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

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 processes and control mechanisms which regulate this process. 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 processes. Figure 1.1 shows the most important of these processes and how they convert between the three most important classes of molecules in the cell.

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

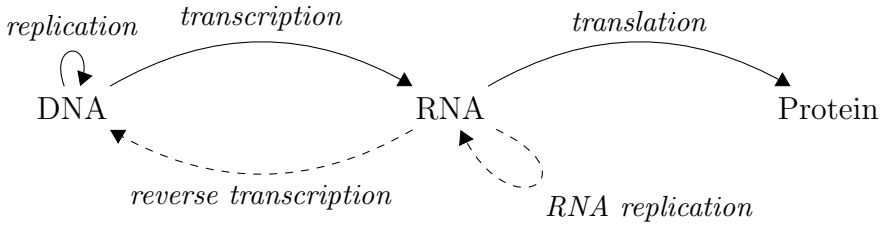

Figure 1.1: The main processes in molecular biology. The three most common are shown using solid lines while two important but less common processes are shown in dotted lines.

contains a nucleobase which is either Adenine, Guanine, Thymine or Cytosine (A, C, G or T) and it is the sequence of so called 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, 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.

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 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 happens as soon as possible, 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 to 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 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

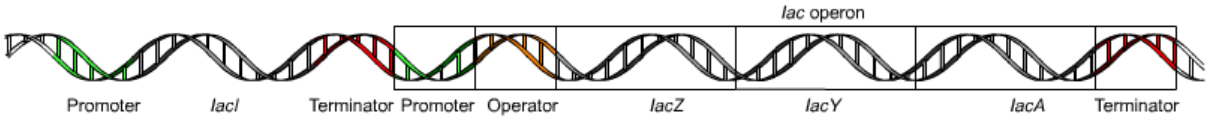

Figure 1.2: Annotated diagram of the lac operon. It contains two transcription units and a total of four genes. The first (leftmost) unit contains *lacI* and is expressed constitutively (continuously). The protein which is produced is called the lac repressor, and in the absence of lactose it binds tightly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

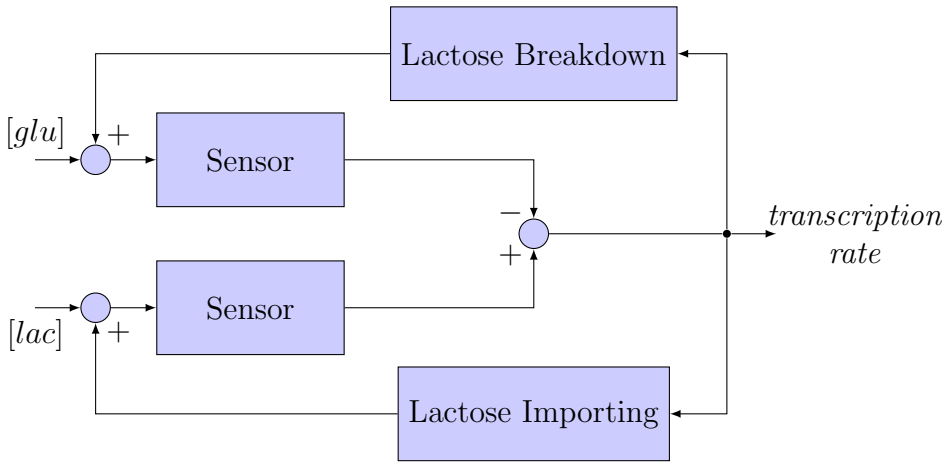

Figure 1.3: Simplified block diagram of the lac operon, showing only the most important interconnections. In the presence of lactose, transcription is turned on and more extracellular lactose is pumped into the cell, causing a positive feedback loop. Simultaneously, lactose is broken down into glucose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

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.

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. 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 pathway involved had to be up-regulated, which 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.

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. Achieving anything remotely close to this figure in a project such

as the production of artemisinin would help reduce the vast number of people who die of this treatable disease each year (roughly 2,000 deaths a day in 2010 [4]).

1.2.1 The PPR Protein

Brief introduciton to what a PPR protein is and why they are interesting, reference Section 2.2 heavily.

Chapter 2

Literature Review

2.1 Hidden Markov Models

2.1.1 Mathematical Description

A description of the Hidden Markov Model (HMM) - a Markov Model whose internal state is unknown, but which emits a symbol from an alphabet with a distribution dependant on the state.

2.1.2 Use in Bioinformatics

A description of how HMMs have been successfully used to predict genes in a number of studies.

2.1.3 HMMER

A description of the HMMER implementation and the algorithms required for HMMs, including:

- Training the model transition and emission probabilities from a data set
- Efficiently representing an HMM in memory
- Calculating a score for a sequence given a set of transition and emission probabilities
- Choosing a decision threshold for promoter regions

Also discuss the limitations of the HMMER software for scripting and its vast memory use issues.

2.2 The PPR Family

Discuss detection/prediction vs extraction. Introduce references.

2.2.1 In *A. Thaliana*

Explain the development of the PPR detection routine for Arabidopsis, as well as why Arabidopsis made a good test organism. Compare directly with the paper about this.

Chapter 3

Experimental Methods

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

How to spot a chain of PPR repeats

3.1.3 Extraction

How to extract that chain as a PPR

3.1.4 Comparison to Existing Data

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

3.1.5 Expantion 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] WHO. World malaria report 2011. Technical report, World Health Organization, 2011.