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: _____

A Cross-Genome Study of the Pentatricopeptide Repeat Protein

by

Haydn King (JE)

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

Technical Abstract

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Contents

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. Naturally, many of these processes are highly complicated and 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.

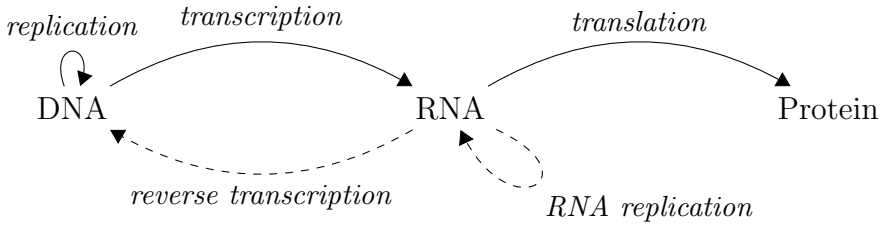

Figure 1.1: The main processes in molecular biology. The three most common are shown using solid lines while two important but less common processes are shown in dotted lines.

Molecules of DNA are the cell's long term storage mechanism – recent research estimates the half-life of DNA to be 521 years[1]. DNA molecules are long sequences of simple nucleotides which encode all the genetic information of the cell. Each nucleotide contains a nucleobase which is either Adenine, Guanine, Thymine or Cytosine (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 chain which is only read in one direction, known as the 5' to 3' direction. Each base in the chain forms a hydrogen bond with a particular base from a complementary chain of DNA, forming a double stranded structure. These 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 an 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 a chain of amino acids which fold into a very specific shape and perform many important functions within the cell. The region of DNA which encodes 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 one-to-one to that of RNA, $\{A, C, G, U\}$, where thymine is replaced with uracil. Transcription is clearly 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 tran-

scription unit that initiate transcription by causing RNA polymerase to bind. They are terminated by terminator regions, which cause the RNA polymerase to cease transcription and release the mRNA. Modulating promoter activity in response to the concentration of another molecule is a common control motif.

Transcription units often contain non-coding regions called *introns* which are removed from the message in a process called RNA splicing before translation. Introns do not contain any useful sequence and are often present in genes and tend to complicate matters significantly as efforts to predict their location accurately and reliably have thus far failed.

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 lower organisms (prokaryotes) translation begins immediately, possibly concurrently with 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 64 possible codons and only 20 amino acids, the code is over complete – several different codons map to the same amino acid. As well as coding for amino acids, three special codons (UAG, UAA and UGA) are known as stop codons as they 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.

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 transcribed at a constant rate for the concentration of its protein to remain constant.

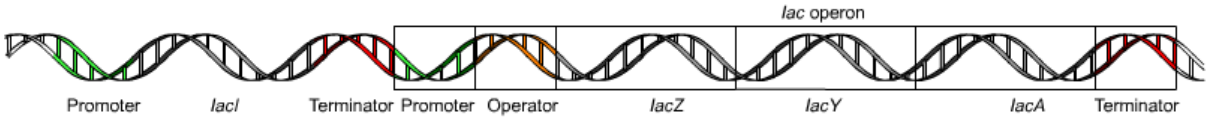

Figure 1.2: Annotated diagram of the lac operon. It contains two transcription units and a total of four genes. The first (leftmost) unit contains *lacI* and is expressed constitutively (continuously). The protein which is produced is called the lac repressor, and in the absence of lactose it binds tightly 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 allowing transcription of the second unit.

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.

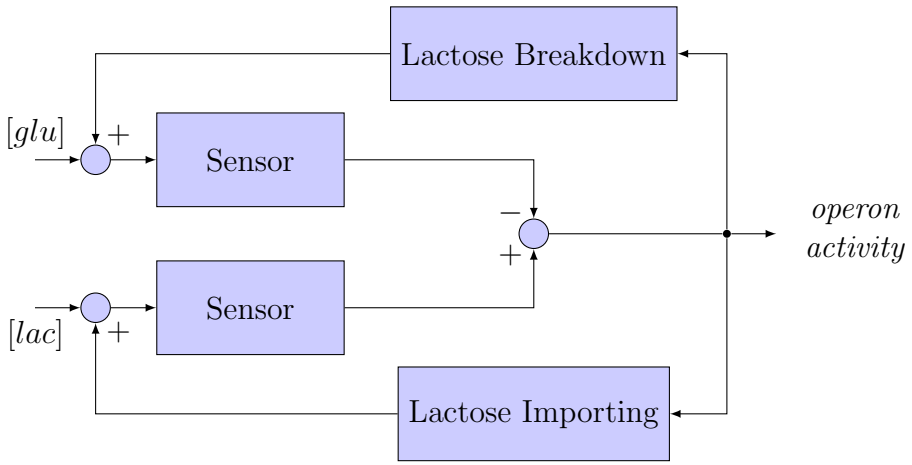

Figure 1.3: Simplified block diagram of the lac operon, showing only the most important interconnections. In the presence of lactose, transcription is turned on and more extracellular lactose is pumped into the cell, causing a positive feedback loop. Simultaneously, lactose is broken down into glucose (and galactose) which inhibits transcription, causing a negative feedback loop.

1.1.3 Control

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. The cell would prefer 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 it's lactose processing machinery when only lactose is available. This is achieved by the lac operon as described in figure 1.2 and shown schematically in figure 1.3.

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. Synbio 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 design new organisms, but tinkers with existing ones without trying to understand the underlying principals.

Synbio can be thought of as programming, but with DNA instead of machine code. An example project which captures this idea is Tabor's bacterial edge detector[2]. 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 is exposed to a pattern of light and dark, the messenger is produced in the dark regions and diffuses into the light, where it stimulates the production of the pigment, leading to an edge detection like effect.

While this and other such simple demonstrations show some of the potential of synbio, they lack immediate application and are of somewhat limited scope. A major problem

in expanding this work is the lack of targeted reporter molecules. 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. Since reusing the same promoter/signal combination in the same cell is impossible due to cross-talk, there are simply not enough signalling modalities 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 more applicable synbio 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 in 2010 caused roughly 2,000 deaths *per day*, mainly because it mostly affects the developing world where access to anti-malarials is poor. 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[3]. Artemisinin produced in this manner has yet to be approved for sale, but it is hoped that it should be available at some point during 2013, at a considerably lower price than any other known method of production.

The major limiting factor in this project 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 directed though it. This led to a difficult balance – too little 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. It is desirable therefore 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 and the PPR Family

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 cells can be found within each plant cell, where they make an excellent target for synbio. They are similar to previous synbio hosts, but with access to the more sophisticated plant cell machinery and superb potential for biosynthesis. The native enzyme RuBisCO is so abundant in the chloroplasts that it can be up to 50% of overall soluble leaf protein.

Chloroplasts contain their own DNA and expression machinery, separate from that of the plant cell, however, they do not appear to make use of the type of control seen in the lac operon – instead mRNA levels appear to be constant[4] implying that transcription in chloroplasts is always on. However there are clear variations in the protein levels in chloroplasts (for example during the day/night cycle) and so control must instead be implemented using post-transcriptional RNA processing and/or interactions with translation.

Understanding how expression control is achieved in plant chloroplasts is a key step in developing the potential of these organisms. Without such an understanding, attempts to engineer synthesis pathways in them would be reduced to genetic engineering.

One such control mechanism is the PPR family of proteins, a class of proteins which are very common in plants but almost entirely absent in other organisms. PPR proteins are encoded in the nucleus and are transported to organelles such as the chloroplast by the cell. Once in the organelle, PPRs bind to mRNA molecules and modulate the translation rate of the proteins encoded in the mRNA molecule. PPRs are a mechanism for the cell nucleus to directly control activity in its chloroplasts.

A thorough review of the PPR family is given in section 2.1.

Chapter 2

Literature Review

A summary of all the information currently known about the PPR protein is given in the first part of this chapter, demonstrating their potential for applications in synthetic biology. The remainder of the chapter introduces the Hidden Markov Model (HMMs) and the main software tool used by bioinformaticians when modelling using HMMs.

2.1 The PPR Family

2.1.1 The Chloroplast

The chloroplast is thought to be the result of an ancient symbiosis where small energy producing cells were enveloped by the larger plant cells. Chloroplasts contain limited genetic information – they contain an expression system capable of transcription and translation as well as several genes vital to the photosystem, but many of the proteins found in them are expressed by the plant nucleus and then imported into the chloroplast [5, 6].

Most genes in the chloroplast are transcribed constitutively[4] and are thus only controlled at a post-transcriptional level. It is 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 correct location [7, 8].

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 [9]. 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 [10]. 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.

2.1.2 Discovery and Classification of the PPR Family

The PPR family was first identified by Small and Peeters in the year 2000 and are a large family of similar proteins commonly found in the nuclear genome of most plants. They are defined by their tandem degenerate repeating motifs[11]. These repeat motifs are referred to as PPR motifs and share many similarities with the tetratricopeptide repeat (TPR) motif which are known to aid protein-protein binding.

The typical PPR protein sequence contains three regions, the first of which is a signal peptide which targets the protein to a particular organelle. This mechanism 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 the repeating PPR motif array which contains between 2 and 30 PPR motifs. 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[12].

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 motifs 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 [13]. The precise function of the tail remains unknown, but evidence suggests that it is related to the known RNA editing functions of some PPRs[14].

2.1.3 Known interactions with mRNA

PPRs can regulate gene expression and are involved in a variety of post-transcriptional RNA processing steps such as RNA editing, splicing and stability[15, 16].

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[14, 17]. A particular protein can be responsible for several edits by having multiple binding sites within the chloroplast genome[18], 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[19, 20]. 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)[21], which is of commercial importance in breeding. PPRs prevent sterility by preventing the production of specific proteins which cause the condition[22]. While the specific interaction preventing translation is unknown, it is thought to be due to cleavage of the mRNA transcript or degradation[23]. It is also possible that the PPR out competes the ribosome when binding to the ribosome binding site.

Binding Rules

The mechanism behind PPR-RNA interactions 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

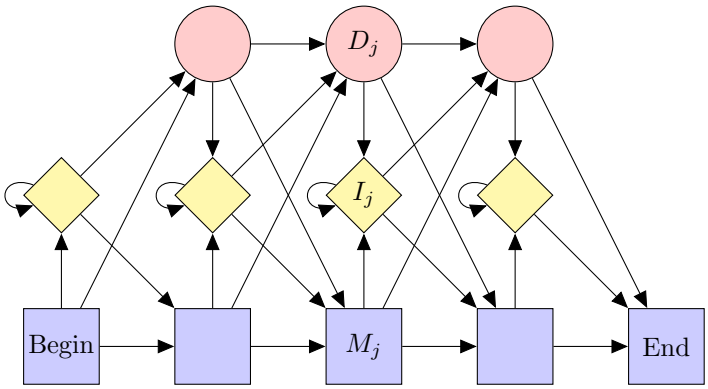

Figure 2.1: A profile HMM. Squares are *match* states, which emit a symbol according to the consensus sequence, diamonds show *insert* states, which insert one or more symbols into the sequence and circles are *delete* states which emit no symbols and effectively skip one or more symbols in the model.

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.

structure[24, 25], suggesting that PPR binding might be similar to that of TALE and PUF repeats[26].

As of writing, two major theories on the PPR binding rules exist, one due to Barkan[27] and another due to Yagi[28]. Both are based on statistical inference on the small number of characterised PPR-RNA interactions and are discussed in more length in section 3.2.

2.2 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[29]. 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.

2.2.1 HMMER

HMMER is a collection of command line tools implementing all the basic algorithms required for using HMMs in a biological context[30]. 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 2.1 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 [29].

Searches are performed using the *hmmsearch* program which makes heavy use of heuristics in order to efficiently search 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.

Chapter 3

Automating PPR discovery

A brief overview of extraction and comparison

3.1 Automated PPR Detection and Extraction

Introduce the problems of detection and extraction.

3.1.1 pyHMMER

Explain the need for a HMMER wrapper and discuss the development of pyHMMER, drawing attention to the github repo.

3.1.2 Detection, Extraction and Annotation

How to spot a chain of PPR repeats

3.1.3 Comparison to Existing Data

Compare the number of PPRs found to those found in the paper in arabidopsis.

3.1.4 Expansion to Other Plants

Discuss the expansion to other plants, and the problems faced (larger genomes, bugs in HMMER).

3.2 Predicting PPR Binding regions

Introduce the problem and the data available.

3.2.1 Direct HMMs

Explain this method and why it failed

3.2.2 Direct PSSM

Explain why this method was more successful, but why it fails to recognise a precise binding region and how this problem was overcome.

3.2.3 Comparison of PSSMs

Explain the method, its strengths and its limitations.

Chapter 4

Discussion of Results

4.1 PPR Survey

A discussion of the results of the PPR survey, the number of PPRs in each plant and their connection.

4.2 PPR Homology

A discussion of any extra homology found between PPRs with similar binding preferences.

Chapter 5

Conclusions and Further Work

5.1 Summary of the Work

Summary of everything that was done as a whole, including the key contributions to the field.

5.2 Future Work and Directions

A discussion of what else needs doing and what can be done to improve the characterisation of the promoters and to improve the usefulness of the software written during the project.

Bibliography

[1] Morten E. Allentoft, Matthew Collins, David Harker, James Haile, Charlotte L. Oskam, Marie L. Hale, Paula F. Campos, Jose A. Samaniego, Thomas P. Gilbert, Eske Willerslev, Guojie Zhang, R. Paul Scofield, Richard N. Holdaway, and Michael Bunce. The half-life of dna in bone: measuring decay kinetics in 158 dated fossils. *Proc R Soc B*, 279:4724–4733, 2012.

[2] Jeffrey J. Tabor, Howard M. Salis, Zachary Booth Simpson, Aaron A. Chevalier, Anselm Levskaya, Edward M. Marcotte, Christopher A. Voigt, and Andrew D. Ellington. A synthetic genetic edge detection program. *Cell*, 137:1272 – 1281, 2009.

[3] Dae-Kyun Ro, Eric M. Paradise, Mario Ouellet, Karl J. Fisher, Karyn L. Newman, John M. Ndungu, Kimberly A. Ho, Rachel A. Eachus, Timothy S. Ham, James Kirby, Michelle C. Y. Chang, Sydnor T. Withers, Yoichiro Shiba, Richmond Sarpong, and Jay D. Keasling. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440:940–943, April 2006.

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

[5] C G Kurland and S G Andersson. Origin and evolution of the mitochondrial proteome. *Microbiology and molecular biology reviews : MMBR*, 64(4):786–820, December 2000. ISSN 1092-2172.

[6] Debashish Bhattacharya, John M Archibald, Andreas P M Weber, and Adrian Reyes-Prieto. How do endosymbionts become organelles? Understanding early events in plastid evolution. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 29(12):1239–46, December 2007. ISSN 0265-9247. doi: 10.1002/bies.20671.

[7] M Sugiura, T Hirose, and M Sugita. Evolution and mechanism of translation in

chloroplasts. *Annual review of genetics*, 32:437–59, January 1998. ISSN 0066-4197. doi: 10.1146/annurev.genet.32.1.437.

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

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

[10] Sota Fujii and Ian 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.

[11] 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.

[12] Keiko Kobayashi, Masuyo Kawabata, Keizo Hisano, Tomohiko Kazama, Ken Matsuoka, Mamoru Sugita, and Takahiro Nakamura. Identification and characterization of the RNA binding surface of the pentatricopeptide repeat protein. *Nucleic acids research*, 40(6):2712–23, March 2012. ISSN 1362-4962. doi: 10.1093/nar/gkr1084.

[13] Claire Lurin, Charles Andrés, Sébastien Aubourg, Mohammed Bellaoui, Frédérique Bitton, Clémence Bruyère, Michel Caboche, Cédrig Debast, José Gualberto, Beate Hoffmann, Alain Lecharny, Monique Le Ret, Marie-Laure Martin-Magniette, Hakim Mireau, Nemo Peeters, Jean-Pierre Renou, Boris Szurek, Ludivine Taconnat, and Ian Small. Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *The Plant cell*, 16(8):2089–103, August 2004. ISSN 1040-4651. doi: 10.1105/tpc.104.022236.

[14] Yusuke Yagi, Makoto Tachikawa, Hisayo Noguchi, Soichirou Satoh, Junichi Obokata, and Takahiro Nakamura. Pentatricopeptide repeat proteins involved in plant organellar RNA editing. *RNA biology*, 10(9), May 2013. ISSN 1555-8584.

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

[16] Takahiro Nakamura, Yusuke Yagi, and Keiko 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.

[17] Kenji Okuda, Fumiyoshi Myouga, Reiko Motohashi, Kazuo Shinozaki, and Toshiharu 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.

[18] Kenji Okuda and Toshiharu 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.

[19] Jeannette Pfalz, Omer Ali Bayraktar, Jana Prikryl, and Alice 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.

[20] Jana Prikryl, Margarita Rojas, Gadi Schuster, and Alice 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, January 2011. ISSN 1091-6490. doi: 10.1073/pnas.1012076108.

[21] Stephane Bentolila, Antonio A Alfonso, and Maureen 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, August 2002. ISSN 0027-8424. doi: 10.1073/pnas.102301599.

[22] Tomohiko Kazama, Takahiro Nakamura, Masao Watanabe, Mamoru Sugita, and Kinya 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, August 2008. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2008.03529.x.

[23] Zhonghua Wang, Yanjiao Zou, Xiaoyu Li, Qunyu Zhang, Letian Chen, Hao Wu, Dihua Su, Yuanling Chen, Jingxin Guo, Da Luo, Yunming Long, Yang Zhong, and Yao-Guang 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, March 2006. ISSN 1040-4651. doi: 10.1105/tpc.105.038240.

[24] Rieke Ringel, Marina Sologub, Yaroslav I Morozov, Dmitry Litonin, Patrick Cramer, and Dmitry Temiakov. Structure of human mitochondrial RNA polymerase. *Nature*, 478(7368):269–73, October 2011. ISSN 1476-4687. doi: 10.1038/nature10435.

[25] Michael J Howard, Wan Hsin Lim, Carol A Fierke, and Markos 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, October 2012. ISSN 1091-6490. doi: 10.1073/pnas.1209062109.

[26] Emily H Rubinson and Brandt F Eichman. Nucleic acid recognition by tandem helical repeats. *Current opinion in structural biology*, 22(1):101–9, February 2012. ISSN 1879-033X. doi: 10.1016/j.sbi.2011.11.005.

[27] Alice Barkan, Margarita Rojas, Sota Fujii, Aaron Yap, Yee Seng Chong, Charles S Bond, and Ian Small. A combinatorial amino Acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS genetics*, 8(8):e1002910, August 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002910.

[28] Yusuke Yagi, Shimpei Hayashi, Keiko Kobayashi, Takashi Hirayama, and Takahiro Nakamura. Elucidation of the RNA recognition code for pentatricopeptide repeat

proteins involved in organelle RNA editing in plants. *PloS one*, 8(3):e57286, January 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0057286.

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

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