A Cross-Genome Study of the Pentatricopeptide Repeat (PPR) Protein

Haydn King*

Department of Engineering
University of Cambridge
hjk38@cam.ac.uk

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

I. Introduction and Motivation

Synthetic biology is a relatively new engineering discipline whose goal is to apply standard engineering techniques such as standardisation, characterisation and encapsulation of function to biology. Synbio aims to use 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. For a brief introduction to those principals, see appendix B.

Synbio is often referred to as programming, but with DNA instead of computer code. An example project which captures this idea is Tabor's bacterial edge detector[1]. Bacteria were programmed to produce a colourless chemical messenger in the absence of light and to produce a dark pigment in the presence of light and the chemical messenger. When a film these bacteria is exposed to a pattern of light and dark, the messenger diffuses out from the dark regions and into the light, where it stimulates the production of the pigment, leading to an edge detection effect.

While this and other such simple demonstration shows some of the potential of synbio, they lacks immediate application and are somewhat limited. A major problem in expanding this work is the lack of targeted reporter molecules. In the edge-detector example, two molecular signal are produced when light is not present – AHL, a cell-to-cell signalling molecule and cl, a transcriptional repressor molecule. Both AHL and cl affect the promoter $P_{lux-\lambda}$; while AHL stimulates expression, cl 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 cl on $P_{lux-\lambda}$ is one of a small but growing number of well understood control motifs. Since

*Supervised by: Dr Jorge Gonçalves (Dept. of Engineering), Dr Jim Haseloff (Dept. of Plant Sciences)

reusing the same promoter/signal in the same cell is impossible due to cross-talk, there are simply not enough signalling modalities available to perform more complex calculations 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 are of the cells metabolism to malfunction with undesirable consequences.

Another successful synbio project is 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[2]. 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 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 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 expressed in the chloroplasts where it makes up up to 50% of 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 [3]).

II. Project Background

Understanding how gene expression is controlled in chloroplasts is a key step in achieving this goal. Unlike previous synbio targets, chloroplast genes are generally expressed constitutively (continuously) leading to constant mRNA levels rather than being controlled by promoter regulation[4].

PPR proteins are a class of signalling protein found almost exclusively in plants[5]. They are synthesised in the nucleus and sent to organelles such as the chloroplast where each one binds to a specific RNA sequence. Depending on whether the protein binds to the untranslated region within the RNA message or over the ribosome binding site, expression is either increased (by preventing exonucleases from destroying the mRNA) or decreased by reducing ribosome activity[6].

PPR proteins are made up of a short targeting region followed by a series of repeating regions which form the RNA binding site. Often, the protein also contains a tail region which can be classified as belonging to one of 3 different categories, but the function of which is hitherto unknown[7].

Discovering the target region of a given PPR protein is a difficult problem as nothing is known of the 3D structure of the repeat domains. Statistical analysis of a large set of experimentally verified PPR/RNA pairs has shown a strong link between the amino acids at position 6 in repeat domain n and position 1 in repeat domain $n+1$ (hitherto refered to position 1′) and the bound nucleotide at position n[8]. Weight was added to this conjecture after a PPR protein with known binding target was mutated such as to change its binding

preference in a predictable way[8].

Ultimately, being able to design a PPR protein to bind to an arbitrary RNA sequence with a pre-specified affinity would be of great use to synthetic biology. It would both improve our knowledge of the chloroplast and give us the necessary tools to control expression within it and provide a convenient way to precisely control the expression of a gene without the possibility of cross-talk or interference.

Unfortunately, such an undertaking is infeasible for such a short project, and so the goals of this project are:

- Find and predict the binding targets of PPR proteins from several plant genomes and compare those which correspond to similar binding targets and discover which features of the protein are preserved
- Verify that PPR-style control can be performed in a bacterial setting such as in *E. coli*, by designing and performing a simple test with known PPR/RNA binding pairs

III. Preliminary Results

Table 1: *Example table*

Name		
First name	Last Name	Grade
John	Doe	7.5
Richard	Miles	2

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

$$e = mc^2 \tag{1}$$

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

IV. Further Work

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed

porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Appendices

A. Hidden Markov Models and the HMMER Package

A Hidden Markov Model (HMM) is a statistical model of a Markov Process where the sequence of states is unknown but a symbol is emitted from each state. An HMM has a set of N states, $\Omega = \omega_{1...N}$ and an alphabet of M symbols, $\Psi = \psi_{1...M}$. The probability of emitting the symbol ψ_i from state ω_j is defined as $\theta_{\psi_i|\omega_j}$. Similarly, the probability of transitioning from state ω_j to state ω_i is given by $\phi_{\omega_i|\omega_j}$. The model results in a sequence of states $x(t) \in \Omega$ which are not observed, and a sequence of symbols $y(t) \in \Psi$ which are observed for some range of t.

HMMs have proved useful in a number of fields, but have been particularly useful in modeling biological sequences. In general, bioinformaticians use a special case of the HMM called a profile-HMM or a pHMM. A pHMM is an HMM whose network topology is fixed, as shown in. They contain a number of nodes, each of which contains an emission state, an insert state and a mute delete state, which either emit a single symbol, emit one or more symbols or emit no symbols before moving to the next node respectively.

Profile-HMMs have numerous practical advantages over general HMMs. Firstly,there is a significant reduction in the number of transition states which must calculated and stored and secondly it possible to automatically generate a pHMM from a sequence alignment using the Expectation-Maximisation algorithm as the topology is fixed. More information is available about HMMs and other aspects of bioinformatics in [9].

Many of the algorithms required to build and manipulate pHMMs are implemented in the HMMER[10] package, a free and open source software package available from hmmer.janelia.org.

B. Molecular Biology

Molecular biology is the study of the molecular basis of biology. While the field itself is rather broad, much of it is underpinned by what is refered to as the central dogma of molecular biology. 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 1960s to understand these processes. Below is a brief introduction, aimed at the information or control engineer.

Molecules of DNA are the cell's long term storage mechanism – recent research estimates the half-life of DNA to be 521 years[11]. The first process is called *translation*, where the DNA molecule is 'read' by an RNA polymerase, producing an RNA copy of a section of the DNA. The RNA molecule is called messenger-RNA as it is a short-lived (minutes to hours) message. This message is read by a ribosome, a molecule which translates the mRNA into a protein, a process refered to as *translation*. Proteins then fold into a very specific shape determined by their sequence, and go on to perform many important functions within the cell. The processes of transcription and translation are typically very tightly controlled by the cell, as this is the main way of influencing the levels of various proteins within the cell.

DNA consists of a sequence of four different nucleotides recorded as G,A,T and C. When DNA is transcribed to mRNA, thymine is replaced with uracil, such that the RNA alphabet is represented as G,A,U and C. 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 base pairs

of RNA, which are referred to collectively as a codon. Since there are 4^3 possible codons and only 20 amino acids, the code is 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.

The DNA region which codes for a protein is called a gene, and is marked by a promoter region, to which the RNA polymerase binds at the start of transcription. Control is often achieved by modulating the activity of the promoter, either to enhance or hinder the binding of RNA polymerase. In prokaryotes, the promoter region is usually a short distance upstream from the gene or genes to be transcribed, such that the mRNA sequence contains a short untranslated region, followed by one or more genes and then another short untranslated region.

Ribosomes bind to the mRNA, reading the gene and creating the appropriate protein before detaching from the mRNA. mRNA is more fragile than DNA but is also targeted by exonucleases, a class of enzyme which degrade the RNA molecule, preventing it from producing more protein. Similar process 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 stable.

References

[1] Jeffrey J. Tabor, Howard M. Salis, Zachary Booth Simpson, et al. A synthetic genetic edge detection program. *Cell*, 137:1272 – 1281, 2009.

[2] Dae-Kyun Ro, Eric M. Paradise, Mario Ouellet, et al. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440:940–943, April 2006.

[3] WHO. World malaria report 2011. Technical report, World Health Organization, 2011.

[4] Mamoru Sugita and Masahiro Sugiura. Regulation of gene expression in chloroplasts of higher plants. *Plant Molecular Biology*, 32:315–326, 1996.

[5] 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.

[6] 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.

[7] Claire Lurin, Charles Andrés, Sébastien Aubourg, et al. Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *The Plant cell*, 16(8):2089–103, August 2004.

[8] Alice Barkan, Margarita Rojas, Sota Fujii, et al. A combinatorial amino Acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS genetics*, 8(8):e1002910, August 2012.

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

[10] S.R. Eddy. *HMMER User's Guide*, March 2010.

[11] Morten E. Allentoft, Matthew Collins, David Harker, et al. The half-life of dna in bone: measuring decay kinetics in 158 dated fossils. *Proc R Soc B*, 279:4724–4733, 2012.