Computational Detection and Analysis of Polyadenylation Signals

Ashwin Hajarnavis

Darwin College
&
The Wellcome Trust Sanger Institute

March 2005

A dissertation submitted for the degree of
Doctor of Philosophy
at the University of Cambridge

Preface:

This dissertation is the result of my own work and includes nothing, which is the outcome of work done in collaboration except where specifically indicated in the text.

No part of this thesis is being submitted for any other qualification or at any other University.

Acknowledgements:

I should like to thank my supervisor, Richard Durbin, for his constant advice and guidance. The Sanger Institute has been an incredible place to work, on account of the many people I have met there. Numerous though they are, some deserve a special mention. Firstly, Kevin Howe and Ian Korf, for sharing so much of their time, and discussing everything from C to *C. elegans*. Also Lachlan Coin and Thomas Down, for being inspirational sources of technical help. Access to data was made simple, thanks to help from all at WormBase, in particular Daniel Lawson and Keith Bradnam. I am grateful to Sam Griffiths-Jones and Alex Bateman for many ideas, and for the trip into the world of RNA. Finally, current and former members of our lab; Diego DiBernardo, Marc Sohrmann, Irmtraud Meyer, David Carter, Avril Coghlan, and Mark J Minichiello.

This work was funded by the Medical Research Council and The Wellcome Trust.

Summary:

Many computational techniques exist for the prediction of genes from genome sequence, and for their functional characterisation. Less well understood, however, are the sequences that cause processing and regulation of these genes. One such sequence is the polyadenylation signal, which is required for the expression of most eukaryotic genes. The ability to detect polyadenylation signals accurately means that genomes can be annotated to a greater extent. Although this can be carried out in the laboratory, a computational method is much faster and cheaper, especially considering the acceleration in the sequencing of whole genomes.

A particular gain is that a polyadenylation signal prediction also provides a predicted end to the untranslated region (UTR) lying downstream (3') of a gene's stop codon. This region can contain regulatory motifs, which can dictate properties such as when, where, and how a gene is expressed. Knowledge of gene regulation is as important as gene function if we are to try and gain a full understanding of systems biology from genome sequencing.

In this thesis, I present the development of a piece of software for detecting sequence signals in genome sequence.

I then develop a model for the polyadenylation signal in the nematode worm *Caenorhabditis elegans* and show that the predictions are accurate, leading to the publication of good quality 3' UTR data sets.

Models are then built for three other species, and a comparison made with existing methods.

A comparison between polyadenylation signals of *C. elegans* and the closely related *C. briggsae* follows, which leads onto the discovery of a putative regulatory motif, conserved between the ribosomal protein 3' UTR sequences of both species.

Contents