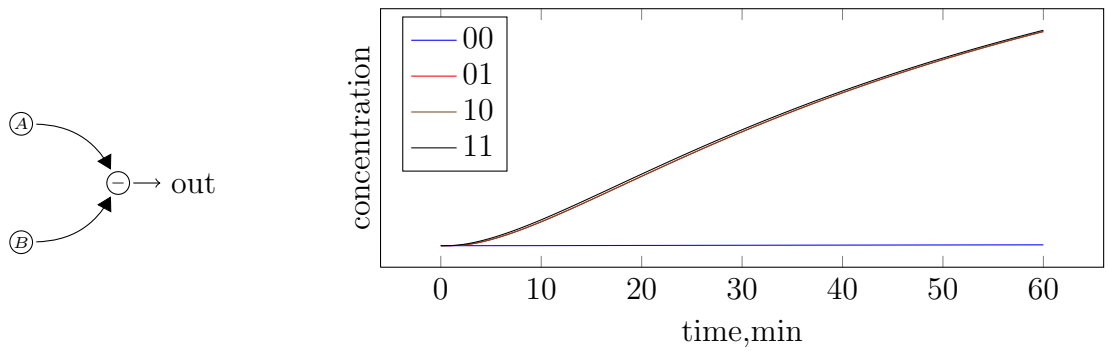

Figure 4.3: **An implementation of an OR gate.** Expression of either A or B causes excitation of the output. This can be easily expanded to N inputs using a total of N (possibly identical) PPRs.

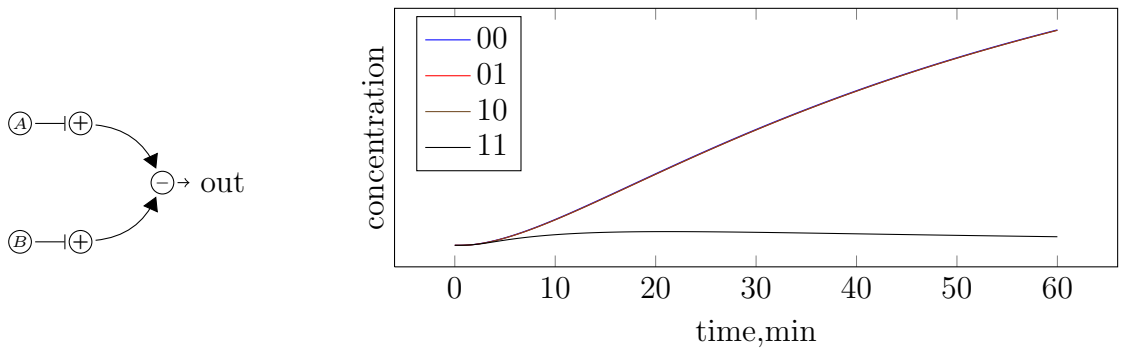

Figure 4.4: **An implementation of a NAND gate.** Output is produced unless both A and B are high. This can be implemented for N inputs with $2N$ PPRs.

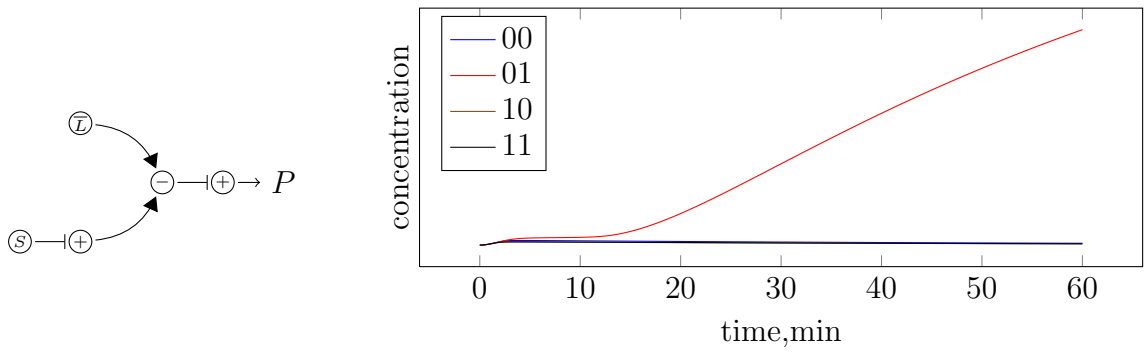

Figure 4.5: **An implementation of Tabor's bacterial edge detector.** Pigment is only produced in response to light ($\overline{L} = 0$) and signal ($S = 1$).

using De Morgan's law.

Figure 4.5 shows an implementation of this logic using simulated PPR proteins. The key advantage of this solution is that knowledge of a particular promoter which responds to the chosen cell-cell signal molecule used the output of the light detector is not required, we merely need to be able to detect the two signals independently. This would allow us to freely choose another cell-cell signalling molecule (perhaps with a different diffusion rate) in order to change the behaviour of the system – all we require is a promoter which will respond to it.

Chapter 5

Conclusions and Further Work

A brief summary of the work carried out during the project and a description of the work required to make PPR technology a reality are given in this chapter.

5.1 Summary of the Work

Pentatricopeptide repeat (PPR) proteins are a vital class of proteins for a large number of plants. During this project, software was developed to automatically identify, extract and characterise PPRs from unannotated DNA sequence. It was found that the majority of plants for which nuclear genomes are available contain roughly the same number of PPR proteins, suggesting that the function of these proteins is consistent across these plants [1].

A comparison of current theories regarding the prediction of PPR binding preferences was made, and it was shown that while current models of PPR binding sites score a genuine match highly, they allow too much variation in the sequence such that any reasonably long random sequence will contain matches which are just as good if not better, making accurate prediction impossible.

It was also shown that there appears to be considerable conservation of PPR binding sites across different organisms. If this is confirmed experimentally then there must also be considerable conservation of both PPR function and PPR binding preferences across the same organisms. This conservation would make experimentally extracting and characterising new proteins from these organisms relatively simple, which would drastically increase the number of PPR-RNA pairs of which we are aware.

In order to assess the potential of the PPR protein as a tool for the wider field of synthetic biology, a general model of PPR activity based on known interactions was created. By simulating various interaction networks, it was shown that PPRs can be used

[1] All extracted PPRs are available from *https://github.com/haydnKing/4th-year-project-data/tree/master/PPRs*

to perform various simple logic operations at the translation level within the cell. Since each PPR can be programmed to recognise long sequences of RNA (up to at least 30bp) this leads to an enormous address space for signalling (a maximum of 4^{30}), effectively solving the problem of cross-talk in molecular circuit design.

5.2 Future Work and Directions

Increasing the number of known PPR-RNA will allow us to improve the prediction of how binding specificity is encoded, and will eventually enable us to design PPRs which can bind to any sequence of appropriate length of our choice.

However, a thorough understanding of the rules which govern binding domain preference is only part of the way to realising the potential of PPR-based technology. We must also understand the molecular basis behind the interactions with translation and mRNA stability in order to exploit these effects for our own goals.

Although plausible models exist suggesting RNA stabilisation, ribosome recruitment and degradation of mRNA, and there is evidence for some of these interactions *in vitro* (see Prikryl *et al.*, 2011), there is little characterisation of these mechanisms *in vivo*, nor of how PPRs should be designed to exploit these mechanisms.

Acknowledgements

My thanks to my external supervisor, Dr Jim Haselof (Plant Sciences), who was most helpful in providing direction and inspiration throughout the project, Dr Gos Micklem (Genetics) who provided helpful advice on the bioinformatics of the project, and to my internal supervisor, Dr Jorge Conçalves who provided useful input from an engineering perspective.

Appendix A

Probability of a Binding Domain Appearing by Chance

We are interested in the probability of a particular sequence S of length N appearing within a larger sequence of length M one or more times. Both sequences are assumed to be made up of an alphabet of size K, and the larger sequence is assumed to be a series of independent, uniform draws where each symbol has equal probability $\frac{1}{K}$.

In order to avoid some rather involved combinatorics, we calculate the expected number of draws until S appears. Letting a equal the expected number of draws until we draw the first symbol in S, we see that

$$\begin{aligned} a &= 1 + \frac{K-1}{K}\, a \\ \Rightarrow a &= K \end{aligned}$$

since we must draw at least one symbol and there is a $\frac{K-1}{K}$ chance of us being returned to a.

Now let b equal the number of draws until we find the first two symbols in S. The second symbol is drawn on the $(a+1)^{\text{th}}$, and these is a $\frac{K-1}{K}$ chance of us returning to b so that

$$\begin{aligned} b &= (a+1) + \frac{K-1}{K}\, b \\ \Rightarrow b &= K(K+1) \end{aligned}$$

Having seen this pattern, we can easily invoke proof by induction to show that the

expected number of draws until the entirety of S is drawn is

$$\sum_{n=1}^{n=N} K^n$$

This, the expected number of occurrences of S in a random sequence of length M is

$$\frac{M}{\sum_{n=1}^{n=N} K^n}$$

We can now use the Poisson distribution to calculate the probability of at least one occurrence of S, which is given by

$$P = e^{-\frac{M}{\sum_{n=1}^{n=N} K^n}}$$

Since out binding site can be located on the forward or reverse strand, we are actually interested in finding the sequence S itself, or the reverse complement of S, which is equally likely as out probability distribution is uniform. The probability of finding at least one instance of S or it's reverse complement is thus given by

$$2P - P^2$$

Appendix B

Risk Assessment Retrospective

The risk assessment submitted was an accurate representation of the project, which as a purely computational project did not incur any risks other than the normal problems associated with extended computer use.

Bibliography

M. E. Allentoft, M. Collins, D. Harker, J. Haile, C. L. Oskam, M. L. Hale, P. F. Campos, J. A. Samaniego, T. P. Gilbert, E. Willerslev, G. Zhang, R. P. Scofield, R. N. Holdaway, and M. Bunce. The half-life of dna in bone: measuring decay kinetics in 158 dated fossils. *Proc R Soc B*, 279:4724–4733, 2012.

J. B. Andersen, C. Sternberg, L. K. Poulsen, S. P. Bjorn, M. Givskov, and S. Molin. New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. *Applied and environmental microbiology*, 64(6):2240–6, June 1998. ISSN 0099-2240.

A. Barkan, M. Rojas, S. Fujii, A. Yap, Y. S. Chong, C. S. Bond, and I. Small. A combinatorial amino Acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS genetics*, 8(8):e1002910, Aug. 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002910.

S. Bentolila, A. A. Alfonso, and M. R. Hanson. A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. *Proceedings of the National Academy of Sciences of the United States of America*, 99(16):10887–92, Aug. 2002. ISSN 0027-8424. doi: 10.1073/pnas.102301599.

B. Castandet and A. Araya. RNA editing in plant organelles. Why make it easy? *Biochemistry (Moscow)*, 76(8):924–31, Aug. 2011. ISSN 1608-3040. doi: 10.1134/S0006297911080086.

R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

S. Eddy. *HMMER User's Guide*, Mar. 2010. URL `ftp://selab.janelia.org/pub/software/hmmer3/3.0/Userguide.pdf`.

O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, 2000.

S. Fujii and I. Small. The evolution of RNA editing and pentatricopeptide repeat genes. *The New phytologist*, 191(1):37–47, July 2011. ISSN 1469-8137. doi: 10.1111/j.1469-8137.2011.03746.x.

T. Gaj, C. A. Gersbach, and C. F. Barbas. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in Biotechnology*, May 2013. ISSN 01677799. doi: 10.1016/j.tibtech.2013.04.004. URL `http://www.cell.com/trends/biotechnology/fulltext/S0167-7799(13)00087-5`.

M. J. Howard, W. H. Lim, C. A. Fierke, and M. Koutmos. Mitochondrial ribonuclease P structure provides insight into the evolution of catalytic strategies for precursor-tRNA 5' processing. *Proceedings of the National Academy of Sciences of the United States of America*, 109(40):16149–54, Oct. 2012. ISSN 1091-6490. doi: 10.1073/pnas.1209062109.

T. Kazama, T. Nakamura, M. Watanabe, M. Sugita, and K. Toriyama. Suppression mechanism of mitochondrial ORF79 accumulation by Rf1 protein in BT-type cytoplasmic male sterile rice. *The Plant journal : for cell and molecular biology*, 55(4): 619–28, Aug. 2008. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2008.03529.x.

K. Kobayashi, M. Kawabata, K. Hisano, T. Kazama, K. Matsuoka, M. Sugita, and T. Nakamura. Identification and characterization of the RNA binding surface of the pentatricopeptide repeat protein. *Nucleic acids research*, 40(6):2712–23, Mar. 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr1084.

C. Lurin, C. Andrés, S. Aubourg, M. Bellaoui, F. Bitton, C. Bruyère, M. Caboche, C. Debast, J. Gualberto, B. Hoffmann, A. Lecharny, M. Le Ret, M.-L. Martin-Magniette, H. Mireau, N. Peeters, J.-P. Renou, B. Szurek, L. Taconnat, and I. Small. Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *The Plant cell*, 16(8):2089–103, Aug. 2004. ISSN 1040-4651. doi: 10.1105/tpc.104.022236.

T. Matsuo, K. Onai, K. Okamoto, and et al. Real-time monitoring of chloroplast gene expression by a luciferase reporter: Evidence for nuclear regulation of chloroplast circadian period. *Molecular and Cellular Biology*, 26:863–870, 2006. ISSN 3. doi: 10.1128/MCB.26.3.863-870.2006.

T. Nakamura, Y. Yagi, and K. Kobayashi. Mechanistic insight into pentatricopeptide repeat proteins as sequence-specific RNA-binding proteins for organellar RNAs in plants. *Plant & cell physiology*, 53(7):1171–9, July 2012. ISSN 1471-9053. doi: 10.1093/pcp/pcs069.

K. Okuda and T. Shikanai. A pentatricopeptide repeat protein acts as a site-specificity factor at multiple RNA editing sites with unrelated cis-acting elements in plastids. *Nucleic acids research*, 40(11):5052–64, June 2012. ISSN 1362-4962. doi: 10.1093/nar/gks164.

K. Okuda, F. Myouga, R. Motohashi, K. Shinozaki, and T. Shikanai. Conserved domain structure of pentatricopeptide repeat proteins involved in chloroplast RNA editing. *Proceedings of the National Academy of Sciences of the United States of America*, 104 (19):8178–83, May 2007. ISSN 0027-8424. doi: 10.1073/pnas.0700865104.

J. Pfalz, O. A. Bayraktar, J. Prikryl, and A. Barkan. Site-specific binding of a PPR protein defines and stabilizes 5' and 3' mRNA termini in chloroplasts. *The EMBO journal*, 28(14):2042–52, July 2009. ISSN 1460-2075. doi: 10.1038/emboj.2009.121.

J. Prikryl, M. Rojas, G. Schuster, and A. Barkan. Mechanism of RNA stabilization and translational activation by a pentatricopeptide repeat protein. *Proceedings of the National Academy of Sciences of the United States of America*, 108(1):415–20, Jan. 2011. ISSN 1091-6490. doi: 10.1073/pnas.1012076108.

R. Ringel, M. Sologub, Y. I. Morozov, D. Litonin, P. Cramer, and D. Temiakov. Structure of human mitochondrial RNA polymerase. *Nature*, 478(7368):269–73, Oct. 2011. ISSN 1476-4687. doi: 10.1038/nature10435.

D.-K. Ro, E. M. Paradise, M. Ouellet, K. J. Fisher, K. L. Newman, J. M. Ndungu, K. A. Ho, R. A. Eachus, T. S. Ham, J. Kirby, M. C. Y. Chang, S. T. Withers, Y. Shiba, R. Sarpong, and J. D. Keasling. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440:940–943, Apr. 2006.

E. H. Rubinson and B. F. Eichman. Nucleic acid recognition by tandem helical repeats.

Current opinion in structural biology, 22(1):101–9, Feb. 2012. ISSN 1879-033X. doi: 10.1016/j.sbi.2011.11.005.

C. Schmitz-Linneweber and I. Small. Pentatricopeptide repeat proteins: a socket set for organelle gene expression. *Trends in plant science*, 13(12):663–70, Dec. 2008. ISSN 1360-1385. doi: 10.1016/j.tplants.2008.10.001.

J. Schmutz, S. B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D. L. Hyten, Q. Song, J. J. Thelen, J. Cheng, D. Xu, U. Hellsten, G. D. May, Y. Yu, T. Sakurai, T. Umezawa, M. K. Bhattacharyya, D. Sandhu, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Barry, M. Futrell-Griggs, B. Abernathy, J. Du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X.-C. Zhang, K. Shinozaki, H. T. Nguyen, R. A. Wing, P. Cregan, J. Specht, J. Grimwood, D. Rokhsar, G. Stacey, R. C. Shoemaker, and S. A. Jackson. Genome sequence of the palaeopolyploid soybean. *Nature*, 463(7278):178–83, Jan. 2010. ISSN 1476-4687. doi: 10.1038/nature08670. URL `http://www.ncbi.nlm.nih.gov/pubmed/20075913`.

I. D. Small and N. Peeters. The PPR motif - a TPR-related motif prevalent in plant organellar proteins. *Trends in Biochemical Sciences*, 25(2):46–47, 2000. ISSN 0376-5067.

L.-H. So, A. Ghosh, C. Zong, L. A. Sepúlveda, R. Segev, and I. Golding. General properties of transcriptional time series in Escherichia coli. *Nature genetics*, 43(6):554–60, June 2011. ISSN 1546-1718. doi: 10.1038/ng.821.

M. Sugita and M. Sugiura. Regulation of gene expression in chloroplasts of higher plants. *Plant Molecular Biology*, 32:315–326, 1996. ISSN 0167-4412. doi: 10.1007/BF00039388.

M. Sugiura, T. Hirose, and M. Sugita. Evolution and mechanism of translation in chloroplasts. *Annual review of genetics*, 32:437–59, Jan. 1998. ISSN 0066-4197. doi: 10.1146/annurev.genet.32.1.437.

N. Sun and H. Zhao. Transcription activator-like effector nucleases (TALENs): A highly efficient and versatile tool for genome editing. *Biotechnology and Bioengi-*

neering, pages n/a–n/a, Mar. 2013. ISSN 00063592. doi: 10.1002/bit.24890. URL `http://www.ncbi.nlm.nih.gov/pubmed/23508559`.

J. J. Tabor, H. M. Salis, Z. B. Simpson, A. A. Chevalier, A. Levskaya, E. M. Marcotte, C. A. Voigt, and A. D. Ellington. A synthetic genetic edge detection program. *Cell*, 137:1272 – 1281, 2009.

Z. Wang, Y. Zou, X. Li, Q. Zhang, L. Chen, H. Wu, D. Su, Y. Chen, J. Guo, D. Luo, Y. Long, Y. Zhong, and Y.-G. Liu. 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. *The Plant cell*, 18(3):676–87, Mar. 2006. ISSN 1040-4651. doi: 10.1105/tpc.105.038240.

WHO. World malaria report 2011. Technical report, World Health Organization, 2011.

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*, 8(3):e57286, Jan. 2013a. ISSN 1932-6203. doi: 10.1371/journal.pone.0057286.

Y. Yagi, M. Tachikawa, H. Noguchi, S. Satoh, J. Obokata, and T. Nakamura. Pentatricopeptide repeat proteins involved in plant organellar RNA editing. *RNA biology*, 10 (9), May 2013b. ISSN 1555-8584.

W. Zerges. Translation in chloroplasts. *Biochimie*, 82(6-7):583–601, June 2000. ISSN 03009084. doi: 10.1016/S0300-9084(00)00603-9.