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

A brief introduction into the *central dogma* of Molecular biology, aimed at the engineer. References Appendix for a more thorough introduction

1.2 Synthetic Biology

An introduction to SynBio and the philosophy behind it as an engineering discipline. Include information about what is meant by a part and how they can be used, paying particular attention to the promoter.

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.