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.

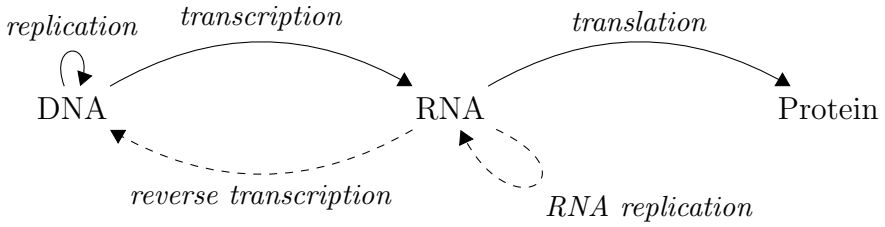

Figure 1.1: The main processes in molecular biology. The three most common are shown using solid lines while two important but less common processes are shown in dotted lines.

Molecules of DNA are the cell's long term storage mechanism – recent research estimates the half-life of DNA to be 521 years(Allentoft et al., 2012). 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 transcription 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.

Transcripts often include non-coding regions at either end called the 5' and 3' untraslated regions (UTR) respectively. They also contain other non-coding regions called *introns* which often found within gene sequences and are removed from the 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.

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. The point where the ribosome binds is calls a ribosome binding sequence (RBS), and one of the most common sites is a Shine-Dalgarno sequence, which is typically found a few bases upstream of a start codon.

Because of the 3:1 nature of translation, proteins are sensitive to frame shifts – where translation is shifted by one or two bases, the resulting amino acid is effectively unrelated

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.

to the encoded one. This can be the case if an 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 transcribed 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. The cell would prefer to directly metabolise glucose if it is available as lactose is harder to process, and so the cell can save

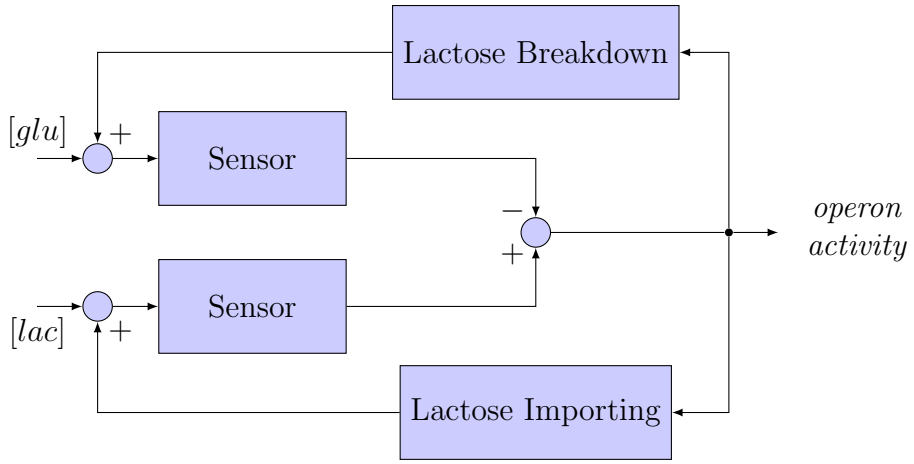

Figure 1.3: Simplified block diagram of the lac operon, showing only the most important interconnections. In the presence of lactose, transcription is turned on and more extracellular lactose is pumped into the cell, causing a positive feedback loop. Simultaneously, lactose is broken down into glucose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

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(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 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(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 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

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 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 does not appear to be used 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 correct location(Sugiura et al., 1998; Zerges, 2000).

This does not reflect the protein levels found in the chloroplast, which vary considerably during the chloroplast's circadian cycle – energy is stored during the day and only consumed at night. Since mRNA levels are constant, this control 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.

1.4 Nucleotide 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 joints between amino acids in a protein are very flexible, 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 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 discovered 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 recognise and bind to a specific triplet of nucleotides, and have been reverse-engineered such that they can now be designed to bind to effectively arbitrary sequences(Gaj et al., 2013).

However, zinc fingers have limited modularity as each motif binds to three bases with

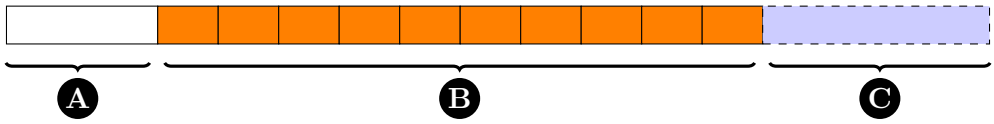

Figure 1.4: The three regions within a PPR protein. (**A**) The signal peptide acts as a header, essentially encoding the address of the target location of the PPR within the cell (**B**) The repeat regions contain 2-30 PPR motifs which specify the binding preferences of the protein (**C**) The tail region is optional contains either E, E andE+ or E, E+ and DYW motifs.

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 code dependent on two hyper-variable amino acids within the repeat motif. TALEs are inherently easier to design than zinc fingers and have been used for numerous novel applications, such as designing new transcription activators which work on arbitrary sequences or manipulating the genome to allow for novel studies of protein function(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 PPR family is a group of proteins commonly found in plants and are known to bind specific RNA sequences and display many similarities with TALEs. They contain tandem degenerate repeating motifs which are referred to as PPR motifs and share many similarities with the tetratricopeptide repeat (TPR) motif which are known to aid protein-protein binding(Small and Peeters, 2000).

PPRs are found exclusively in the nuclear genome, and are commonly targeted 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 1.4. The first 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(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 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 (Lurin et al., 2004). The precise function of the tail remains unknown, but evidence suggests that it is related to the known RNA editing functions of some PPRs(Yagi et al., 2013a).

1.5.2 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(Schmitz-Linneweber and Small, 2008; Nakamura et al., 2012).

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(Yagi et al., 2013a; Okuda et al., 2007). A particular protein can be responsible for several edits 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

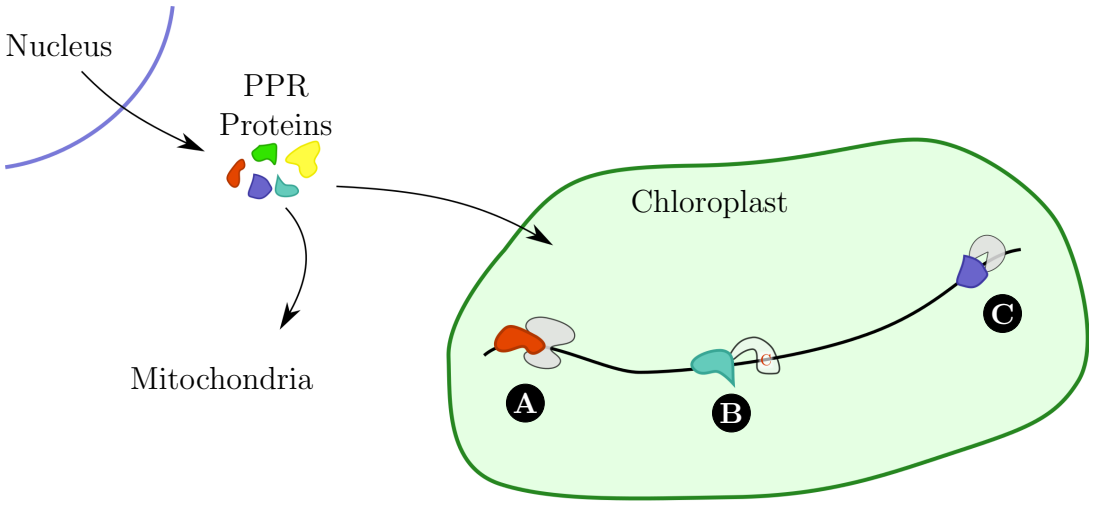

Figure 1.5: Summary of PPR activity once in the chloroplast. Similar processes are known to occur in the mitochondria. (**A**) Increased translation – PPRs can promote ribosome recruitment and increase the translation rate of the transcript (**B**) RNA editing – PPRs edit the RNA transcript, as well 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.

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

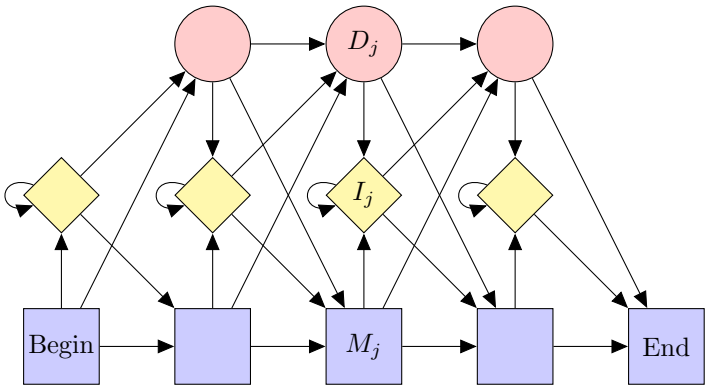

Figure 1.6: A profile HMM. Squares are *match* states, which emit a symbol according to the consensus sequence, diamonds show *insert* states, which insert one or more symbols into the sequence and circles are *delete* states which emit no symbols and effectively skip one or more symbols in the model.

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.

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 structure(Ringel et al., 2011; Howard et al., 2012), suggesting that PPR binding might be similar to that of TALE and PUF repeats(Rubinson and Eichman, 2012).

As of writing, two major theories on the PPR binding rules exist, one due to Barkan(Barkan et al., 2012) and another due to Yagi(Yagi et al., 2013b). 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.

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 1.6 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).

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 2

Automating PPR discovery

Most sequenced genomes are also annotated with information about the genes they encode a specific locations. These annotations are usually very 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 usually fairly unreliable and certainly 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, it is interesting to try and predict their binding locations within the 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 can only compare a protein model with a protein target, and so the genome must be translated before it can be searched. 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 can sometimes cause problems under particular circumstances, the most problematic of which 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. Python was chosen as the main language for this project 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 map the returned matches to the correct location
- Automatically terminate HMMER processes which attempt to allocate more memory than the system can sensibly be expected to provide and then call HMMER sequentially with subsections of the target[2]
- 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 validated PPRs and the best chosen.

[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

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, and still returns relatively 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. Group the motifs into clusters such that motifs which are on the same strand and are within a certain distance are put in the same cluster

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 only 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 very large, 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 only the most probable matches and only a few of the 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 starts 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 1500*bp* 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 yet. For envelopes on the reverse strand, the sequence is extracted such that it reads correctly from left-to-right. 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 orders of magnitudes smaller than the first pass and so the chances of finding a high-scoring match 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 to 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 is not the case 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. 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 extracted 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. (2013b).

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 for the score calculations. This involves calculating the expected score of the model for random sequences in order to assess the significance of a particular match. A program exists called *hmmsim* exists in the HMMER package for just this purpose, however the program's man page gives a hint of things to come –

> "*hmmsim* is not a mainstream part of the HMMER package. Most users would have no reason to use it.
>
> ...
>
> "Because it is a research testbed, you should not expect it to be as robust as other programs in the package. For example, options may interact in weird ways; we haven't tested nor tried to anticipate all different possible combinations."

Unfortunately, one particular unanticipated use case is trying to normalise 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

Discussion of Results

3.1 PPR Survey

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

3.2 PPR Homology

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

Chapter 4

Conclusions and Further Work

4.1 Summary of the Work

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

4.2 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

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.

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.

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

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.

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.

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.

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

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.

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

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.

Thomas Gaj, Charles A. Gersbach, and Carlos 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`.

Ning Sun and Huimin Zhao. Transcription activator-like effector nucleases (TALENs): A highly efficient and versatile tool for genome editing. *Biotechnology and Bioengineering*, pages n/a–n/a, March 2013. ISSN 00063592. doi: 10.1002/bit.24890. URL `http://www.ncbi.nlm.nih.gov/pubmed/23508559`.

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.

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.

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.

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 2013a. ISSN 1555-8584.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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 pentatri-

copeptide repeat proteins. *PLoS genetics*, 8(8):e1002910, August 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002910.

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 2013b. ISSN 1932-6203. doi: 10.1371/journal.pone.0057286.

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

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

Olof Emanuelsson, Henrik Nielsen, Sren Brunak, and Gunnar von Heijne. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, 2000.