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

1. An Introduction to 3' Ends and Polyadenylation Signals

1.1. Preamble

The high-throughput sequencing of major eukaryotic genomes has led to a sudden abundance of sequence information. This wealth of data represents an extremely useful resource for the scientific community. A genome contains the inherited information required to determine the physiology of an organism. If we can access and interpret the genome, then we can have a much better understanding of the biological processes defining that organism. In its raw, un-interpreted form, a genome sequence does not prove to be a particularly intelligible resource. However, once the genome sequence is subject to interpretation by biological or computational methods, it quickly becomes a collection of many sources of information that can further our knowledge of molecular biology. For instance, an organism's full set of protein coding genes can be found by the use of computer programmes in conjunction with transcript-mapping techniques. For maximum accuracy these methods require supervision by an expert human annotator, who can best integrate computational and biological evidence for accurate delineation of genome sequence. Once the protein repertoire is known, we have a better idea as to the physiological constituents and processes that are possible. The availability of annotated genomes of multiple species allows us to reconcile empirical differences, such as between mice and humans, and interactions, such as those between malaria and mosquitoes, at the level of molecular biology.

A genome contains far more information than that coding for proteins. Some types of sequence, whilst not specifically coding for a protein, are no less important. The reason for this is that the information for when and where proteins are expressed

must somehow be coded in the DNA. Although our current understanding of the phenomenon of protein coding is reasonable, finding protein coding genes only informs us as to what physical processes might be possible at some point in the life cycle. For a full understanding of the molecular biology of a system, it is necessary to know not only what components are involved, but also the circumstances under which each is required, the location, and the amount. This regulatory information is encoded in the DNA sequence of the genome, but interpreting it is not as straightforward as the *in-silico* translation of a coding sequence into a protein sequence.

The expression of a eukaryotic gene is an extremely complex process, starting with chromatin remodelling, transcription, mRNA processing, mRNA transport, translation, and post-translational modification (Alberts et al. 2002). Each of these processes can be regulated separately, thus there are very many factors that have an effect on gene expression. An example is the initiation of transcription (Gill 2001), in which the coordinated and sequential binding of proteins to the promoter region, assembles the transcriptional machinery on the DNA lying upstream of the coding region. These proteins are able to bind the promoter on account of having affinity to particular sequence motifs, which are called binding sites. One particular example of DNA encoding a regulatory signal is in the case of heat shock promoters (Morimoto 1993). Genes preventing cellular damage during heat shock have an increased transcriptional activation during such stress on account of a protein heat shock factor binding to nGAAn inverted repeats, which increases transcriptional initiation activity. Thus the DNA sequence in this region not only codes for a protein with some stress-related function, it also contains signals that specifically indicate this function to the cell. Thus, if a protein of unknown function is shown to have such a regulatory

element, this provides some evidence that can be used to aid functional annotation, add confidence to an existing annotation, or improve an existing gene prediction.

Many other such signals, some very specific and some much more ubiquitous, are also encoded in the DNA. Although our knowledge of proteins, the sequences that encode them, and the tools available for their analysis is commendable, a full understanding of biology relies on our understanding of regulatory sequences and the different mechanisms of regulation. Protein sequences are encoded by a well-understood trinucleotide codon signal, reviewed in (Nirenberg 2004). Sequence characteristics are also responsible for specifying splice sites, restriction sites, (Alberts et al. 2002), DNA bending propensity (Brukner et al. 1995), nucleosome position (Thastrom et al. 2004), and much more.

Building a high-confidence protein repertoire for an organism requires good gene predictions, which can only perform as well as our knowledge of the underlying biology allows (Makarov 2002; Mathe et al. 2002). It has been shown that refining parts of gene prediction models to closer resemble the observed biology results in better gene prediction (Stanke et al. 2003). Hence studying the biological signals that cooperate to specify a gene aids our ability to predict genes and thus further increases our knowledge about an organism's physiology.

It has been suggested that the increase in complexity between organisms such as *C. elegans* and *H. sapiens* cannot be explained by the increase in size of their respective proteomes (Mattick 2001). Furthermore, this paper argues that the difference between the proteomes of individuals cannot account for phenotypic differences, and that it is the regulation of gene expression, particularly that mediated at the RNA level, that adds this layer of complexity. This RNA regulation may exist as non-coding RNA genes (Eddy 2001), or regulatory elements encoded within

transcribed sequence (Griffiths-Jones et al. 2005). Incorporating such information further complicates the already incompletely understood concept of gene regulatory networks, which at the moment tends to focus on transcription factor binding networks (Pritsker et al. 2004) and protein-protein interactions (Walhout et al. 2001).

In this chapter, I aim to set the scene for the research that will follow. I will introduce the biology that is to be studied and extended. I discuss eukaryotic gene structure, in particular the importance of the 3' untranslated region (3' UTR). The polyadenylation signal is found within this region, and I go on to discuss what it is for, and why we might want to be able to detect it.

Unless otherwise stated, notably in chapter 5, the work in this thesis has been carried out on *C. elegans* (The C. elegans Sequencing Consortium) on account of its relatively well annotated genome, and the availability of well-designed tools for accessing genomic information (Stein et al. 2003; Chen et al. 2005). Although there are other model organisms, accurate gene predictions, coupled with good coverage of transcript information, make this an ideal organism for the analyses in subsequent chapters.

1.2. Overview of untranslated region molecular biology

As the name suggests, untranslated regions are not translated into protein. They are, however, transcribed and to understand them better, we must first gain an insight into transcription.

1.2.1. Transcription and eukaryotic gene structure

1.2.1.1. Transcript termination

Figure 1 shows the processes involved as a primary protein coding gene transcript matures into a processed mRNA ready for translation. Following the gene's transcription, introns are spliced out, leaving the region spanning the start of transcription to the translational start site, the coding sequence itself, and a downstream region. Of this whole sequence, only the coding sequence gets translated, and thus the upstream and downstream sequences are known as 5' and 3' untranslated regions, respectively.

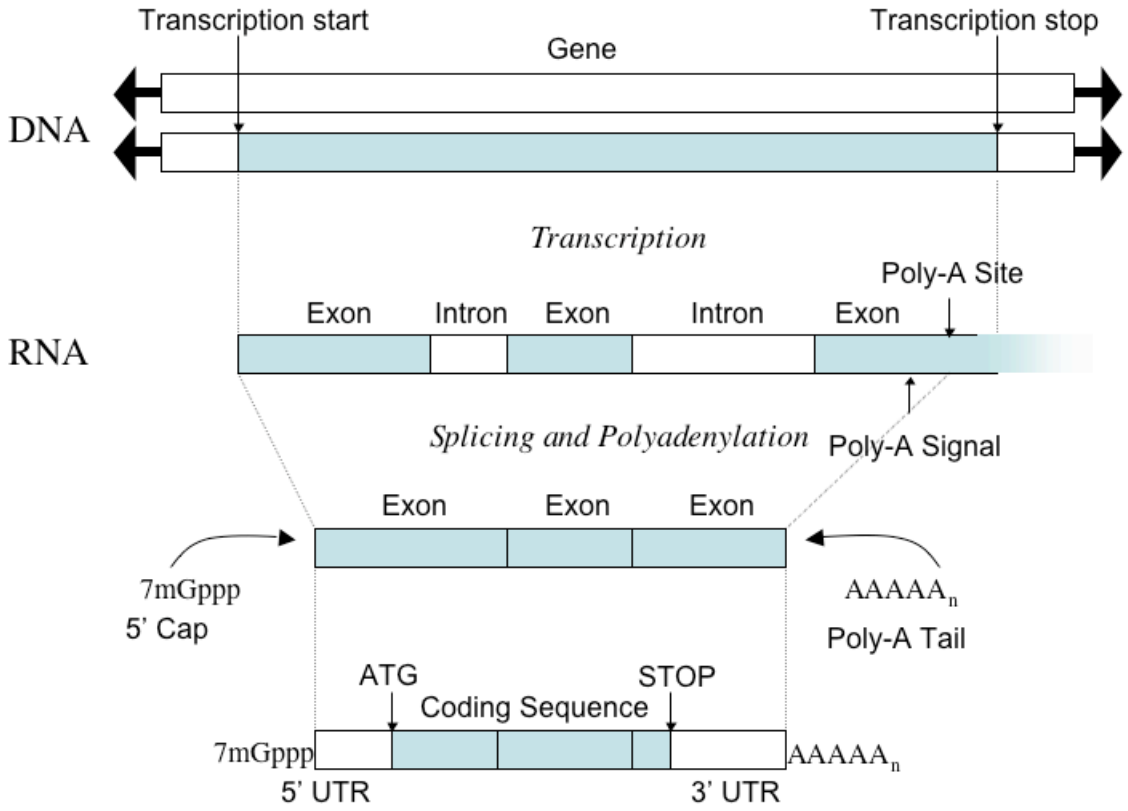

Figure 1. Main steps in the expression of a typical eukaryotic protein coding gene, showing transcription, splicing, and processing.

The termination of RNA polymerase II transcription is a complex process, of which our understanding is still incomplete (Proudfoot et al. 2002). Both computational and experimental transcription stop site annotation have proven to be difficult. Part of this complexity arises from signals which are upstream of the eventual transcriptional termination point. The 3' end of a mature mRNA is not the end of transcription. The RNA polymerase II continues past the known 3' end (Ford et al. 1978). A crucial part of the process of mRNA maturation is the separation of the nascent mRNA from the transcriptional apparatus. This occurs by the cleavage (Colgan et al. 1997) and polyadenylation (reviewed in (Scorilas 2002) of the mRNA. The cleavage separates the transcript from RNA polymerase II, so it can be exported out of the nucleus and translated. The addition of a long polyadenylate tail - of up to 250 nucleotides in mammals (Wahle et al. 1993) - is thought to stabilise the transcript, as it is known that one of the first processes in degradation of such mRNAs is the de-adenylation of the tail (Ford et al. 1997). The RNA lying to the 3' of the cleavage site is eventually degraded and the RNA polymerase II complex is recycled. The primary signal for the recruitment of the cleavage and polyadenylation complex is called the polyadenylation signal; in this thesis we will call this the AATAAA or AAUAAA motif (see chapter 3). A description of this signal and an overview of cleavage are given below, but to appreciate the importance of the polyadenylation signal, it is necessary to understand the sequence context within which it appears.

1.2.1.2. The 3' untranslated region

The 3' untranslated region (3' UTR) is defined as the sequence extending from a protein coding gene's stop codon (UAG, UAA, UGA) up to the point at its 3' end where the transcript is cleaved (Figure 1). As the coding sequence is constrained to code for protein, any regulatory sequence elements required at the post-transcriptional level are much more likely to be encoded in the untranslated regions, which are under much less selective pressure. It is well established that repressor proteins can bind to the 5' UTRs to mediate translational control (Gray 1998; Wilkie et al. 2003), but other factors involved in control of translation of mRNA stability bind to the 3' UTR, as we shall discuss later. *C. elegans* 5' UTRs tend to be short (~75% under 50 nt) on account of the phenomenon of *trans*-splicing (Blumenthal 1995), so we concentrate instead on the 3' UTR.

Regulation by sequence elements in the 3' UTR can have many types of function. These include regulating stability (Xu et al. 1997) of powerful signalling agents in the immune system, and inhibiting translation (Olsen et al. 1999) of developmental genes in appropriate stages of development. A characterised 3' UTR motif allows mRNA localization (Gavis et al. 1996) to specify the *Drosophila* posterior pole. Additionally, in the case of selenoproteins (Hubert et al. 1996), a 3' UTR stem-loop allows the alternative interpretation of a UGA stop codon into an insertion site for Sec-tRNA$_{Sec}$. Mutations in the 3' UTR are known to cause human diseases, notably in the cases of myotonic dystrophy (Timchenko 1999), and alpha-thalassaemia (Higgs et al. 1983).

All of these forms of post-transcriptional regulation are essential for understanding the biology of eukaryotes. No amount of protein sequence analysis can possibly elucidate the control mechanisms involved, and for this reason, sequencing

and functional characterisation of 3' UTRs is as important as that of coding sequences.

A number of regulatory elements identified by a variety of biochemical analyses and computational verification have been collected into a database (Mignone et al. 2005). However, the size and specificity of these motifs makes it impossible to search for most of them accurately at the genome level. There are too many false positive matches to the consensus pattern. To restrict the search space, it is necessary to search just within 3' UTR sequences. Similarly, if we are to try and discover novel regulatory motifs by computational methods, then it is again necessary to discard the non-3' UTR genome from any such analysis. It is therefore important to identify the end point of the 3' UTR, the cleavage and polyadenylation site.

1.2.2. Reliable 3' UTR sets

The standard method to identify 3' UTR sequences is to align cDNAs such as expressed sequence tags (ESTs) back to genome sequence. We also need gene annotations showing the coding regions. cDNAs are typically made from mRNAs by using an oligo dT primer to bind to the polyA tail of the mRNA, which then forms a substrate for reverse transcription into DNA. Theoretically, the full length mRNA is thus copied into DNA, which can be amplified and sequenced. Thus, a high throughput EST project provides evidence for what parts of the genome are transcribed. As mentioned earlier, the whole 3' UTR is transcribed, and thus aligning ESTs to the genome can give us the end point of the 3' UTR. To obtain the start of the UTR, we need to identify the stop codon from the genes' annotation.

Theoretically, a genome sequence, coupled with gene annotations and ESTs, should be enough to build a set of 3' UTRs for all genes. However, there are four further points preventing the establishment of a perfect set. Firstly, the organism in question needs a high throughput EST project. *C. elegans* has one (Kohara, unpublished), but the related nematode *C. briggsae*, for example, does not. Secondly, the project needs to cover a large proportion of the genes in the whole genome. By its nature, the manufacture of cDNAs is difficult for genes expressed in very small amounts or in highly specialised conditions. Hence, there is only EST coverage for approximately half the *C. elegans* gene set. Thirdly, a small but significant problem is that of internal priming; if a gene contains an internal poly-A tract, perhaps because of a poly-lysine tract in the protein, then the oligo-dT primer may map to this tract, instead of the polyadenylate tail at the end of the transcript. The final and most significant problem with ESTs from *C. elegans* (and other organisms) is that a large number of them have been clipped at the 3' end for reasons of sequencing accuracy. As we shall see in chapter 3, some UTRs have been clipped up to 80 nt short of the real cleavage and polyadenylation site. All of these factors serve to reduce the size and accuracy of the search space within which known and novel 3' UTR regulatory elements occur.

A solution to the species, coverage, and end-clipping problems is to predict the site at which cleavage and polyadenylation occurs. This method requires only a good gene coding sequence annotation, and can generate full-length 3' UTR sequences. In the case of end-clipping, the prediction can be used in conjunction with partial EST coverage to identify cleavage sites with higher confidence.

1.2.3. Polyadenylation signals and cleavage sites

The 3' ends of most eukaryotic protein-coding transcripts terminate with a poly-A tail (Darnell et al. 1971; Edmonds et al. 1971; Lee et al. 1971) that is important for nuclear export, stability, and efficient translation (Bousquet-Antonelli et al. 2000; Proudfoot 2001). The tail is added via a multi-protein complex that recognizes sequence elements in the 3' UTR, cleaves the nascent transcript, and adds adenylate residues in a template-independent reaction. The biochemical details of the process have been studied most intensively in mammals and yeast (Guo et al. 1996; Colgan et al. 1997; Zhao et al. 1999).

The local sequence features thought to recruit the polyadenylation and cleavage apparatus show some conservation across phyla. In mammals, the two sequence features that are most important are a highly conserved AAUAAA motif located 10-30 nucleotides upstream of the cleavage site and a GU-rich element located 20-40 nucleotides downstream of the cleavage site. Together, these two elements specify the location of the cleavage site. The Cleavage and Polyadenylation Specificity Factor (CPSF) has been shown to bind to the AAUAAA motif and Cleavage Stimulation Factor (CstF) to the GU-rich element. There is evidence in *C. elegans* that the binding of CstF to the element is not necessary for at least some genes, (Huang et al. 2001), though RNAi analysis has shown that knockout of CstF itself is lethal (Simmer et al. 2003).

In *Saccharomyces cerevisiae,* the 3' UTR features are slightly different. The AAUAAA motif is not as highly conserved and there is no downstream GU-rich element. Instead, there is a UA-rich sequence upstream of the AAUAAA motif. The protein that binds the AAUAAA motif is Rna15, which is orthologous to CPSF; the

UA-rich sequence is bound by Hrp1 (Kessler et al. 1997; Chen et al. 1998; Gross et al. 2001). The cleavage site is 10-30 nucleotides downstream of the AAUAAA motif and has the sequence $Y(A)_n$. In addition to these features, U-rich sequences immediately flanking the cleavage site also appear to be important (Dichtl et al. 2001).

The formation of the 3' end processing complex is linked to transcription by RNA polymerase II; it has been shown that the RNA polII C-terminal Domain (CTD) is essential in mRNA polyadenylation (Hirose et al. 1998). Additionally, it is thought to bind to CstF at transcription initiation. As CPSF is known to interact strongly with transcription factor TFIID (Dantonel et al. 1997), it appears that both these essential 3' end complex proteins are involved in mRNA processing right from the initiation of transcription.

Other proteins involved include two cleavage factors, a poly-A polymerase, and a polyA-binding protein which stabilises the polyadenylated mRNA (Zhao et al. 1999). Figure 2 shows an overview of the 3' end processing complex.

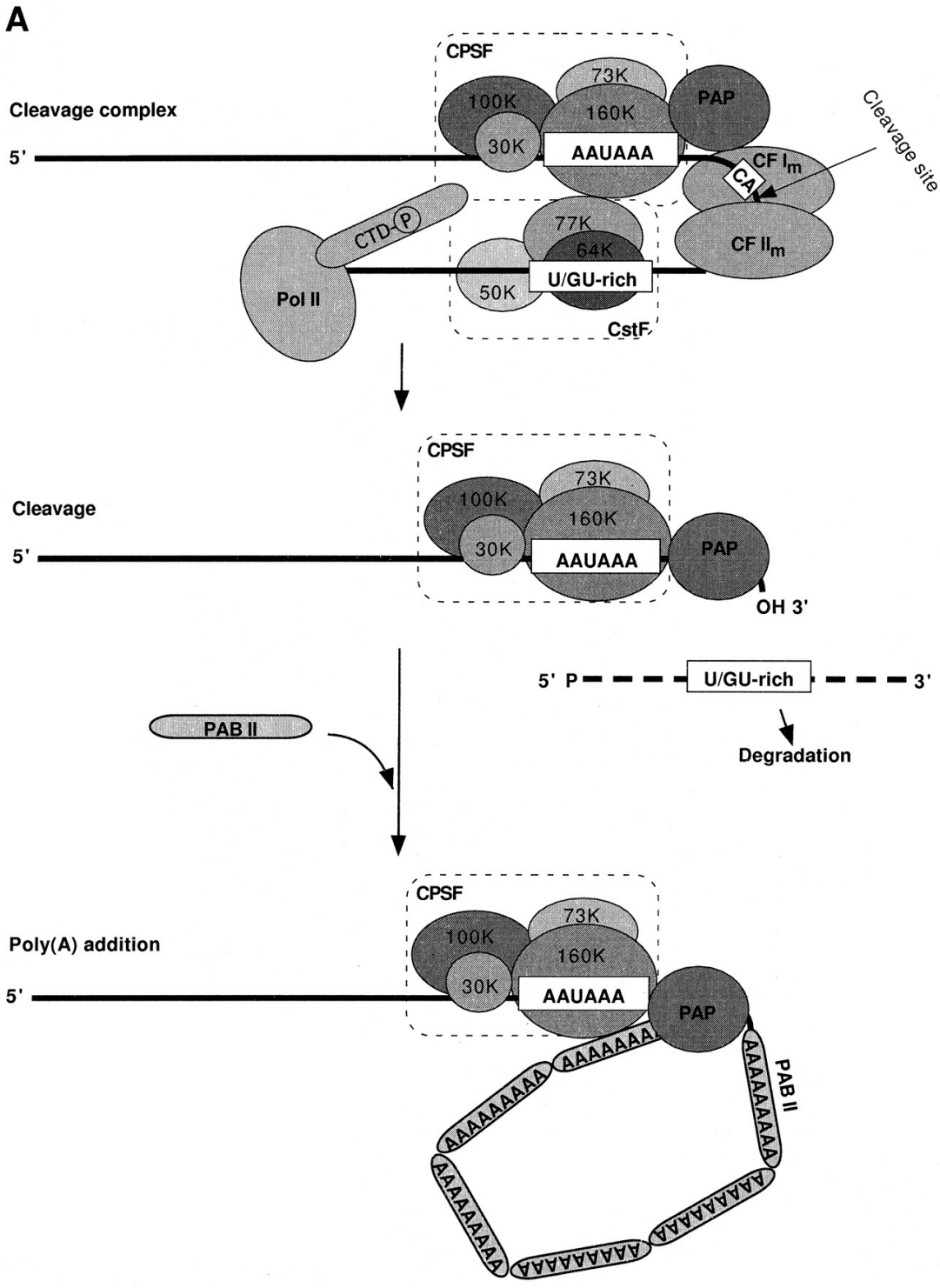

Figure 2. An overview of some of the proteins involved in mammalian 3' end processing. We can see the four subunits of the Cleavage and Polyadenylation Specificity Factor (CPSF), Poly-A Polymerase (PAP), Cleavage Factors I and II (CFI, II), Cleavage Stimulation Factor (CstF), RNA Polymerase II (RNAPol II) with its C-Terminal Domain (CTD). Image taken from (Zhao, Hyman, et al 1999)

1.2.4. Polyadenylation and splicing

According to the currently understood model of exon definition (Berget 1995), each exon is defined by the upstream acceptor (3') splice site and the donor (5') splice site at its end. Initial and terminal exons are missing functional initial acceptor and final donor splice sites respectively, and it is thought that the function of these splice sites is accounted for by the 5' methyl-guanine cap (Ohno et al. 1987) and some component of the polyadenylation complex (Niwa et al. 1991) respectively.

It has now been established that polyadenylation is closely linked to the splicing of the final intron (Cooke et al. 1996). The U1 spliceosomal ribonuclear protein (RNP), which is involved in early recognition of donor splice sites, has been shown to interact with Cleavage Factor I (Awasthi et al. 2003). Additionally, another part of the U1 complex, U1A protein, is known to bind to CPSF and stabilises its binding to polyadenylation signals (Lutz et al. 1996). Another factor involved is the U2AF protein, which binds to poly-A polymerase (Vagner et al. 2000). This protein helps specify acceptor splice sites, and may suggest that more components of the spliceosome are recruited to the cleavage and polyadenylation apparatus. An interesting connection between splicing and polyadenylation pathways is the involvement of Poly-pyrimidine Tract Binding protein (PTB). This has a known function in competing with U2AF for the poly-pyrimidine tract found at the 3' end of introns, and is thus thought to be one of the factors responsible for alternative splicing (Lin et al. 1995). It appears that PTB also competes with the CstF binding site, which can be GU- or pyrimidine-rich (Castelo-Branco et al. 2004). Although this competition causes repression of polyadenylation when PTB is overexpressed, depletion of PTB by RNAi abrogates 3'end processing at certain types of

polyadenylation signal, such as that of the human Complement C2 gene, as does mutation of the PTB binding site (Moreira et al. 1998).

1.2.5. Alternative polyadenylation

Some genes contain multiple polyadenylation signals (Edwalds-Gilbert et al. 1997). This can lead to formation of multiple transcripts, some having extra 3' UTR sequence, such as described by (Qu et al. 2002). This difference is enough to increase translational efficiency of one variant. Alternatively, polyadenylation signals can appear in introns, meaning that different transcripts contain different coding exons in a manner similar to alternative splicing (Alt et al. 1980). An example of the latter includes the mouse immunoglobulin M heavy chain gene, where the switching of polyadenylation signals from one in the 'terminal' 3' UTR to one in an intron causes the deletion of a C-terminal hydrophobic region responsible for membrane anchoring. This changes the protein product from being a membrane-bound protein to a secreted one. More cases are reviewed in (Edwalds-Gilbert et al. 1997).

1.2.6. Polyadenylation signal detection

1.2.6.1. The need for signal prediction

One reason for computational prediction of 3' UTR sequence was given earlier; to restrict searches for mRNA regulatory motifs. However this information is also useful for integrating into other sequence analyses. Knowledge of the extent of the 3' UTR can aid in gene prediction and genome annotation. As the majority of protein coding genes have a polyadenylation signal, each good prediction represents a

piece of high confidence evidence for a gene. The existence or lack of a predicted signal could be the difference that convinces an annotator as to the veracity or otherwise of a gene prediction. Although it is outside the scope of this thesis to write a full gene-finding program, the results of predictions could be integrated into a genefinder that uses many sources of evidence e.g., (Howe et al. 2002), which could use the extra information to improve gene prediction relative to a program that does not model 3' UTRs.

In *C. elegans,* polyadenylation signal prediction will make up for the ~50% coverage missed by EST projects. Now there are genome projects without deep EST projects, such as 5 new nematodes and 10 new flies during 2005. Assuming that the characteristics of polyadenylation signals are conserved between closely related species, we can improve gene prediction in newly sequenced genomes by extending terminal exon predictions to include 3' UTRs. This computational method means that 3' UTR sets can be made without the need for EST projects. The coordinated analysis of the 3' UTRs of orthologous genes, in particular the statistical reinforcement provided by having multiple functional alignments will hopefully improve detection of diffuse conserved regulatory sequences in 3' UTRs.

Computational polyadenylation signal prediction has been carried out to some success in *S. cerevisiae, H. sapiens,* and *M. musculus* (see below). No such work, beyond the suggestion of a naïve model, has been carried out previously in *C. elegans.* In addition to providing improved datasets to the scientific community (http://www.sanger.ac.uk/Projects/C_elegans/POLYA), polyadenylation signal prediction, be it tuned for a given species or no, presents an interesting computational and biological problem.

1.2.6.2. Existing computational methods

Computational polyadenylation signal prediction has been previously attempted by several groups, though this work has mainly been carried out in *H. sapiens*. An early approach was to use a linear discriminant function (Salamov et al. 1997). This method looks for matches to a polyadenylation signal and downstream element consensus, surrounded by characteristic hexamer and triplet frequencies. There is a preferred distance between the signal and the element. The linear discriminant function weighs each of these coefficients according to maximising discriminatory power on a training set. The most important elements were thought to be the polyadenylation signal itself and the hexamer frequencies in the downstream region.

Another group (Tabaska et al. 1999) used a more complex quadratic discriminant function to learn weight matrices for the AAUAAA motif and the GU rich element. The downstream GU rich element and its distance from the polyadenylation signal were once again found to be discriminating, alongside the separation between the two, and the dinucleotide frequencies of the downstream region.

A third study assembled weight matrices from alignments of a large number of sequences containing AAUAAA motifs discovered from EST data (Legendre et al. 2003). This group adjusted the width of putative weight matrices up and downstream of the AAUAAA motif to optimise prediction accuracy, though maximum discrimination was found using just the AAUAAA motif and the local downstream region. This was a similar observation to that gained in the first two studies; in human and mouse, there appears to be little discriminatory information upstream of the polyadenylation signal.

As an alternative to using weight matrices, an investigation into 3' end processing in *S. cerevisiae* (Graber et al. 2002) used a hidden Markov model (HMM) to describe nucleotide frequencies in well-characterised words in the vicinity of the cleavage site, linked by background frequencies elsewhere. This resulted in a model of three informative hexamer words, a pentameric cleavage site and a downstream hexamer word. These words were linked by states having some background nucleotide frequency distribution and a preferred length.

Less predictive work has been carried out in *C. elegans*. The current model for sequence features involved in 3' end formation in *C. elegans* is focussed entirely on the AAUAAA motif (Blumenthal et al. 1997). From a predictive standpoint, this means that one typically scans a weight matrix across the sequence and annotates those sites scoring over a particular threshold. This is not a reliable method of prediction, as the hexamer does not carry enough information to define a polyadenylation signal specifically, compared to the background frequency of AATAAA and similar motifs in the genome. This simple weight matrix model cannot interpret context information, should there be any present.

We now proceed to develop software capable of detecting polyadenylation signals. We can use this to predict signals in *C. elegans* (chapter 3), *C. briggsae* (chapters 5 and 6), *D. melanogaster, H. sapiens,* and *M. musculus* (all chapter 4). In addition to providing a solution to the polyadenylation signal problem, these predictions enable us to study 3' UTR sequence evolution (chapter 5) and help us find a putative 3' UTR motif (chapter 6).

2. PAjHMMA – Parameter Adjustable Java Hidden Markov Model Architecture

2.1. Introduction

In this chapter, we present a flexible software framework, PAjHMMA, for detecting signals in nucleotide sequences. The resulting model is based on the observation that regions of different biological function can have constrained sub-sequences or nucleotide distributions. By varying the parameters in the model, it is possible to model many different encoded biological signals. The technique used allows us to model both long-range, diffuse sequence information, as well as exact or stringent matches to well-characterised sequence motifs. The flexibility of the framework is provided by the separation of the model from the algorithms used to search for model hits. For this reason, it is possible to search sequences for different biological motifs quickly. As the model file is provided in a very simple syntax, a model can be changed or developed afresh with no change required to the decoding software. PAjHMMMA is available for download from http://www.sanger.ac.uk/Software/analysis/pajhmma.

2.2. An overview of hidden Markov models (HMMs)

2.2.1. Background

As discussed in chapter 1, biologically meaningful signals are encoded within biological sequences as discrete regions of given function, which can range from being a local exact sequence match (such as an in-frame STOP codon), to much more

diffuse, poorly-defined motifs, such as eukaryotic promoters (Down et al. 2002). Hidden Markov models have been used successfully to detect signals of varying strengths and combine them together, such as in assigning proteins to various families based on the position and order of protein domains (Bateman et al. 2004), and in gene finding (Burge et al. 1997; Zhang 2002).

2.2.2. Hidden Markov models

An HMM is a statistical model, which has been used in diverse fields in which information occurs in sequences of discrete emissions. Examples of such uses include speech recognition (Rabiner 1989), music recognition (Raphael 1999), gene finding (Burge et al. 1997), and in other biological sequence analysis (Durbin et al. 1998). The model has a finite number of states, each of which has a distinctive frequency distribution over the 'alphabet' of possible emissions. The states are connected to one another by a set of probabilities. A state can be analogous to some kind of functional or characterised feature, such as a season of the year, which has a distinctive emission frequency for particular types of weather.

We can think of an HMM as a machine generating a sequence left to right according to an underlying state path that is hidden from us. The machine has a finite number of interconnected states, and a fixed alphabet. Hence the machine is constrained to emit symbols existing only in that alphabet, and at the frequencies prescribed by the current state. At each stage, the model emits a nucleotide according to the nucleotide emission characteristics of the model's current state. It also moves into a new state (possibly the same one) according to a state-state transition probability.

The HMM in Figure 3 shows a model that can generate a sequence of **a** and **b** emissions. Each of the two states has a characteristic emission. The transition probabilities between the states dictate the order and lengths of the states, and also the overall topology of a pass through the model. An example of a model-directed topology is shown; a pass through the model must end in state B. An HMM is a natural method to model DNA sequence, as regions of different functions typically have different nucleotide frequencies, and can have functionally constrained positions relative to each other. The simple model in Figure 3 models two different states with different emission characteristics, typically encourages state A to emit fewer emissions than state B, and constrains the model such that a state sequence can swap between A and B as many times as it likes, but can only terminate when in the B state.

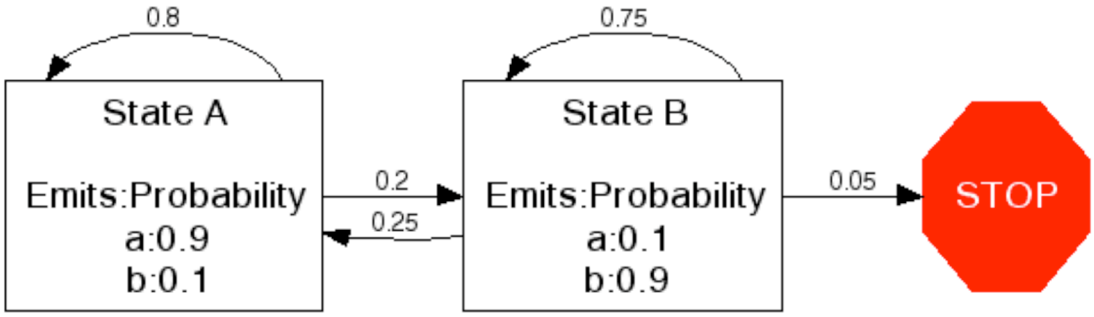

Figure 3. A diagram of an HMM that generates a sequence of a and b emissions. Note that states A and B have different emission frequencies for possible emissions a and b. State A has a tendency to emit a, and B a tendency to emit b. States A and B may transition into themselves or into each other with the given probabilities. State B can transition into an end state. Emission probabilities dictate the constitution of each state.

2.2.3. HMMs for prediction

Rather than using the model as a sequence generator, it is possible to score a given sequence according to all possible state paths through the model. We scan the sequence, and place each emission in one of the states. We score an observed **a** with

0.99 if we decide to put it in state A and 0.01 if it is to be put in state B. The opposite scoring scheme is used if we encounter a **b**. At any point, the model must either transition to a different state, or elect to stay in the same one. This transition has a characteristic cost, depending on which of the possible transitions occurs. We note the state into which each emission can be placed, trying to maximise the score at all times. The state path that scores the highest, represents a demarcation of the sequence into partitions of different emission frequencies. The highest scoring state path, for a sequence

aaaaaabbbbbbaaaaaabbbbbb

would be

AAAAAABBBBBBAAAAAABBBBBB

This state path would also be the highest scoring for a sequence

aabaaabbbabbaaabaabbabbb

but not for

aabbaabbaabbaabbaabbaabb

for which the highest scoring state path would be

AABBAABBAABBAABBAABBAABB

The difference in the two highest scoring state paths is that in the final example, it is deemed preferable to transition into a different state rather than accommodate a suboptimal emission, which is tolerated in the previous sequence.

Given a sequence, therefore, the inference of this state path represents a predicted annotation for the sequence. If this is DNA sequence, and the states score

nucleotides with emission probabilities characteristic of exons, splice sites, and introns etc., then the highest scoring state path represents a gene prediction.

Each sequence emission in any given state is given a transition and emission score which is equal to the probability of that transition and emission. The score of a sequence path through an HMM represents a joint probability distribution over state paths and sequences. Given a sequence, it is possible to find the most likely state path using the Viterbi algorithm (Viterbi 1967). Rather than evaluate every sequence element - nucleotides in the case of a DNA sequence - in every possible state, which would be a brute-force approach, we can use dynamic programming to search the state-space more efficiently. The Viterbi algorithm accomplishes this by a left–to–right sweep to establish partial scores of matching a sequence prefix, given it ends in some state. At any time during the forward sweep, only the best score so far ending in each state is stored. This is followed by a trace back from right to left, extracting the most likely state path. This means that sub-optimal solutions are not evaluated, and thus the total search time actually used is much smaller than that would be used by a brute force method.

Additional information available from using an HMM comes from the ability to calculate the probability that nucleotide n was generated by a state k, using the forward and backward algorithms, which will be covered later. Later in this chapter, I give the equations used to implement these algorithms.

2.2.4. HMMs for gene finding

We can partition DNA sequence into exons, introns, intergenic sequence etc. These have different nucleotide properties on account of their different biological

functions. As a result, HMMs have been used successfully as gene finders. In a given genome sequence, the state path of biologically functional regions is hidden in emissions made up of nucleotides, dinucleotides, or higher order emissions, such as coding triplets. Using sets of manually annotated genes, it is possible to build states representing each biological entity, such as an exon, by finding its characteristic emissions. These states are then connected to each other in a biologically meaningful manner. By this, we mean that the design of the model must obey the rules of biology as we understand them. Hence, a reasonable gene prediction must not contain in-frame stop codons, nor can a model pass in and out of an intron without flanking it with donor and acceptor splice sites.

HMMs are ideal constructs for modelling 3' UTRs and polyadenylation signals, because the region to be modelled can be partitioned into functional areas of distinctive nucleotide properties, as we shall observe in chapter 3.

Polyadenylation signal prediction can therefore be viewed as a logical extension to gene prediction, as virtually all protein coding genes have a 3' UTR and polyadenylation signal.

2.2.5. Length issues

In a simple HMM, when one state accounts for several bases, it does so by having some transition back to itself, and some into the next state. The more sequence is emitted from this state, the more times the transition back to itself must have been chosen. If the probability of transition to itself is p, then the transition of leaving is $(1-p)$. Disregarding emission probabilities, the probability of remaining in a state for m nucleotides

$$P(m\ residues) = (1 - p)p^{m-1}.$$

This leads to the length of sequences emitted by the state being distributed according to the geometric distribution; they decay exponentially. Although certain biological sequence lengths, such as 3' UTRs, can be approximated using this distribution, HMMs are weaker at modelling features with distinctive non-geometric length distributions. There are various strategies allowing a combination of geometric-length states to model a non-geometric phenomenon (Durbin et al. 1998), but a PAjHMMA model allows the user to specify an explicit length distribution for a given state when required. We implement this using a *generalised* HMM.

2.2.5.1. Generalised HMM

In a generalised HMM a state can emit a region of sequences in one step, rather than one nucleotide at a time. This means that we can specify the length distributions for sub-sequences emitted by given states, so we can model states with non-geometric length distribution shapes more accurately. For example, this technique can be used to impose a minimum length on introns, which is appropriate, as these have a biologically constrained minimum length.

The algorithm is extended so that when it transitions into a state with an explicit length distribution, the whole sequence region is emitted and scored, according to the provided length distribution. This means that the number of emissions coming from a particular state is influenced by an observed length distribution rather than a transition probability.

2.2.6. Model training

To build an accurate HMM, it is necessary to determine emission and transition probabilities that closely reflect the biological feature being modelled. As we shall discover in chapter 3, polyadenylation signals can be found experimentally for a small number of genes. For some species, this number is big enough to find emission probabilities by counting nucleotide frequencies. Transition probabilities can also be found, by counting occurrences of an annotated transition event. Manual training is also possible, by calculating a transition probability to approximate an observed length distribution.

Some state sequences correspond to a pass through a weight matrix. There is one state per weight matrix column. This means that each of these states emits exactly one nucleotide, so the transition probability from one state to another is always set to 1.

2.3. Software design

To annotate a sequence against an HMM, the user provides a sequence in FASTA format and a generalised HMM with topology and parameters in a file described below. We describe the format of the model and the pre-processing of the sequence, before describing the dynamic programming algorithm.

2.3.1. Objects

The code is written in the Java programming language. This was partly due to the availability of the BioJava project (http://www.biojava.org/), which provides easy access to software libraries for computational biology. Of interest to this project were classes used for handling FASTA files and DNA sequence utilities.

Having access to objects allows an intuitive setting and access of parameters for the model. We declare a GeneralisedHMM object, which gives us access to a number of States, connected by transition probabilities. Each State is also an object, having methods to return its characteristic nucleotide frequencies, and length distribution, if it has one.

After a few initialisation steps, dynamic programming is carried out by first principles. Sequences are parsed using BioJava utilities and converted into streams of integers. Transition probabilities and state emission frequencies are stored as elements in two dimensional arrays. Modelling the DNA sequence as a sequence of integers means that for each emission, emission and transition scores can easily be looked up by using array indices. Calculated scores are stored in and looked up from a dynamic programming matrix, which is also a 2D array.

2.3.2. Model

The model to be used for sequence decoding is provided as a simple tab-delimited text file. It consists of an HMM declaration, followed by a list of states. Each state consists of a state declaration, followed by the transition, emission, and (optionally) length parameters. Once the file is parsed, a GeneralisedHMM object and a collection of State objects is constructed. This allows the formation of lightweight 2D double arrays containing emission and transition frequencies. As we will be converting the DNA sequence into numbers (ints), this means that all lookups in the

dynamic programming matrix fill stage will be array lookups, rather than hash lookups, which are slower.

2.3.2.1. The HMM declaration

```
HMM C.elegans-3'end 14
```

This declaration informs the constructor of the GeneralisedHMM object how many states there are. Although this information is redundant, it is useful to instruct the program how much memory to allocate. Java memory allocation is automatic, but pre-specifying the number of states allows the use of arrays. These are used in preference to ArrayLists, as the latter's gain in flexibility comes at a price of poorer performance.

2.3.2.2. The State declaration

```
State UTR 1    0   2   0   0
Transition UTR  0.99
Transition A1   0.01
Emissions  0.270   0.198   0.127   0.405
EndState
.
.
.
State SP 0 0   1   0   30
Transition C1  1.0
Emissions  0.271   0.138   0.137   0.454
Length /Users/ashwin/models/length_distribution.txt
EndState
```

The state declaration contains the name of the state, and five numbers. The first two are flags with 0/1 values for being the initial and terminal states. The third

value gives the number of states to which this state has legitimate transitions. The fourth figure gives the order of the state, though mixed order models are not currently implemented.

The ability to model non-zero order (dinucleotides, trinucleotides, etc) emissions is important, as certain features are either not coded in mononucleotides (Gardiner-Garden et al. 1987), or better modelled using a higher order alphabet (Salzberg et al. 1999). Building higher order models requires more data than zero order models, on account of the need to avoid overfitting.

The last number in the state declaration is zero for geometric states, but when a length distribution is explicitly specified, then this value is equal to the length of the length distribution. As with the HMM declaration, this up-front declaration allows more efficient parsing of the model file into a Java object.

2.3.2.3. The state specification

For each state, there is a list of legal transitions and their probabilities. All other transitions are set to zero. There then follows a list of emission frequencies for each nucleotide, given in alphabetical order. If the *StateOrder* is the order of the emissions from a state, the number of emissions expected is $4^{StateOrder+1}$, so with a first order model, 16 dinucleotide frequencies should be given.

In the example above, the SP state has a length distribution specified explicitly. The length declaration is the path to a file containing tab-delimited text in the form of a length and frequency or count. The distribution counts are normalised to add up to one.

2.3.2.4. Model attributes

Once the model file has been parsed, two 2D matrices are created. One is the transition matrix, and the other is the emission matrix. The transition matrix has dimensions to contain transition probabilities from each state to every other state in the model, including itself. Transitions disallowed by the model topology are set to zero. All values are stored as log-probabilities for arithmetic reasons (discussed further below).

The emission matrix contains the emission probabilities for each nucleotide (dinucleotide etc.) within each state. These are also stored as log probabilities.

2.3.3. Sequence pre-processing

The DNA sequence to be annotated is converted to an array of integers depending on the order of the model. For a zero order model, this array contains values 0-3, corresponding to A, C, G, and T. There are 16 values if the model is first order. This pre-processing allows us to avoid the use of hash lookups in the dynamic programming loops, in favour of array lookups, which are faster. The conversion of the sequence into a numerical form means that if we refer to a state by a number, given a model and a sequence, we can call a particular emission probability from the emission matrix by giving the state number and the nucleotide number as a lookup from the 2-dimensional array.

2.3.4. Dynamic programming algorithms

The three algorithms commonly used in annotating a sequence with states from an HMM are well documented (Durbin et al. 1998). As the HMM used here is generalised – states are allowed to have an explicitly specified length – modifications are needed to the Viterbi, forward, and backward algorithms when finding maximal or sum evaluations in a state with an explicit length.

Given a sequence of length L having emission x at position i., we require a dynamic programming matrix to store the evaluation of sequence emissions in each state. For each position in the sequence, this matrix, $v_k(i)$, stores the probability of the highest scoring path ending with the i-th nucleotide in state k.

2.3.4.1. Viterbi algorithm for geometric states

The standard Viterbi algorithm used for states with geometric length distributions is reproduced with slight alteration below from (Durbin et al. 1998).

$$\text{Initialisation } (i = 0): \quad v_0(0) = 1, v_k(0) = 0 \ \textit{for } k > 0.$$

Then for all nucleotides in a given state l, where the previous nucleotide was in state k, a score v_l is calculated. There are two components to this score. The first is the emission score $e_l(x_i)$ of nucleotide x at position i in state l. This is constant regardless of the value of the previous state k. The second component is the maximum value found by evaluating, for all values of k, the product of the previous maximal score at the previous nucleotide ($v_k(i-1)$) and the transition probability a_{kl} from state k to l. Only storing the maximal of all previous values means that at each extension by one nucleotide, the extension is being carried out only on the optimal prefix, rather than on all prefixes. It is for this reason that the Viterbi algorithm finds the optimal

path, termed π*, much faster than a brute-force evaluation. The two components are multiplied together to give the score v_l at nucleotide position *i*. To keep track of which state transitions were occurred at which positions in the optimum path π*, the value at each *l*, of the optimal previous state, $\arg\max_k(v_k(i-1)a_{kl})$, is stored in an array of pointers.

$$\text{Recursion } (i = 1...L): \quad v_l(i) = e_l(x_i)\max_k(v_k(i-1)a_{kl});$$
$$pointer_i(l) = \arg\max_k(v_k(i-1)a_{kl}).$$

$$\text{Termination}: \quad P(x,\pi^*) = \max_k(v_k(L)a_{k0});$$
$$\pi_L^* = \arg\max_k(v_k(L)a_{k0}).$$

Then by tracking backwards through the sequence, we know at each nucleotide, which state the preceding nucleotide was in for an optimal scoring path, so following this pointer through the whole sequence will give the state annotation that scores highest, given each state's characteristic nucleotide emission and the transition probabilities between the states.

$$\text{Traceback } (i = L...1): \pi_{i-1}^* = pointer_i(\pi_i^*).$$

For a zero order model, an emission equates to one nucleotide, but for higher order models, the sequence must be tokenised into higher order emissions, such as hexamers, and the emission scores trained on appropriate measurements.

2.3.4.2. Posterior decoding

The Viterbi algorithm detailed above gives the most probable state path through the sequence. However, for modelling a biological system, it is often more appropriate to find sub-optimal matches to a model. An example of this is in splice site analysis, in which it has been shown that an exact match to a consensus causes reduction of splicing activity on account of the U1 snRNA binding too strongly to its binding site (Lund et al. 2002). The most probable path is also not particularly informative if there are many high probability paths with very similar probabilities. In these circumstances, it may be better to capture every occurrence of a sequence being in a particular state above a threshold probability, so it is informative to know the probability of being in a particular state at a given nucleotide. This is the posterior probability, and requires us to calculate the forward and backward probabilities $f_k(i)$ and $b_k(i)$, which are explained below.

The posterior probability of a particular nucleotide at position *i* in a particular state *k* is calculated as

$$P(\pi_i = k \mid x) = \frac{f_k(i)b_k(i)}{P(x)},$$

which is the probability of the observed sequence *x* over all paths up to nucleotide *i* in state *k* (the forward probability) multiplied by the probability of the observed sequence in all paths following nucleotide *i* being in state *k* (backward probability), divided by the probability of the sequence. To work out these values, we use the forward and backward algorithms.

2.3.4.3. The forward algorithm

The forward algorithm calculates the probability of all possible paths through the sequence instead of the most likely. It is similar to the Viterbi algorithm, but for a particular position in a particular state, it evaluates the sum of all the possible paths leading up to the one in question, rather than finding the maximum. The initialisation is the same, but the recursion and termination stages of the forward algorithm are therefore summations, not evaluations of the maximum. The probabilities, up to and including point *i,* of all paths putting nucleotide *i* in state *k* are stored in a forward matrix ($f_k(i)$).

$$\text{Recursion } (i = 1...L): \quad f_l(i) = e_l(x_i)\sum_k f_k(i-1)a_{kl};$$

$$\text{Termination}: \quad P(x) = \sum_k f_k(L)a_{k0};$$

2.3.4.4. The backward algorithm

The backward algorithm $b_k(i)$ calculates the probability of the sequence following *i*, given that nucleotide *i* was put in state *k*. This is like the forward algorithm, but has a backward recursion though the sequence, and for a given nucleotide, score sums are evaluated over which state the next nucleotide can be put.

$$\text{Initialisation } (i = L): \quad b_k(L) = a_{k0} \text{ for all } k.$$

$$\text{Recursion } (i = L-1,...,1): \quad b_k(i) = \sum_l a_{kl}e_l(x_{i+1})b_l(i+1);$$

$$\text{Termination}: \quad P(x) = \sum_l a_{0l} e_l(x_1) b_l(1).$$

2.3.5. Modifications required for decoding explicit length states

The length duration of an HMM state is usually implied by the transition probability for leaving that state. For a state with out-transition probability P, the probability of remaining in that state after N transitions is $(1\text{-}P)^N$. The length of a sequence that is modelled this way is thus geometrically distributed with mean 1/(1-P). Geometric length distributions are reasonable approximations for certain biological sequences, such as eukaryotic intergenic regions (Burge et al. 1997), but not for others, such as the region between the *C. elegans* polyadenylation signal and the cleavage site, as we shall show in chapter 3.

States with an explicit length distribution add a complication to the three dynamic programming algorithms discussed above. One way to deal with them is to have different out-transition probabilities dependent on how many emissions have been made from that state. However, no length information is stored by the Viterbi algorithm, and thus if a preceding state has a non-geometric length distribution, it is necessary to evaluate all possible lengths from that state separately.

For a given state transition being evaluated in PAjHMMA, the dynamic programming loop has a switch that asks whether the next emission/set of emissions should be scored in an explicit duration state or a geometric one. This information is stored in each state of the model.

2.3.5.1. Viterbi algorithm for explicit length states

According to the Viterbi algorithm, for each nucleotide, it is necessary to find the state that scores it and its prefixed state path maximally. Now in addition, each time an evaluation is made in a non-geometric state, it is necessary to evaluate (and maximise over) not just a single emission in that state, but a compound score calculated over all possible sequences of emissions with that starting point. There are two components to the extra information required in explicit length states. For a state length distribution *d* with *D* elements, the first is an emission score for the whole sequence being evaluated within the state, and a length score based on the frequency of that length in the distribution (Figure 4).

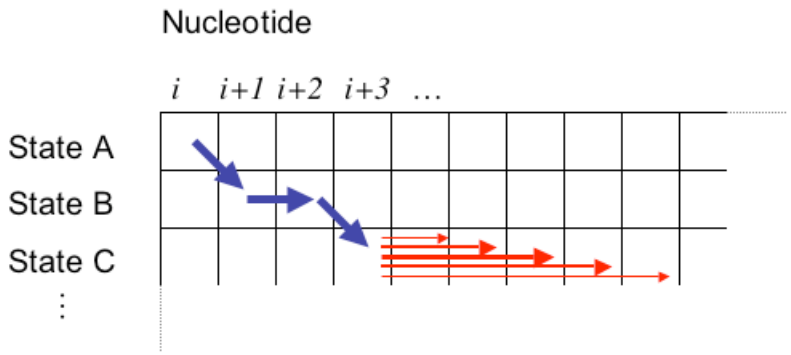

Figure 4. A diagram of a dynamic programming matrix showing one of many possible paths. States A and B are normal states having geometric length distributions. The maximal scoring path at any stage is maximised over which state the next nucleotide should be put in. State C has an explicit length distribution, shown with red arrows. Certain lengths in this distribution are favoured. Scores are found for all possible lengths of sequence in this state. The maximal scoring path is maximised over the combination of (a) the emission score of all these lengths and (b) a scaling factor according to frequency of each length in the length distribution.

All possible lengths within the length distribution are evaluated. At any particular sampling length *m,* the score of the maximal prefix to this sequence is found by looking back *m* nucleotides. The length score is *d(m).* If *l* is a state with a specified length distribution,

$$v_l(i) = e_l(x_i)\max_{\substack{k;\\ m\in\{1,\ldots,D\}}} v_k(i-m)a_{kl}d(m)\prod_{q=1}^{m} e_l(x_{i-q}).$$

Each time the algorithm attempts a transition into an explicit length state, the pointer containing the optimal source state is kept as before, but it also stores the length of sequence causing the maximum score. This value is required in the traceback procedure.

2.3.5.2. Posterior decoding with explicit length states

As with geometric length states, posterior decoding requires the forward and backward probabilities. The forward algorithm for explicit length state *l* is once again similar to the Viterbi; it is simply the sum of all the terms from which the maximum was stored previously.

$$\text{Forward}: \quad f_l(i) = e_l(x_i)\sum_{\substack{k;\\ m=1}}^{m=D} f_k(i-m)a_{kl}d(m)\prod_{q=1}^{m} e_l(x_{i-q})$$

The backward algorithm differs from the non-explicit length version simply by having extra terms for the emissions from the *m* residues being scored in state *l*, which

is $e_l(x_{i+m})$, the sum of backward scores up to that point $b_l(i+m)$, and the length score $d(m)$.

$$\text{Backward}: \quad b_k(i) = \sum_{\substack{l; \\ m=1}}^{m=D} b_l(i+m)a_{kl}d(m)\prod_{q=1}^{m} e_l(x_{i+q});$$

2.3.6. Preventing numerical underflow

All the algorithms given above feature the multiplication of probabilities. In particular, in the Viterbi and forward algorithms, the maximum score at nucleotide *i* is a product of all the previous maximum scores. The minimum value of a Java `double` is 2^{-1074}. A straightforward implementation of even a simple HMM would therefore underflow (depending on parameters) within a few thousand nucleotides at the most. A joint sequence probability with a mean probability per nucleotide of 0.5 would underflow after 1074 residues. To prevent such problems, the standard solution is to work in log space. This turns all of the multiplications into sums. In the standard Viterbi algorithm, as the score value is just the product of all the score components, using logs is simple enough.

$$\ln v_l(i) = \ln e_l(x_i) + \max_k(\ln v_k(i-1) + \ln a_{kl}).$$

In the forward and backward algorithm, an added complication is that calculated path scores have to be summed, so these log values need to be re-exponentiated before they are summed, and the logarithm found again. Hence for the forward algorithm.

$$f_l(i) = e_l(x_i)\sum_k f_k(i-1)a_{kl}$$

$$\ln f_l(i) = \ln e_l(x_i) + \ln \sum_k f_k(i-1)a_{kl}$$

$$= \ln e_l(x_i) + \ln(\exp(n_1) + \exp(n_2) + ... + \exp(n_k)),$$

$$where\ n_k = \ln f_k(i-1) + \ln a_{kl}.$$

However, these exponentiations are likely to underflow, preventing an accurate summation, so we instead rearrange, using the following observation:

$$\ln \sum_{x=a}^{b} \exp(x) = \ln\left(\exp(q)\sum_{x=a}^{b}\frac{\exp(x)}{\exp(q)}\right)$$

$$= q + \ln \sum_{x=a}^{b} \exp(x-q)$$

If we choose the value of the scaling factor q to be the smallest of all the exponents (n_1,n_2…n_k), then the smallest exponent becomes 0.

2.3.7. Methods of usage

PAjHMMA has two principal output forms; one is a traceback through the pointers matrix, which annotates each nucleotide to a state according to the most probable path through the model, with the state boundaries and sequence being printed. Alternatively, the dynamic programming algorithm makes the posterior decoding matrix available to the user. It is thus possible to list the probability of the sequence being in any given state along its length. For a given sequence, all paths

having probability greater than some threshold in some state can be output. The use of this can be seen in Chapter 3.

2.4. Conclusion

In this chapter, I have presented a flexible framework for building an HMM for nucleotide sequences. The resulting HMM is based on a series of interconnected states of different possible types. It can be used to annotate a sequence according to its most probable state path through the model, or the posterior probability of each base matching a particular state. This software is used throughout this thesis to predict polyadenylation signals. PAjHMMA allows the user to specify an HMM in a model file containing the number of states, their characteristic emission frequencies, and their transition probabilities. A model file is provided; this can contain any reasonable number of states, each having characteristic nucleotide emission frequencies. One particular motivation for the design of this software was to support states with an explicit length distribution.

The standard decoding algorithms have been modified to allow states to have a user-defined length distribution. Although the software described here was originally designed for the prediction of *C. elegans* polyadenylation signals, it is possible, given a manually built model with nucleotide frequency and length parameters for each state, to annotate any sequence for any feature having a sequence of states with distinctive nucleotide frequencies.

I next proceed to use the software described to predict polyadenylation signals in *C. elegans*. Chapter 3 explains how to build an accurate model of the *C. elegans*

polyadenylation signal. Based on the success of this, I build models for other species, which eventually enables me to carry out analyses on orthologues (chapters 5 and 6).

3. A Probabilistic Model for 3' End Formation in *C. elegans*

3.1. Introduction

In this chapter, we analyse the polyadenylation and cleavage site from a large number of *C. elegans* genes. By aligning cDNAs that diverge from genomic sequence at the poly-A tract, we accurately identified a large set of true cleavage sites.

Analysis of these cleavage sites showed that in addition to the well known AAUAAA motif, characteristic nucleotide biases were also seen in well-defined regions up- and downstream of cleavage sites. Sequences were demarcated according to the mean lengths of these regions, which were identified manually, and a PAjHMMA model created.

This model is successful at identifying polyadenylation signals when given a 3' UTR and downstream genomic DNA (Hajarnavis et al. 2004). The model is also able to identify sites of alternative polyadenylation. In addition, in an attempt to model molecular biology in a more realistic manner, a simple coding model was introduced upstream of the 3' UTR model, and tested against virtual transcripts, consisting of the spliced coding sequence, the 3' UTR, and downstream genomic. This model showed minimal loss of accuracy versus restricting the search to the 3' UTR, and overwhelmingly outperformed the only previously available regime of scanning sequences with an AATAAA weight matrix.

In cases where there are many mRNAs for a gene we can frequently see that the cleavage site, itself downstream of the AAUAAA, is not clearly defined but occurs in one of a distribution of sites in a defined interval downstream of the motif.

For these genes, the posterior probability of a cleavage site prediction at a particular point as derived from our model appears to mirror closely the observed frequency of cleavage at that point.

For the work described in this Chapter, I gratefully acknowledge the help of Dr. Ian Korf, who built the datasets and provided some of the figures. This work was published in Nucleic Acids Research in 2004 (Hajarnavis et al. 2004), and the figures are adapted from that paper.

3.2. Background

3.2.1. Polyadenylation signals

We are interested in understanding 3' end formation in *Caenorhabditis elegans*. Previous studies on cDNAs have found the presence of the AAUAAA motif 7-22 nucleotides upstream of the cleavage site but none of the other common elements (Blumenthal et al. 1997; Huang et al. 2001), such as a GU-rich region. Furthermore, in this set, only approximately 50% of identified polyadenylation signals are AAUAAA; many single base variants are seen, especially AAUGAA. One unusual feature of 3' end formation in *C. elegans* is that the process is associated with trans-splicing when genes are in operons. In these circumstances, 3' end formation of the upstream gene has been shown to be functionally upstream of SL2 trans-splicing of the downstream gene (Evans et al. 2001). As in mammals, CPSF binds the AAUAAA motif, but unlike in mammals (Chen et al. 1998), there is evidence that efficient 3' end formation can take place in the absence of a putative CstF binding site (Huang et al. 2001). CstF is present, but its role is apparently to increase the local

concentration of SL2 at the trans-splice site and not to specify the position of the cleavage site (Evans et al. 2001).

3.3. Model building

3.3.1. Introduction

Computational methods typically attempt to identify the polyadenylation signal itself, rather than the cleavage site. To build a training set of *C. elegans* polyadenylation signals, it would be necessary to use a large number of known signals. However, there are only 152 *C. elegans* mRNA sequences in EMBL/Genbank with a 'polyA_signal' annotation. A problem with these is that there is no information provided as to what evidence supports that annotation. Possibly as a result of one very influential early paper on *H. sapiens* 3' end processing (Proudfoot 1991), an annotator may have looked for the last occurrence, if any, of an exact match to AAUAAA. Alternatively, there may be mutagenesis evidence that this is indeed the real polyadenylation signal. Bearing this in mind, it is impossible to know whether an annotated signal is real. In contrast, given cDNA evidence, the cleavage site is easy to determine computationally. This is the point where the sequence of a polyadenylated mRNA ceases to be a copy of the genomic sequence in the 3' UTR, and turns into a run of adenylate residues. Hence, any model should be built on sequences with a correctly annotated cleavage site. Although the polyadenylation signal will still be a part of the model, this method ensures that each one is upstream of a verified cleavage site.

3.3.2. Experimentally derived cleavage sites

The 3' UTR of a *C. elegans* gene starts at its stop codon. One of our prior analyses of 3' UTRs (as dictated by EST alignments for about 9,000 genes) showed that 97% of 3' UTR sequences are under 1 kb long. Hence it is reasonable to assume for the purposes of model building that the cleavage site will be included if we take the 1,000 nucleotides 3' of the stop codon. Current sequencing technology allows for reads of up to 1000 nt, and WormBase annotators do not annotate a 3' UTR unless the 3' EST read extends into the coding sequence. Thus, real 3' UTRs above 1000 nt would not be represented in the database. However, the shape of the length distribution of 3' UTRs (Figure 9), suggests that there are an insignificant number of these. Those which do appear above this length are likely to be mapping errors.

22,156 candidate 3' UTRs up to 1000 bp long were extracted from WormBase release WS110 (http://ws110.wormbase.org). Sequences were truncated if they overlapped downstream genes on the same strand. 216,943 *C. elegans* transcripts (cDNAs and ESTs) were retrieved from EMBL/GenBank. The transcripts were processed with a Perl script that used the following rules to identify transcripts containing a poly-A tail.

The transcript had to be at least 200 nt long. Any sequence with 6 or more terminal As was kept, and for those sequences without, since the vector may be present at the end of the sequence, sequences with runs of mostly As near the end were also kept. The Perl regular expression used to define the run of As with a potential sequencing error and up to 30 bp of vector was

```
/(A{3,1000}.?A{3,1000})(.{0,30})$/
```

5,306 transcripts passed these tests and the 3'-most 200 nt were searched against the candidate 3' UTRs with BLASTN version 2.0MP-WashU 23-May-2003 (W. Gish unpublished, http://blast.wustl.edu) using parameters `W=30 M=1 N=-3 Q=3 R=3`. These BLAST parameters mean that no alignment is even seeded unless there is an exact match of 30 contiguous nucleotides between the mRNA and the genomic sequence. Point mismatches are penalised greater than usual (match (M) /mismatch (N) values are usually 5/-4). The change in Q (gap opening penalty, default 10) and R (gap extension penalty, default 10) means that insertions and deletions are penalised at the same rate as mismatches. This is three times the match value, meaning that our BLAST parameters are extremely stringent. Parameters such as the large word size mean that mRNAs only align to those parts of the genome where the query and target sequences are virtually identical. Thus we can be very confident that a particular aligned mRNA represents a transcript from a particular gene.

Following this process, 1,810 3' UTRs had matching transcripts. Some of these sequences are duplicates, on account of having different gene isoforms. By insisting that each sequence had a unique spacer sequence, these duplicates were removed, leaving 1,468. This seems like a small number, given the size of the genome and the amount of cDNA coverage. Approximately half of *C. elegans* genes do have some cDNA evidence, normally in the form of Expressed Sequence Tags (ESTs), but most *C. elegans* ESTs in GenBank have no initial poly-T tract corresponding to the poly-A tail because the initial part of the sequencing read was clipped off before submission for reasons of sequence quality. The traces are not publicly available.

Multiple alignments of each unique candidate 3' UTR were created with their matching transcripts using a Perl script that employed Bioperl libraries (Stajich et al.

2002). 1,156 had at least one matching transcript that diverged from the genomic sequence in what appeared to be a poly-A tail.

3.3.3. Variety of cleavage types

Looking at the cleavage site for each of the genes where there was a clear dissociation of the mRNA from the genomic into a run of As showed that there were four classes of cleavage site (Figure 5).

Figure 5. Four classes of cleavage site, as found by the BLAST analysis. The cleavage site is where the mRNA diverges from genomic sequence. AATAAA motifs are boxed in yellow. Green boxes show the range of possible cleavage sites in the cases where the cleavage occurs adjacent to a genomic A. (a) a cleavage between two G residues. (b) a cleavage that could have occured in any of seven positions. (c) the two mRNAs map to different places in one gene- this gene has more than one polyadenylation signal and cleavage site. (d) a gene with many mRNAs. This shows that the cleavage site caused by a given signal is not always precisely positioned.

Figure 5a shows an example where the cleavage site is clearly visible between two G residues. There are many cases, however, where the cleavage occurs just upstream or downstream of a genomic A, or in a run of genomic As. In this case, the alignment will look the same, regardless of the exact point in the run of As that the

polyadenylated mRNA switches from templated to non-templated As (Figure 5b). The precise cleavage site in these circumstances is ambiguous. Figure 5c shows an example of alternative polyadenylation – there are two separate mRNAs mapping to different parts of the sequence. The final example shows a case of a gene with many mRNAs all mapping to approximately the same place, but showing that the cleavage is an imprecise event.

Of our 1156 cleavage sites found in this way, 156 were of type (a). 855 had a cleavage occurring within a run of genomic As as in Figure 5b. The remaining sequences had multiple mRNA hits; 30 distinct (type (c)) and 115 staggered (type (d)) and these were not used in model building. Given the relatively low coverage of the genome by polyadenylated mRNAs, the relatively large occurrence of non-staggered cleavage sites is more likely to be a result of the scarcity of mRNAs relative to genes, rather than there being an overrepresentation of precise cleavage for biological reasons. Only 262 genes had more than one mRNA aligned.

3.3.4. The problem with ambiguous cleavage sites

To build an accurate model, it is important to train on reliable data. In this case, we wanted to identify the exact polyadenylation signal and the precise cleavage site. One hurdle, therefore, was that the majority of the training set contained ambiguous cleavage sites. We therefore decided to look at those 156 sequences where the cleavage site was known with certainty.

3.3.5. Sequences with well-defined cleavage sites

3.3.5.1. Information at the cleavage site

Superimposition of the three nucleotides flanking the cleavage site to make a weight pictogram (http://genes.mit.edu/pictogram.html) showed that there was little information at the cleavage site itself (Figure 6), barring a general T-richness, which is true of the whole *C. elegans* 3' UTR. However, there was a marked suppression of G in the +3 position relative to the cleavage site.

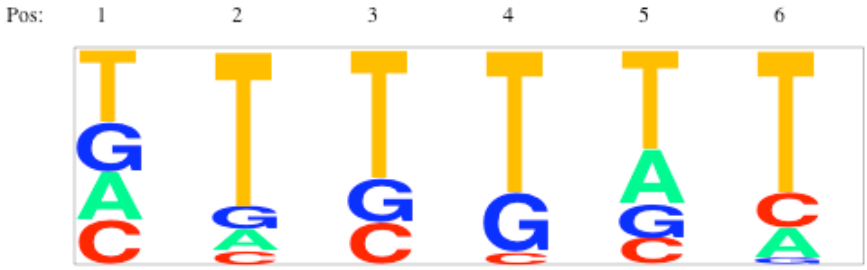

Figure 6. A pictogram of the nucleotide frequencies in the three nucleotides either side of the cleavage site. The cleavage site from 156 sequences occurs between columns 3 and 4. There are no As directly flanking the cleavage site, as it is their absence that defines this class of cleavage. A clear suppression of G is seen in column 6.

3.3.5.2. Length distribution between AATAAA and cleavage site

The 156 sequences with well-defined cleavage sites were isolated, and analysed upstream of the site to look for an exact match to AATAAA. 106 sequences had exactly one non-overlapping exact match within 40 bases upstream of the cleavage site, and no other A-rich hexamer. These AATAAA matches were thus assumed to be real polyadenylation signals. It was observed that the length of

sequence between the polyadenylation signal and the cleavage site had a distinctive distribution (Figure 7).

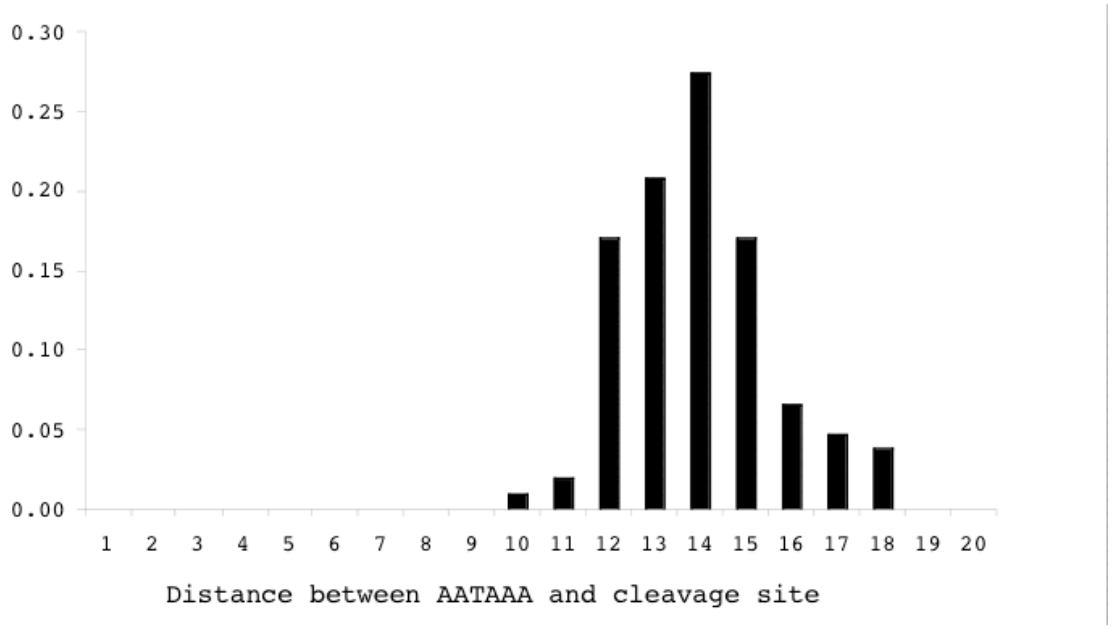

Figure 7. The length distribution of the spacer sequence separating the polyadenylation signal (exact match to AATAAA) and the unambiguously defined cleavage site from 106 sequences.

This suggests that there are preferred separation lengths between the polyadenylation signal and site. Many sources in the literature cite a 10-30nt separation. We see here that the distribution is not flat, but distinctively shaped. The distribution is very tight, ranging from 10 to 18 nucleotides and having mode 14. According to this distribution's Shannon entropy $H(X) = -\sum_i P(x_i)\log P(x_i)$, whereas the flat distribution has 4.39 bits, the observed distribution has 2.63, making it substantially more specific. A normal distribution with mean 13.92 and standard deviation 1.71 has 2.60 bits and is a fairly good fit.

3.3.6. Maximum likelihood determination of cleavage sites

Given this length distribution and a rough idea what a polyadenylation signal should look like, we can use a previously published weight matrix (Blumenthal et al. 1997), to help us annotate cleavage sites that are ambiguous. Given a sequence with a run of As at the 3' end, for every possible cleavage site within the run of As, the weight matrix was evaluated at every length in the length distribution, calculating a weight matrix and length distribution score. The maximum likelihood position of the hexanucleotide and cleavage site was calculated for all 855 ambiguous cleavage sites. To prevent excessive peaking of the observed maximum likelihood scores, the length distribution was smoothed asymmetrically, quartering the frequency at each decrease in length from 10 to 5, and halving it at each increase from 18 to 30. As well as preventing an overly peaked profile, it gives us some prior frequency for outlying lengths, as would have occurred had a much larger set been used, from which to sample the spacer lengths.

Running the maximum likelihood method on the poly-A tail alignments led to the assignment of a unique maximum likelihood polyadenylation signal and cleavage site annotation for 961 sequences. 50 sequences, for which there was no single maximum likelihood (as occurred occasionally when polyadenylation signals overlap), were discarded. A frequency histogram of the 961 observed motifs (Table 1) shows that certain hexanucleotide polyadenylation signals are much more common than others. 21 hexanucleotides, such as AATAAC were observed only once in the entire set, whereas the other 940 sequences had one of 26 different motifs, each appearing at least twice in the whole set. As we can have more confidence in the more frequently occurring motifs, those which appeared only once were regarded as outliers and discarded.

Table 1. Those maximum likelihood polyadenylation signals appearing in the set of 961 more than once.

Hexamer	Counts	Hexamer	Counts
AATAAA	531	AACAAA	6
AATGAA	120	AAGAAA	4
TATAAA	71	TGTAAA	3
GATAAA	43	ACTAAA	3
CATAAA	42	AATAAG	2
TATGAA	22	AATTAA	2
ATTAAA	16	GGTAAA	2
AGTAAA	15	GAAAAA	2
CATGAA	12	TTTAAA	2
AAAAAA	11	ATTGAA	2
GATGAA	8	AAAGAA	2
AATAAT	8	AATATA	2
AATACA	7	TTTGAA	2

531 (56%) were exact AATAAA, 13% were AATGAA, 17% had a single mutation at the first position (TATAAA, CATAAA, GATAAA), 8% had a single mutation elsewhere, and 6% had two mutations. The number of double mutations seems high, and will be discussed in Chapter 5.

3.3.7. Nucleotide frequencies

With a polyadenylation signal and a cleavage site annotated for each sequence, it was possible to anchor all the sequences on their cleavage site and plot nucleotide frequency in the vicinity of this region.

Figure 8 shows the different nucleotide frequencies seen in different parts of the 3' UTR. Note that the body of the 3' UTR has a very distinctive distribution compared to the genome. Globally, because of base pairing, we do not expect one component of a base pair to outnumber its counterpart, so the levels of A and C should be equal to those of T and G respectively. In the 3' UTR, a single stranded

transcribed feature, it is apparent that there is some strand asymmetry with respect to nucleotide frequencies, as there is a clear preference for the pyrimidine of each base pair to be on the sense strand, and the purine on the other. In Figure 8, we can see the different nucleotide distributions; the UTR is T-rich up to about 20 nt upstream of the cleavage site. As well as T being favoured over A, the level of C is greater than that of G. The A-rich region is the AAUAAA motif. Following this, there is a T-rich region of constrained length, leading up to the cleavage site itself, where there is a spike of As, as expected by most cleavages being adjacent to an A. Another T-rich region follows, before the nucleotide distribution returns to genomic levels, some 15-20 nt downstream of the cleavage site.

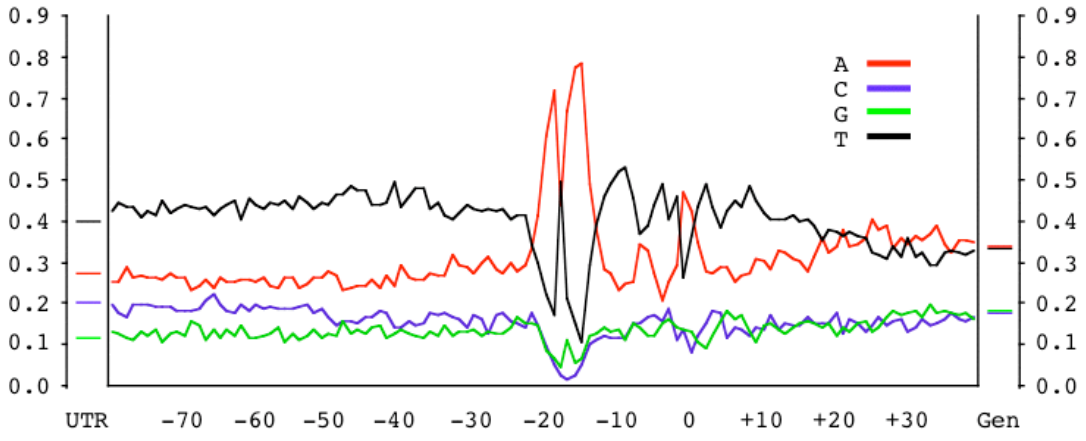

Figure 8. The nucleotide frequencies in 3' UTR, the region -80 to +30 about the cleavage site, and the genomic nucleotide distribution in C. elegans. Each sequence in the training and test sets was annotated into states according to the demarcated zones of distinctive nucleotide frequency.

3.3.8. Building an HMM

Given the information from the anchored alignment, it is possible to build a model, using the PAjHMMA software described in Chapter 2, to represent the characteristic length and nucleotide emission spectrum of each of the distinct regions that can be used to define a 3' end.

All states emit nucleotides at set frequencies, which are characteristic to each state. For each state, these frequencies can be calculated by counting bases of sequences that are split into state sections as described below. The expected length of each state is either set implicitly by its out-transition probability or specified explicitly.

3.3.8.1. 3' UTR state (UTR)

The *C. elegans* 3' UTR has a highly variable length. 97% of sequences are below 1000 nt, and the mean length is 200 nt. For the purposes of the model, let us define the UTR state to run from the STOP codon (inclusive) to the polyadenylation signal (exclusive). The length distribution can be seen in (Figure 9)

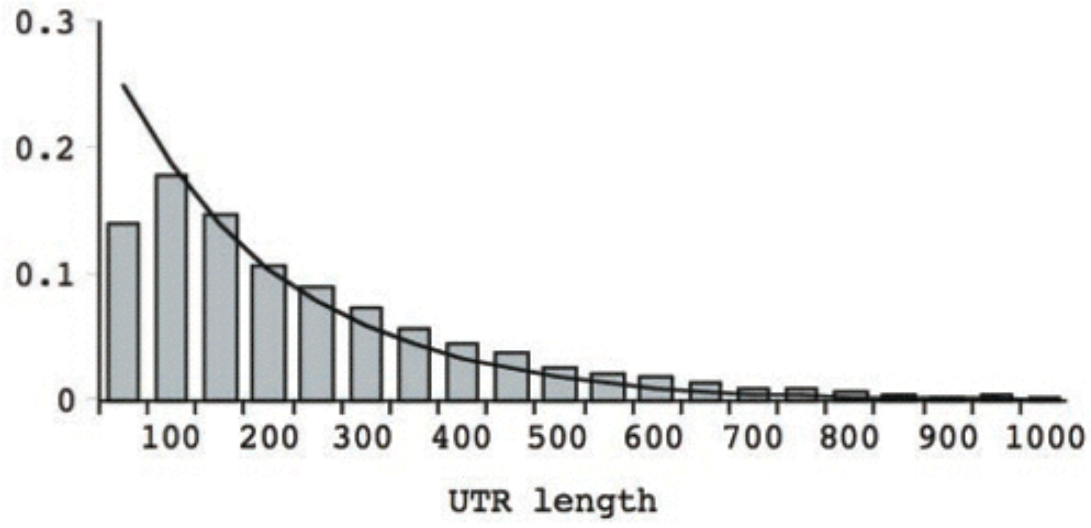

Figure 9. Bars - The length distribution of the 3' UTR sequences in our set of 940. Line - the expectation from a geometric distribution with a mean of 200. The sub-50 nt bar is truncated as a result of our length requirements when building

this dataset. Because of the BLAST word size of 30 nt, no 3' UTRs under this length were sampled. However, our observations in the genome using 3' UTRs from EST aligments show that the line is a fair estimate of the observed frequency of short (< 30 nt) UTRs.

For a state to have its length distributed geometrically, its out-transition probability is related to the mean length $\bar{x}$, such that $P_{out} = \frac{1}{\bar{x}}$, where in this case, $\bar{x}$ = 200nt.

3.3.8.2. Polyadenylation signal (AATAAA)

The information in the 940 observed polyadenylation signals can be modelled using a weight matrix (Table 2). In practice, this evaluates to six consecutive single-column states, each with its own characteristic emission spectrum, and a transition probability of 1 to the next column.

Table 2. Polyadenylation weight matrix. In our implementation, this was modelled by six consecutive single-emission states.

	1	2	3	4	5	6
A	0.778	0.952	0.016	0.819	0.989	0.988
C	0.057	0.003	0.006	0.001	0.007	0.001
G	0.058	0.021	0.004	0.178	0.001	0.002
T	0.106	0.023	0.974	0.002	0.002	0.009

Compared to the background nucleotide distribution Q in the genome, we can find how much information is in each weight matrix column, which has distribution P over the set of nucleotides i. For each column, the relative difference from the genomic distribution of nucleotides (the Kullback-Leibler distance) is:

$$H(P \parallel Q) = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)}.$$

The entropy of this weight matrix relative to a genomic background is 7.58 bits, which is 1.26 bits per column.

3.3.8.3. Spacer state (SP)

Figure 8 shows that the sequence between the polyadenylation signal and the nucleotide 5' of the cleavage site is T rich (rather than pyrimidine rich) and has the distinctive non-geometric length distribution shown in Figure 7. This length distribution is modelled explicitly and the transition probability from this state to the next is 1.

3.3.8.4. Cleavage site (CS)

The cleavage site can be modelled using another weight matrix (Table 3), with the cleavage occurring between the first (-1) and second (1) columns. Cleavages adjacent to As have been reintroduced, resulting in some loss of information relative to Figure 6, though the suppression of G residues is still visible in the +3 position. The out-transition probability from each column is 1, and from the final column, there is obligatory entry to the next state.

Table 3. Cleavage site weight matrix. Four consecutive single emission states. Cleavage occurs between column -1 and 1.

	-1	1	2	3
A	0.483	0.42	0.348	0.276
C	0.131	0.073	0.115	0.145
G	0.137	0.129	0.104	0.086
T	0.249	0.378	0.433	0.49

This matrix contains less information per column (0.09 bits) than does the AATAAA weight matrix.

3.3.8.5. Downstream region (DS)

Figure 8 shows that just 3' of the cleavage site, there is a T-rich section before the nucleotide frequency returns to genomic levels. From the gradual drop seen, it appears that this sequence too has a variable length. This state is thus modelled geometrically with a mean length of 15.

3.3.8.6. Genomic state

The final state we model is one where the nucleotide emission spectrum matches that of the whole genome. After annotation of all the other states, the mean length of these sequences was calculated as 680 nt.

3.3.9. Model topology

The topology of the model is shown in Figure 10.

Figure 10. State transition diagram for *C. elegans* cleavage and polyadenylation site prediction model.

Circular states have geometric length distributions, and thus have out- and self-transitions related to the mean length. For the UTR, DS, and G states, the mean is 200, 15, and 680 respectively.

Boxed states are fixed-length. Once the AATAAA state is entered, there must be exactly six emissions before a mandatory transition to the next state. There is a similar case with the CS state, where there are 4 emissions.

The SP state, shown with the diamond, has a length distribution which is a smoothed version of Figure 7. The length of this state is absolutely restricted to values between 5 and 30 nt. As discussed in Chapter 2, each entry into this state requires evaluation of all possible sequence lengths allowed by the specified length distribution, prior to an obligatory transition into the CS state. This makes HMM decoding algorithms more complex than the standard Viterbi/forward/backward, but the generalised HMM algorithms scale linearly with sequence length.

3.4. Model evaluation

4/5 of the data was used for training and 1/5 for testing. Results for the 5 non-overlapping test sets were averaged. The length parameters were fixed and not estimated for each training set. This is important for the spacer state where the length

distribution was calculated from unambiguous sites that represented a minority of the data. Transition, emission, and length parameters were estimated with a variety of Perl scripts. HMM decoding algorithms were written in Java as discussed in Chapter 2.

3.4.1. Prediction of 3' ends

The performance of the HMM was measured by evaluating sensitivity and specificity measures on a 5-fold cross-validation of a test set, and also by comparing it to heuristic methods based on an AATAAA weight matrix (Blumenthal et al. 1997). Since the location of cleavage sites appears to be imprecise, calculation of accuracy was based on identifying the correct polyadenylation signal and not the cleavage site. Basing accuracy on the polyadenylation signal allows the comparison of the HMM with simple weight matrix methods.

3.4.1.1. Weight matrix strategies

Table 4 shows that a crude scan for all exact matches to AATAAA within 1000 nt of the stop codon correctly identifies 56% of signals, though 46% of the total predictions are spurious.

Table 4. Accuracy of four different weight matrix and two HMM regimes for detecting polyadenylation signals in 3' UTR and downstream sequence. TP true positives, FP false positives, FN false negatives, SN sensitivity (TP/TP+FN), SP specificity (TP/TP+FP).

3' UTR only

Method	TP	FP	FN	SN	SP
All AATAAA	531	453	409	0.565	0.54
First AATAAA	482	286	458	0.513	0.628
First Max Score	562	378	378	0.598	0.598
AATAAA 1 mismatch	883	3034	57	0.939	0.225
Viterbi	662	278	278	0.704	0.704
Posterior > 0.1	767	367	173	0.816	0.676

If we propose that the 5'-most (if there are multiple hits) exact match to AATAAA is the signal, the proportion of signals detected correctly is reduced by 5% but there is an 8% increase in specificity.

Using the first maximum score allows for those sequences that contain a mismatch variant of AATAAA; instead of looking for exact matches to AATAAA, we scan with a weight matrix and call the highest scoring hexamer a hit. In the case of multiple identical hits, the 5'-most one is reported, as this would be the first one exposed on the nascent transcript. This has a sensitivity and specificity of 60%.

A far greater sensitivity (94%) is achieved by reporting all exact matches to AATAAA and all possible single base mismatches, though there is a large penalty to specificity.

3.4.1.2. HMM strategies

Two different strategies were used to evaluate the HMM: Viterbi and posterior decoding. The Viterbi algorithm finds a single maximum likelihood polyadenylation signal in the sequence while posterior decoding determines the probability of the signal at each point in the sequence. Posterior decoding therefore allows one to find the most likely motif and other, less likely ones.

For the posterior, a probability threshold of 0.1 was used, which means that at most 10 AATAAA motifs can be found. The HMM strategies are far more accurate

than the weight matrix methods. The Viterbi algorithm recorded 70% sensitivity and specificity. Posterior decoding maintained a similar 68% specificity but significantly increased the sensitivity to 82%. These results indicate that the context in which a polyadenylation signal appears is an important factor for 3' end formation. Furthermore, it suggests that in cases where the maximum likelihood annotation is incorrect, the observed AATAAA motif can be found by looking at other high-scoring positions.

3.4.2. The stochastic nature of 3'-end site selection

While collecting the data set of unique AATAAA and cleavage sites those genes with high cDNA coverage were unintentionally selected against, as genes containing a larger number of matching transcripts tended to have multiple distinct cleavage sites, such as in Figure 5d.

Figure 11a shows the distribution of cleavage sites at each nucleotide for a 3' UTR with 31 cDNA matches. According to the model, the posterior probability of the AATAAA motif indicates that there is only one such motif in the region. The posterior probability of the cleavage site shows a multi-modal distribution. The frequency of observed cleavage sites is very similar to the posterior probability. Figure 11b shows a case where there are multiple polyadenylation signals and cleavage sites. Here too, the posterior probability of the cleavage site is similar to the observed frequencies. The fact that the model fits the observed distribution so well suggests that it is capturing most, if not all, of the local information used to select the cleavage site.

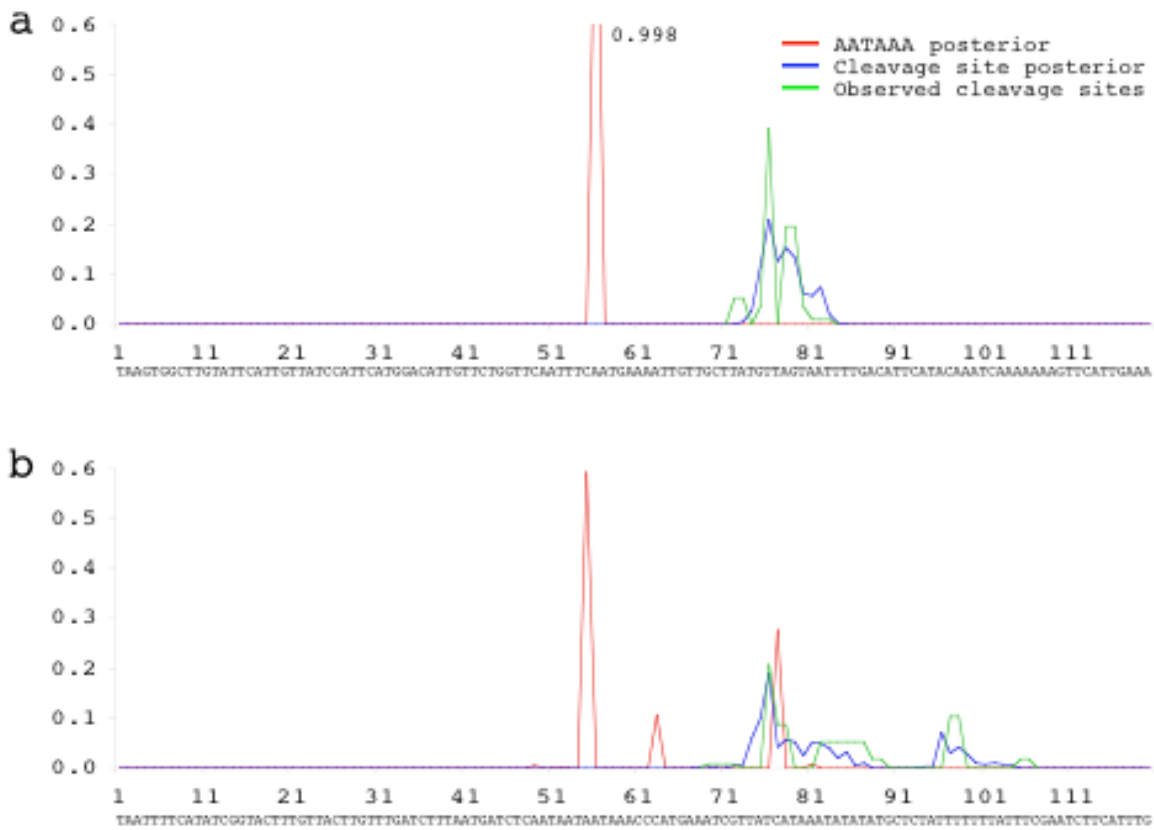

Figure 11. The posterior probability of the AATAAA motif and cleavage site are shown in red and blue lines respectively. The observed frequency of cleavage sites is indicated by a green line. When the cleavage site is ambiguous, the frequency is averaged over the ambiguous positions, which gives the green line a flat peak. (a) 31 mRNAs aligned to gene ZK652.4 show that there are multiple, tightly clustered cleavage sites. (b) 38 mRNAs aligned to gene R09B3.3 show a broad cluster of cleavage sites which are the result of three predicted AATAAA motifs.

3.4.3. Genome-wide scan

The HMM was applied to predict cleavage sites for all the genes in the *C. elegans* genome. There are 22,168 annotated genes in WormBase release WS110 (http://ws110.wormbase.org). For 9,710 of these, a 3' UTR is annotated in WormBase by extending from the stop codon to the 3' end of the 3'-most EST match assigned to the gene. 3'UTRs above 1000 nt are not included. For each gene, the HMM was used

to search the 1000 bases 3' of each annotated stop codon; it annotated the most likely cleavage site as determined by the Viterbi algorithm. We expect 70% of these to be correct, from previous experiments (Table 4). For those genes with 3' UTRs annotated in WormBase, the length of the 3' UTR as determined by ESTs can now be compared with the length predicted by the HMM.

Figure 12 shows the frequency distribution of the distance between WormBase 3' UTRs and the Viterbi prediction for each of their 3' UTR candidates. Peaks are visible around –65 and –10, presumably corresponding to different EST clipping regimes. Based on the graph, we suggest that those predictions that extend the WormBase 3' UTR up to 80 nt are highly likely to be correct because the EST was clipped short. Those predictions that are too short by up to 10 nt are consistent with the local heterogeneity of the cleavage site, and are also likely to be correct. The proportion of predictions falling within the range –80 to +10 is 70%, as expected. This results in a set of 6,570 high confidence identifications of *C. elegans* cleavage sites, which have been made available through WormBase.

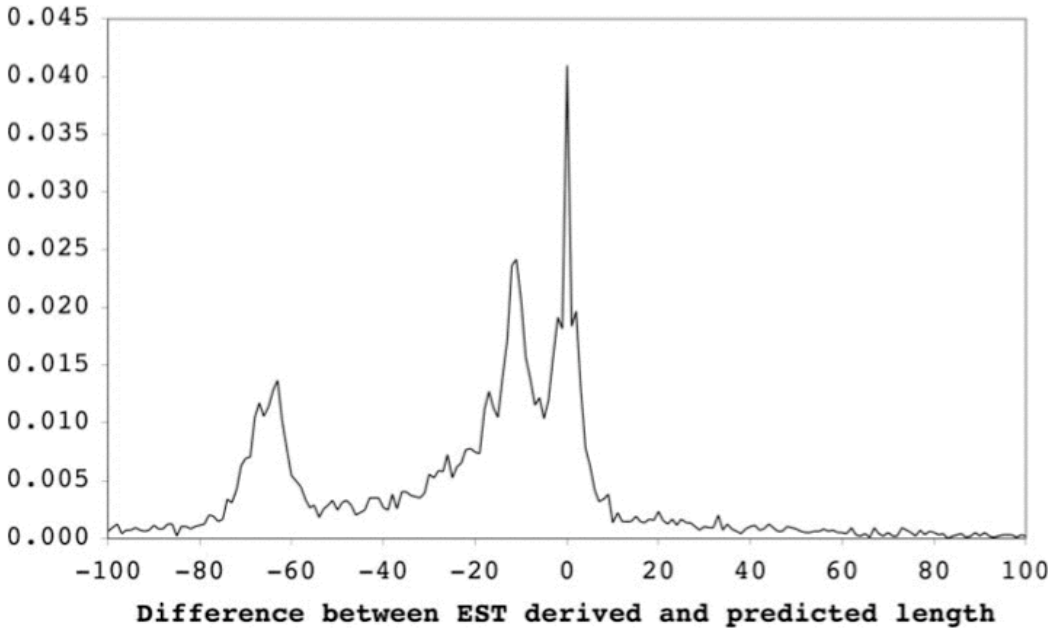

Figure 12. Frequency distribution of the difference between length of 3' UTR as determined by EST alignment and our model.

3.4.4. Posterior probabilities of Viterbi predictions

As stated in the previous section, we have a set of predictions that are likely to be correct, on account of EST support. Each polyadenylation signal prediction is provided with a posterior probability. Figure 13 shows the distribution of posterior probabilities of these Viterbi predictions. We are interested in finding out whether the posterior probability is any indication of the confidence in which we can take the prediction.

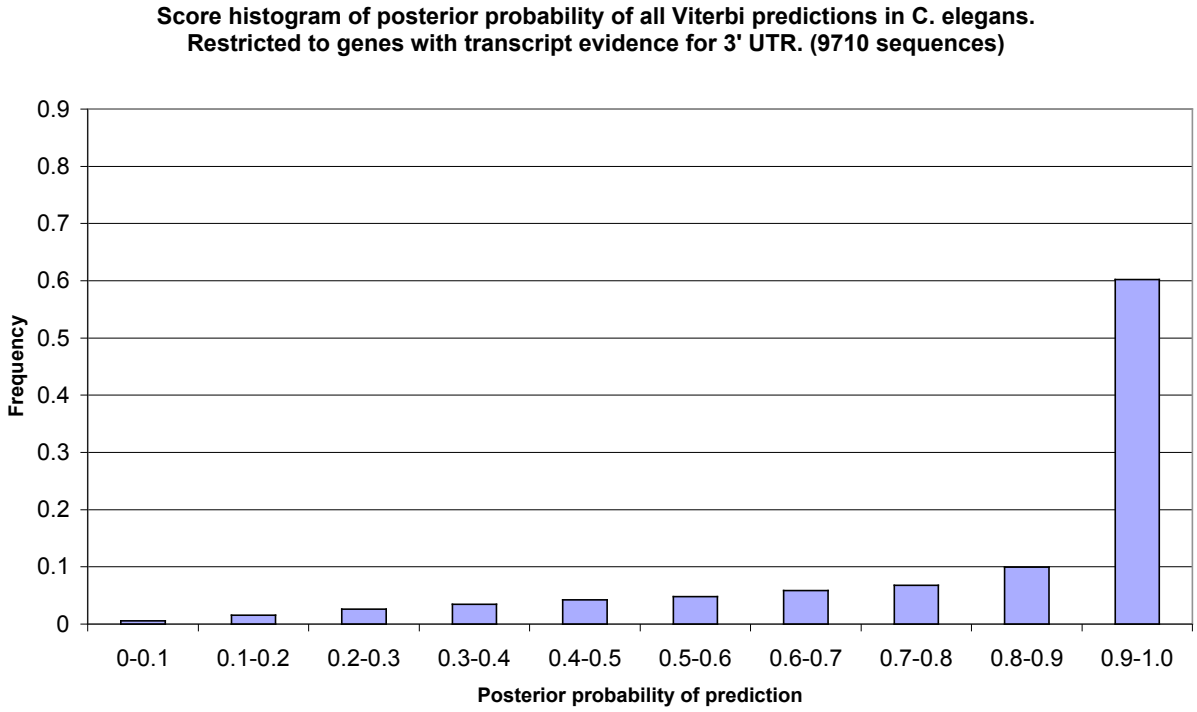

Figure 13. Posterior probability distribution for the Viterbi polyadenylation signal predictions in 9710 3' UTR sequences where a given prediction could be verified by transcript evidence, as ESTs were available.

The posterior probability of Viterbi predictions is highly skewed toward the higher probabilities. This is because all the sequences tested are 3' UTRs and should thus contain at least one polyadenylation signal each.

Of the 9710 predicted signals, 6570 were deemed to be correct from EST evidence, and the rest incorrect. Figure 14 shows that the proportion of these predictions being marked as correct increases with the posterior probability of the prediction. A tenth of all verifiable Viterbi predictions had a posterior probability between 0.8 and 0.9. About 60% of these are correct. 60% of the total predictions have a posterior probability above 0.9 and proportionately, more of these are correct (78%). Again, a number of these will be correct but will not be reported as such on account of the site not being represented in the EST set.

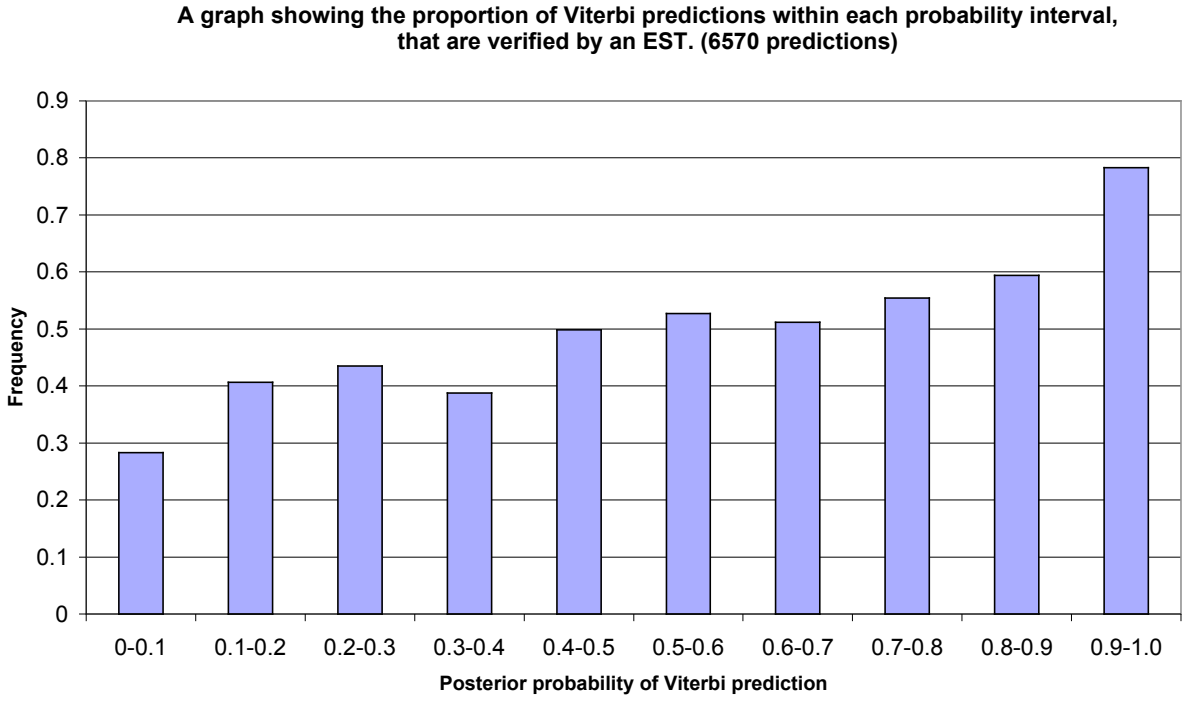

Figure 14. For the 6570 sequences where the position of the Viterbi polyadenylation signal prediction was verified by an EST, this histogram shows

what proportion of all the Viterbi predictions within a particular posterior probability interval were correct.

3.4.5. Testing a scanning model for 3' end recognition

Under a model in which the cleavage and polyadenylation machinery scans along RNA in a 5' to 3' direction, misidentification of the cleavage site may lead to truncated proteins if the cleavage occurs in the coding region. In the experiments above, the only sequence searched for cleavage sites was that found downstream of the stop codon. In order to test weight matrix approaches and the HMM under conditions of a full message scanning model, the methods were evaluated on virtual mature mRNAs containing complete coding sequences plus 1000 nt downstream. In these experiments, the HMM was modified by including an initial group of three coding states, with the third looping into the first, which correspond to the nucleotide frequencies observed in first, second, and third positions within codons. Table 5 shows that the weight matrix methods find a large number of false positives in the coding sequence. However, the specificity of the HMM degrades only slightly; the performance difference of the posterior decoding is particularly small. If the biological machinery scans along the mRNA ‘looking’ for cleavage sites, it is clearly advantageous to ‘see’ more than just the AATAAA motif.

Table 5. Accuracy of various weight matrix and HMM regimes for detecting polyadenylation signals in virtual mature mRNAs. TP true positives, FP false positives, FN false negatives, SN sensitivity (TP/TP+FN), SP specificity (TP/TP+FP), CDS fraction of all signal predictions falling in the coding sequence.

Method	TP	FP	FN	SN	SP	CDS
All AATAAA	525	774	400	0.568	0.404	0.354
First AATAAA	369	436	556	0.399	0.458	0.243
First Max Score	402	523	523	0.435	0.435	0.306
AATAAA 1 mismatch	869	12069	56	0.939	0.067	0.707
Viterbi	632	293	293	0.683	0.683	0.044
Posterior > 0.1	736	405	189	0.796	0.645	0.076

3.4.6. Discussion

In this study, we have made a significant step to improving 3' end prediction in *C. elegans* by developing an HMM that captures global features present in the 3' UTR. HMMs have become popular in the sequence analysis community because they offer a method to incorporate diverse sequence features under a rigorous probabilistic framework, and because they have established decoding algorithms. HMMs are stochastic models and this fits well with cleavage site selection, which appears to be a stochastic process. In cases where there are numerous transcripts aligned downstream of a stop codon, we found that the posterior probability of cleavage sites derived from the HMM mirrors the frequencies of experimentally observed cleavage sites. This suggests that the HMM faithfully represents the local requirements of 3' end formation. It also suggests that the cleavage site is not a precise, locatable entity, and it would be more accurate to refer to a frequency distribution of the most probable sites.

3.4.7. Incorrect predictions

In order to determine why the HMM missed roughly 20% of real polyadenylation signals, the 3' UTRs of the incorrect predictions in were examined in WormBase using ACEDB (http://www.acedb.org). In approximately 30% of cases, there were additional transcripts (without poly-A tails) that supported the predicted 3' end. These 3' ends may therefore fall into the class depicted in Figure 5c with multiple signals. Unfortunately, we do not have access to the raw traces and cannot extend the sequence into the poly-A tails to find the cleavage site. Thus, we believe it is likely that a significant proportion of the false positive predictions are real sites.

Another class where 'incorrect' predictions are real include instances where the predicted and observed AATAAA motifs were just a few nucleotides apart. This occurs in roughly 5% of the incorrect predictions. The original maximum likelihood assignment of the polyadenylation signal and the cleavage site was based on a weight matrix for the AATAAA motif and a probability distribution for the distance to the cleavage site, taken from a subset of the whole training data. As the HMM is a more explicit model of the 3' end, in these cases the HMM prediction may be more accurate than the initial maximum likelihood annotation.

Approximately 25% of the missed predictions (5% of the whole set) resulted from oversights in collecting the data. It was assumed that unlabelled genomic sequence downstream of a terminal exon contains a 3' UTR followed by genomic sequence. This is not always the case. Some 3' UTRs contain tentative evidence for an intron, which means the HMM and the polyadenylation machinery see different sequences, though we should bear in mind that transcripts with introns 3' of the STOP codon are targets for nonsense mediated decay (Chen et al. 2003; Neu-Yilik et al. 2004). Some 3' regions contain transcripts that do not appear to correspond to the 3'

UTR of the labelled gene and instead contain novel genes such as non-coding RNA genes. There were also cases where the aligned transcript had a better match elsewhere in the genome.

In the largest fraction of missed polyadenylation signals, roughly 40% of the errors or 8% of the total, the cause of the error cannot be determined. It may be that with greater transcript coverage some of these 3' ends will turn out to have multiple AATAAA motifs. Alternatively, these 3' ends may form a different class, perhaps with specific factors that direct their positioning. Indeed, we know that for some genes, such as the replication-dependent histones, the cleavage site is determined not by a polyadenylation signal, but by a conserved stem-loop (Dominski et al. 1999). No unusual compositional biases were detected around the missed sites though, so the reason for these incorrect predictions remains a mystery.

Taken together, based on the fact that a number of the incorrect predictions are potentially correct, the HMM is more accurate than we can reliably report, with likely over 90% sensitivity.

3.4.8. Biological implications

The HMM contains states for the polyadenylation signal, the cleavage site, and regions on either side of these features. It does not explicitly model other sequence elements, but it may be taking these into account. For example, the downstream state is T-rich and this roughly corresponds to a CstF binding site. Whether or not CstF binding downstream of the cleavage site is actually required for 3' end formation is not known and may be dependent on the nature of the AATAAA motif (MacDonald et al. 2002).

Correct identification of full-length transcripts is important both for studying the process of 3' end formation and for interpreting and integrating experimental results, such as Northern blots, SAGE tags, and microarrays. Another implication for this work is that it may improve the quality of gene prediction. One of the difficulties in gene prediction is identifying the terminal exon. Misidentification can cause single genes to be split or neighbouring genes to be fused. Employing a more descriptive model of 3' ends should help reduce this problem.

4. Polyadenylation Signal Prediction in Other Eukaryotes

4.1. Introduction

In chapter 3 we showed that the short and long range signals encoding the site for *C. elegans* transcript cleavage and polyadenylation can be robustly modelled by PAjHMMA. In this chapter, we are interested in seeing (a) how the specification of this signal may vary in different organisms – especially given the variation in nucleotide compositional biases across different genomes, and (b) whether a PAjHMMA HMM can successfully capture this information and thus predict polyadenylation signals accurately in other species.

Nucleotide frequencies around the cleavage site in other species suggest that the global and local signals used to specify polyadenylation sites appear to vary (Graber et al. 1999). Thus the existing *C. elegans* polyadenylation signal model would not be of much use in any other organism - although it does work in the related nematode *C. briggsae* (Chapter 6). Given the flexible nature of PAjHMMA models and the efficacy of the *C. elegans* model discussed previously, we attempt to build such models for other species.

A new method for building cleavage site datasets is introduced, though the logic behind it remains the same as that used in *C. elegans*. There is a large amount of cDNA evidence for mouse and human. This, coupled with the size of the genomes, suggests that it would be easier to obtain datasets of experimentally determined cleavage and polyadenylation sites directly from the Ensembl gene build (Hubbard et al. 2005), rather than repeat the analyses that create the data. Nucleotide frequency plots for these mammalian models show that both species have similar signals

dictating the position of the polyadenylation and cleavage site. There are also significant similarities to the *C. elegans* model in terms of state length and topology, though neither contains the long range pyrimidine rich UTR signal exhibited by the nematode.

Initial data from a previous study gained in this way for *Drosophila melanogaster* shows that the model for the fruitfly is quite different from all those previously observed, on account of its cleavage sites being in a region that is A-rich, rather than T or pyrimidine-rich as observed in the other species. Ensembl does not provide us with enough cleavage sites to build statistically significant models for the fly, but as there are a large number of cDNAs available, a cleavage site dataset was built using the same alignment method as in *C. elegans* detailed in chapter 3.

4.2. Data Acquisition

4.2.1. Mouse and Human

To collect experimentally verified cleavage sites for human and mouse, the relevant Ensembl databases (v25.34.e.1 and v25.33.a.1 respectively – both October 2004) were queried using the EnsJ Java API (Stabenau et al. 2004). This workflow can be summarised as below.

```
Foreach Gene

      Get all Transcripts

            Discard if Gene has more than one Transcript

            Discard if Transcript has more than one ThreePrimeUTR

            For the single Transcript

                  Find all SupportingFeatures

                  Discard those that are not DNADNAAligments

                  For the 3'-most DNADNAAlignment

                        Obtain the cDNA from EMBL

                        Check if the last 50 nt of the Alignment are
                           identical for the genome and the cDNA.

                        Check if the cDNA contains a pure poly-A tail,
                           starting just after the point where the
                           Alignment ends
```

This logic is the same as that used in the *C. elegans* dataset – the region isolated was the genomic sequence flanking the point where a polyadenylated mRNA dissociates from being aligned to genomic sequence into a poly-A tail. Model building was restricted to include only those cleavage sites originating from genes with single

transcripts and single 3' UTRs. This is so that the building procedure would resemble that employed for the *C. elegans* model, in which only single transcript genes were used. As we have already observed, this does not compromise the ability of the model to recognise multiple polyadenylation signals and sites.

Using data from the Ensembl gene build allowed the collection of verified cleavage and polyadenylation sites for 2706 genes in human, and 4051 in mouse.

4.2.2. Fruitfly

Building a polyadenylation signal model for *Drosophila melanogaster* is also of interest, as there are areas of nucleotide bias, such as a diffuse A-rich region including the AATAAA motif, extending from the cleavage site to 40 nt upstream, but there appears to be no long range pyrimidine or purine bias that was characteristic of the *C. elegans* 3' UTR. Another difference is at the cleavage site, where the majority (>90%) of cleavages occur within a run of As.

The dataset was built in a similar manner to that for the worm. A batch download of 3' UTR sequences from EnsMart (Kasprzyk et al. 2004) showed that 95% of fruitfly 3' UTRs are shorter than 2000 nt. Therefore 2000 nt sequence 3' of each predicted gene's stop codon was isolated. These sequences were truncated if they overlapped into the next gene. 20601 polyadenylated mRNAs were downloaded from EMBL/Genbank and aligned to the extended 3' UTR set as described in Chapter 3. This led to the generation of 3068 cleavage sites.

4.3. Nucleotide Frequencies

Figure 15 shows the distribution of nucleotide frequencies 50 nt either side of the cleavage site in four organisms.

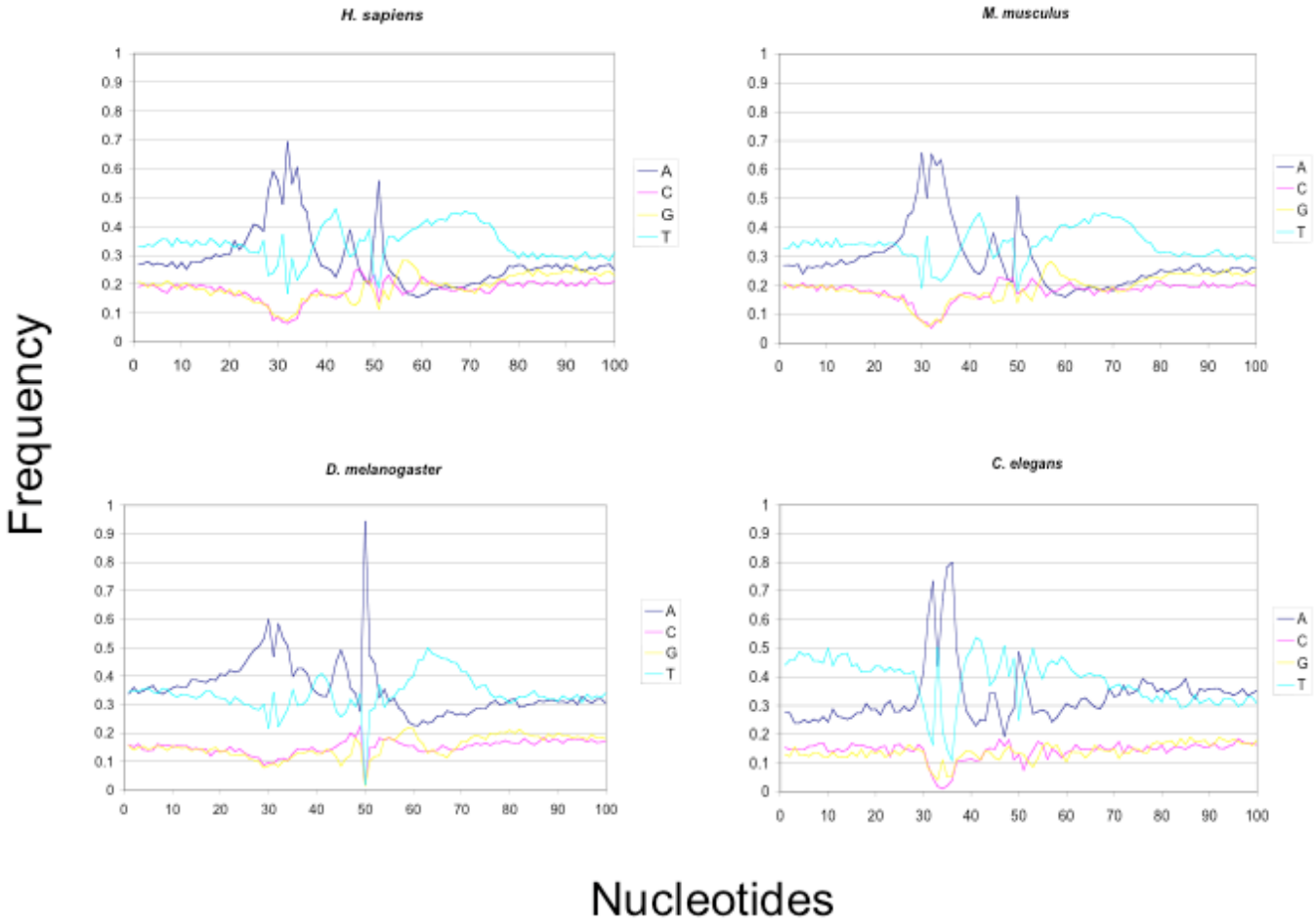

Figure 15. Graphs showing the nucleotide distribution around the cleavage sites of H. sapiens, M. musculus, D. melanogaster, and C. elegans. The maximum likelihood cleavage site occurs at 50 nt in each case.

Figure 16 is an example from mouse, showing how nucleotide frequencies vary over a longer range. A similar graph exists for human (not shown).

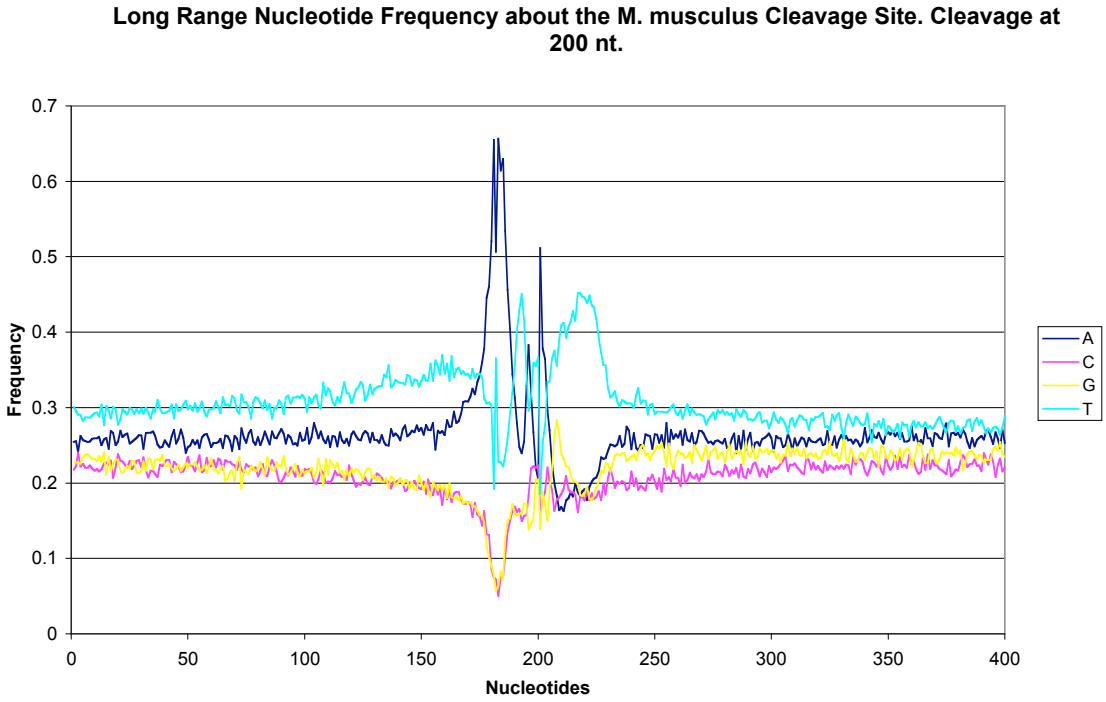

Figure 16. M. musculus graph showing how nucleotide frequency varies over longer ranges.

Figure 15 provides a graphical representation of the local nucleotide frequency signals captured nearest the cleavage site. In both the mammals and the fly, there is a pronounced T-rich region (preceded by elevated levels of G), just downstream of the cleavage site, corresponding to the CStF binding region. Between the polyadenylation signal and the cleavage site, the spacer is T-rich followed by A-rich, followed by T-rich. This latter is also visible to a lesser extent in *C. elegans*. Of the four species shown here, the position of the polyadenylation signal (relative to the cleavage site) seems to be more constrained *C. elegans* than in the others, as can be seen by the relative widths of the A-rich AATAAA motif peaks. The long range nucleotide frequency upstream of the cleavage site – maintained throughout the 3' UTR - is

slightly T-rich and provides some information in mammals, though less than in *C. elegans*.

4.3.1. Long Range 3'UTR (UTR1) and Genomic (G) States

Table 6 shows how much 3' UTR sequence differs from downstream genomic nucleotide frequency levels in different species. The UTR1 state extends from the stop codon to 50 nt upstream of the cleavage site. The genomic state is intended to model the genomic context in which genes appear, and extends from 50 nt downstream of the cleavage site. There is variation between the species as to how much the whole 3' UTR differs from the downstream genomic nucleotide distribution. The worm UTR has a distinctive nucleotide emission profile, with 0.035 bits per base compared to compared to the genomic distribution over an average 215 nt, or 7.67 bits in total. Human only has 0.00108 bits per base, over an average of 815 nt, giving 0.88 bits. The mouse has 0.0011 bits over a similar length, thus providing slightly more information at 0.91 bits. Fly contains more information per base (0.0086 bits) than the mammals, giving 2.51 bits over a mean length of 291 nt

Table 6. Proportions of each nucleotide in several species' UTR1 states and genomic downstream regions. C. elegans has no 50 nt UTR2 state, so extends right up to the polyadenylation signal. The mean length of each organism's UTR1 state used in the model is also given.

	UTR1	Genome	
C. elegans	27.3	32.6	A
(215nt)	19.9	17.5	C
	12.6	17.7	G
	40.3	32.2	T
H. sapiens	26.1	26.4	A
(815nt)	22.3	23.2	C
	22.2	23.6	G
	29.3	26.8	T
M. musculus	25.9	26.9	A
(830nt)	22.5	22.8	C
	22.6	23.0	G
	29.0	27.3	T
D. melanogaster	31.8	27.7	A
(291nt)	19.5	21.4	C
	18.3	21.8	G
	30.4	29.2	T

In *C. elegans,* this long-range nucleotide distribution does not change appreciably between the gene's stop codon and the polyadenylation signal, but for most other species (an example of which is seen in Figure 16), there is a slight change about 50 nt upstream of the AATAAA motif, which we model with a separate HMM state to that modelling the rest of the 3' UTR. This second UTR state is not used in the *C. elegans* model, but it is this state (UTR2) that is visible on the 5' end of the local cleavage models shown in Figure 15.

4.3.2. Second 3' UTR (UTR2) State and purine to pyrimidine asymmetry

The most striking aspect of the nucleotide frequency in the UTR2 state (as indeed with the whole 3' UTR) is the asymmetry of nucleotide bias. This is most apparent in worm, appears to a lesser extent in the mammals, but is not present at all in fruitfly.

For any whole genome, or indeed any double stranded DNA, the number of pyrimidines and purines must be equal. However, we notice in worm, human, and mouse, that the proportion of T bases in the region just upstream of the AATAAA motif is greater than the proportion of As. This asymmetry is possible as the 3' UTR is part of a transcript, which is a single stranded feature. Globally, there is no preferred strand for bases, but transcribed features can have preferred bases on account of the increased mutability of single stranded DNA. It has been suggested (Niu et al. 2003; Touchon et al. 2004) that transcribed sequence should show a C to T mutation bias. This would explain the observed excess of T, but not the less strong excess of C over G seen in *C. elegans*. The HMMs described here are built to recognise features having characteristic nucleotide frequencies. As transcribed DNA is under different mutation pressure to non-transcribed DNA, this long-range asymmetry provides a strong signal that the sequence in question is likely to be transcribed.

4.3.3. A-rich state

All four species show an A-rich peak some 20 nt upstream of the cleavage site. This peak corresponds to an A-rich polyadenylation signal.

In mouse and human, maximum likelihood signal and cleavage sites were calculated as in chapter 3 using previously published data (Beaudoing et al. 2000).

In fruitfly, each sequence had a likely polyadenylation signal annotated, again using the maximum likelihood method. As there was no prior data regarding the distribution of different AATAAA motifs in *Drosophila*, some worm data had to be used. This involved finding the maximum scoring position of the *C. elegans*

AATAAA motif weight matrix, scaled by a fly AATAAA – cleavage length distribution. As with *C. elegans*, this length distribution is found by isolating sequences with an unambiguous exact match to AATAAA for which the cleavage site does not occur adjacent to an A. As only some 6% of cleavages in *Drosophila* can be located exactly (Figure 5a, contrasted with b), this approach was only possible on account of our relatively large dataset, which provided 105 sequences from which to calculate the spacer length distribution.

Figure 17 confirms our earlier observation that there is a wider distribution of spacer lengths in the mammals and the fly, compared to the worm. In addition, the other spacers seems to be slightly longer than in worm, with means of 17 and 16, and 17 nt for human, mouse and fly respectively, compared to 14 in *C. elegans*. This may be as a result of different steric requirements of the proteins in the polyadenylation and cleavage complexes in the four organisms.

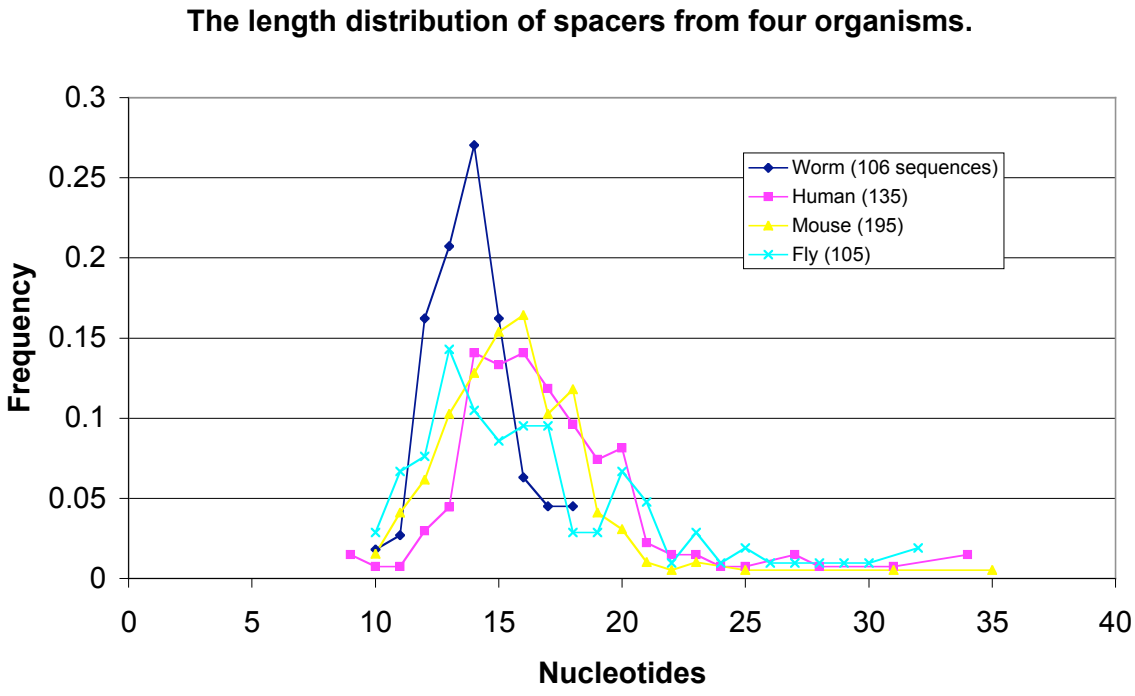

Figure 17. A frequency distribution of the lengths of sequence between unambiguous matches to AATAAA and precisely locatable cleavage sites.

The weight matrices for the four species do show some differences from each other, though the mouse (Figure 18) and human (Figure 19) signals are similar. It is pleasing to see that the fly signal (Figure 20) appears to differ from the worm signal (Figure 21), despite maximal fit to the worm weight matrix being selection criteria for the fly polyadenylation signal.

Figure 18. M.musculus AATAAA motif.

Figure 19. H. sapiens AATAAA motif.

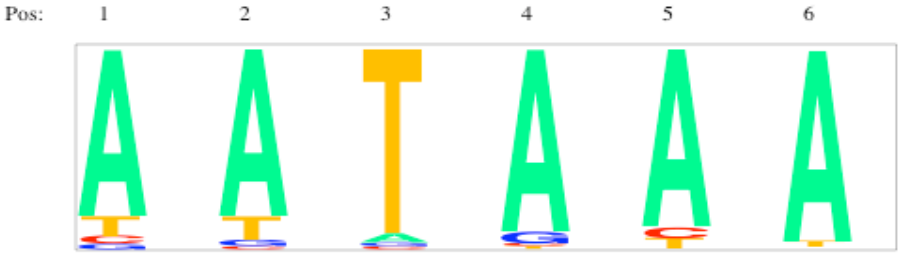

Figure 20. D. melanogaster AATAAA motif.

Figure 21. C. elegans AATAAA motif.

Mouse and human seem less resilient to variations at the first position than the other two species. Interestingly, it appears that the most common non-canonical AATAAA motif differs between species; AATGAA (worm) seems uncommon in vertebrates, which prefer ATTAAA.

4.3.4. Spacer and cleavage site

The spacer is the region between a putative AATAAA motif and the confirmed (or maximum likelihood) cleavage site. In the worm, we used a single T-rich state with an explicitly specified length distribution. In the two vertebrates, there is a peak of As that interrupts a T-rich region. Thus for mouse and human, we have a spacer state with a length distribution calculated as in chapter 3, which extends to cleavage-6. The peak of As, the return to T-richness, and the cleavage site itself are modelled by a weight matrix. All species except the worm exhibit a rise in levels of G just downstream of the cleavage site, so for mouse and human, we use a 16-column weight matrix, capturing 6 nt upstream of the cleavage site, and 10 downstream.

The fruitfly spacer seems to have two parts, a T-rich and an A-rich part. We model these using an explicit length state for the T-rich state, and capture the 8nt upstream of the cleavage site in an 18 nt cleavage site weight matrix.

Figure 22 and Figure 23 show a graphic of how nucleotide frequency varies nearest the cleavage site in human and mouse. The weight matrix captures the second two parts of the three-part spacer (namely the transition from A-richness to T-richness in columns 1-5). Both organisms tend to cleave within a run of As. It has been reported that a CA dinucleotide is favoured prior to the cleavage site, (Sheets et al. 1990), but this study is based on a much simpler strategy for dealing with a cleavage in a run of A, such that the cleavage was always assumed to fall after the first A in a run of As. Additionally, this finding has been refuted by a mutational analysis, (Chen et al. 1995).

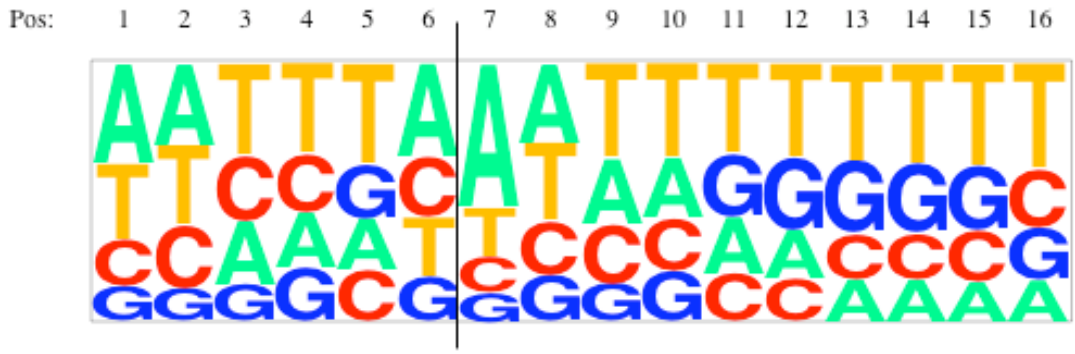

Figure 22. H. sapiens cleavage site weight matrix. Cleavage between positions 6 and 7.

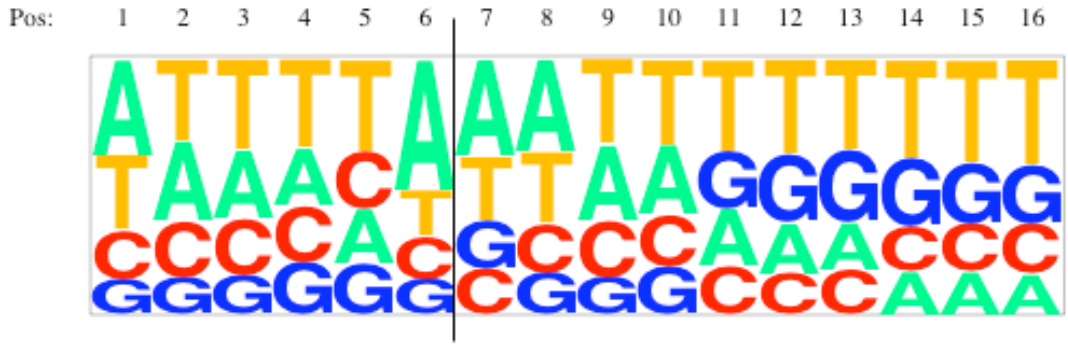

Figure 23. M. musculus cleavage site weight matrix. Cleavage between positions 6 and 7.

Downstream of the cleavage site, the beginnings of a T-rich region can be seen, with G beginning to be preferred to A.

Figure 24 shows the *D. melanogaster* cleavage site weight matrix. The preference for an A before the cleavage site is quite striking. It is unknown whether Cleavage Factors have any sequence specificity, or if they are directed by protein-protein interaction. Cleavage sites seem to be A-rich, which confirms a previous mutational analysis (Chen et al. 1995), though the extreme preference for cleavage 3' of an A seems unusual. Early work from mammals suggests that poly-A polymerase has slight preference for substrates with a terminal A (Bienroth et al. 1993). The reason why the *Drosophila* cleavage site shows such an extreme preference for cleaving after an A is unclear.

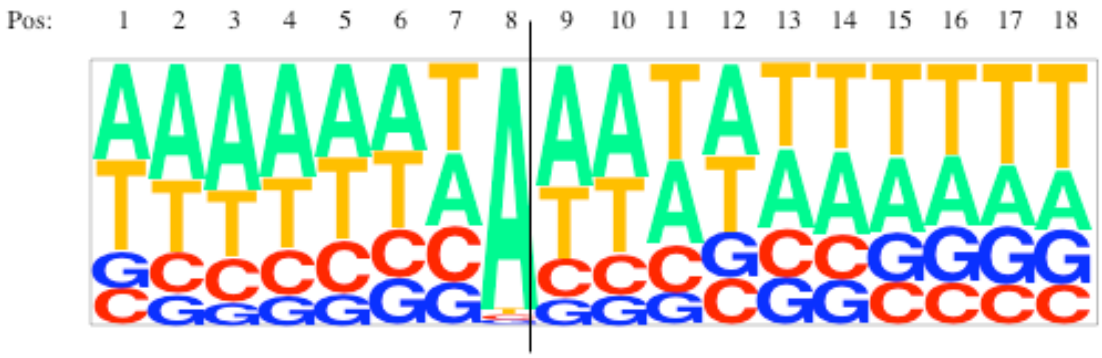

Figure 24. D. melanogaster cleavage site weight matrix. Cleavage between positions 8 and 9

4.3.5. T-rich (T) and Downstream region (DS)

All organisms show a T-richness up to 30 bases 3' of the cleavage site. In *C. elegans*, this is not particularly pronounced compared with the rest of the 3' UTR, but the three other organisms show a definite elevation of T. This is likely to be a CStF binding region. As the mammals and fly (Figure 15) show increased G for 10 nt just 3' of the cleavage site, this region is added to the cleavage site weight matrix, and the 20 nt T-rich region is modelled by a separate state.

Following the T-rich region is another 20 nt region where there is still asymmetry in the nucleotide distributions. This second downstream region is modelled by another state. The rest of the sequence is modelled by the genomic state discussed earlier.

4.4. Model testing

4.4.1. Introduction

The maximum likelihood cleavage sites for each of the three species were split into test and training sets. PAjHMMA models were trained on each of the training sets, and evaluated at the level of AATAAA motif positioning, both in Viterbi (maximal scoring) and posterior decoding (all nucleotides being in AATAAA motif state with probability > 10%) modes. Test sequences for each species were the sequence downstream of each confirmed stop codon such that 95% of 3' UTRs were contained within this length, without the sequence being allowed to extend into the next gene. This length was 4000nt for human and mouse, and 2000nt for fly.

The flexibility of PAjHMMA means that it is easy to change the modelled emissions to dinucleotides; that is, to build a first order Markov model. Dinucleotide datasets were built from the cleavage site datasets mentioned, by counting.

To test the efficacy of the extra information non-AATAAA states, a simple weight matrix scan was also carried out, using the six AATAAA motif states on their own. As reported in chapter 3, the best weight matrix regime, and the only one found to have acceptable accuracy was to report the maximum hit from the AATAAA weight matrix. In the event of multiple, equally scoring hits, the 5'-most hit, being the first to be exposed in the nascent transcript, was reported.

Publicly available software from two previously published methods for human and human/mouse polyadenylation signal prediction, ERPIN (Gautheret et al. 2001), and PolyADQ (Tabaska et al. 1999) were also used for comparison.

4.4.2. ERPIN

This program reports hits to a set of 1st order weight matrices, ranging from the AATAAA motif to 46 nt downstream. This should capture signals encoded by the cleavage site and the downstream rises in G and T. Default parameters were used (http://tagc.univ-mrs.fr/erpin/), which were tuned by the authors empirically to retain sequences with a polyadenylation signal hit with a score greater than 70% of the maximum, and with the downstream region cutoff of 74%. This method does not accept any polyadenylation signal other than AATAAA and the ATTAAA variant.

4.4.3. PolyADQ

A weight matrix for the AATAAA motif and a 10 bp downstream weight matrix were constructed by Gibbs sampling. This algorithm finds all occurrences of AATAAA and ATTAAA in human and mouse, and uses a quadratic discriminant function to decide whether the weight matrix hit is a real polyadenylation signal by considering the downstream hit and the distance between the two.

4.4.4. Results

The accuracy with which each algorithm identifies the correct polyadenylation signal using the HMM, weight matrix and published methods is shown in Table 7.

Table 7. TP, true positives; FP, false positives; FN, false negative; SN, sensitivity (TP/TP+FN); SP, specificity (TP/TP+FP).

Method	Mouse(551)					Human(705)				
	TP	FP	FN	SN	SP	TP	FP	FN	SN	SP
Viterbi	285	266	266	0.517		285	420	420	0.404	
1st order Viterbi	263	288	288	0.477		330	375	375	0.468	
Maximum weight matrix	269	282	282	0.488		347	358	358	0.492	
Posterior >0.1	371	630	180	0.673	0.371	379	949	326	0.538	0.285
1st order Posterior >0.1	310	503	241	0.563	0.381	395	704	310	0.560	0.359
ERPIN	287	605	264	0.521	0.322	344	917	361	0.488	0.273
PolyADQ	403	1049	148	0.731	0.278	391	766	314	0.555	0.338

Method	Fly(500)					Worm(940)				
	TP	FP	FN	SN	SP	TP	FP	FN	SN	SP
Viterbi	193	307	307	0.386		662	278	278	0.704	
1st order Viterbi	243	257	257	0.486		671	269	269	0.714	
Maximum weight matrix	230	270	270	0.460		562	378	378	0.598	
Posterior >0.1	290	749	210	0.580	0.279	767	367	173	0.816	0.676
1st order Posterior >0.1	302	574	198	0.604	0.345	777	254	163	0.827	0.754
ERPIN	-	-	-	-	-	-	-	-	-	-
PolyADQ	-	-	-	-	-	-	-	-	-	-

There is an issue regarding how false positives are calculated. In this work, if the model predicts a polyadenylation signal where there is none annotated according to our data sets, then this has been counted as a false positive. However, as mentioned in chapter 3, there is no way to know whether a given prediction is never used as a real polyadenylation signal. Thus, whilst the false positive rate given may not be an accurate representation of the real value, it does represent a worst-case value. A more realistic rate could be found by finding the number of posterior decoding predictions with greater than 10% probability made per kilobase of random sequence.

At a glance, polyadenylation signal prediction appears to be more difficult in each of these three species than it is in *C. elegans*. The benchmark in chapter 3 was to see if prediction using context information to model the whole 3' UTR was more effective than just looking for a close match to AATAAA. In the worm, a zero order model outperformed the best weight matrix regime by over 10% at sensitivity and

specificity levels. In human and fly, using just the AATAAA weight matrix component of the model outperforms using the whole model, so using context information is misdirecting predictions. Of the three species introduced in this chapter, only in the mouse do zero order Viterbi predictions outperform a weight matrix at the sensitivity level, though this is by less than 3%.

Increasing the order of the HMM to model dinucleotides had different effects on the Viterbi hit in human and mouse. In mouse, the dinucleotide information seems to reduce prediction accuracy a little, whereas it has a beneficial effect in human. In *Drosophila,* a 10% increase in sensitivity and specificity occurred, outperforming the AATAAA weight matrix on its own. This increase was the largest observed, and was unexpected, considering that using dinucleotides in *C. elegans* had a negligible effect on sensitivity.

Posterior decoding reports not the best scoring hit, but rather calculates the probability of each nucleotide being in a particular state. Posterior > 0.1 reports all occurrences of sequences entering the AATAAA motif state with probability > 10%. This predicts an average of 1.5 sites per sequence, though it can predict up to 9 potential polyadenylation signals per sequence. In all four species, this method has increased sensitivity compared to zero and first order Viterbi predictions, and also relative to the weight matrix, whilst maintaining tolerable specificity. As our test sequences were annotated to contain only one polyadenylation signal, we expect a decrease in specificity. However, in *C. elegans,* this decrease is less than 3%, suggesting that posterior decoding is correctly identifying 'weaker', correct polyadenylation signals that were missed by Viterbi predictions. In the three species discussed here, the drop in specificity was considerably higher. In all of them, there were significant gains in sensitivity, though none approached the 82% seen in worm.

Lexicalizing the emissions into dinucleotides in posterior decoding mode had a varied effect on sensitivity (a substantial drop vs zero order posterior decoding in mouse, but a small rise in fly and human), but specificity was consistently increased by the prediction of fewer false positives.

Both ERPIN and PolyADQ are restricted to AATAAA/ATTAAA, meaning that no other variants can be predicted, and that the maximum sensitivity is 80% in human and 86% in our mouse set. PolyADQ is the best performer in mouse, with a sensitivity of 73%.

For each method, accuracy is almost always higher in mouse than in human. One interesting observation here is that ERPIN, despite being trained on human data, also performs slightly better in mouse than in human. This may be explained by our earlier observations that there is much similarity in the human and mouse cleavage site models, but that the mouse cleavage site itself is specified with slightly higher information content than in human, making it slightly easier to detect. Alternatively, it may be a consequence of the set of genes that were selected for the test sets.

The HMM is arguably outperforming PolyADQ in mouse, depending on the relative importance attached to sensitivity and specificity. In human, posterior decoding with dinucleotides outperforms both published methods.

One issue with these two methods is that parts of our test set might have been included in their training data, so their performance scores on our test set may be overestimates.

4.5. Discussion

4.5.1. Sensitivity

Given the success of the zero order hidden Markov model strategy in *C. elegans,* the measured sensitivities in the other species, especially human, are disappointing. It is surprising that a simple weight matrix outperforms a model that adds context information and looks for a global maximum. A partial explanation could be at the level of the polyadenylation signal itself. In human and mouse, the two most frequently occurring signals, (AATAAA and ATTAAA) account for 80 and 86% of all signals in the two respective organisms. This figure is only 69% (AATAAA and AATGAA) in *C. elegans*. This means that the weight matrix contains more information in the two mammals, as it appears to be more constrained. In addition, because the AT composition of the human and mouse genomes is lower than in the nematode, there is a lower probability of an AATAAA occurring by chance, so the probability of a given AATAAA being a real polyadenylation signal is higher. To compensate for the reduced information in the worm weight matrix, context information has to be used. Where it is not required, excess context information can cause incorrect prediction; it has been observed previously in a study on multiple polyadenylation signals, that adding context information from upstream of the human AATAAA motif had a negative effect on prediction accuracy (Legendre et al. 2003).

One of the major factors allowing us to identify the worm polyadenylation signal correctly might be the large amount of long range context information provided by the whole 3' UTR having a very distinctive, biased nucleotide distribution. This striking distribution, constant throughout the whole 3' UTR, is not seen in any of the other species. However, it is not clear whether this is information available to the

biological cleavage process, or a secondary consequence of mutation biases on transcribed sequence.

Analysis of those human polyadenylation sites incorrectly identified showed no markedly different nucleotide composition to those identified successfully, so we do not believe that poor performance is due to a specific type of cleavage site that is a poor fit to our model.

One of the reasons for low sensitivity could be that the Viterbi path used by our model is obliged to make exactly one prediction. It may be that a sequence contains one or more additional as-yet unconfirmed cleavage sites, which have a higher probability under our model than that in our test set.

At least 54% of human mRNAs are subject to alternative polyadenylation (Tian et al. 2005), and as we shall see in the next section, as more transcript data is analysed, this number is likely to increase. With this in mind, for species in which alternative polyadenylation is this common, it might be a good idea to build models specifically modelling mRNAs with 2, 3…n confirmed cleavage and polyadenylation sites. However, the aim of this chapter was to emulate the work carried out on worm transcripts, in which we discarded the small number of transcripts with multiple polyadenylation sites.

Using posterior decoding allows us to predict multiple polyadenylation sites if each site represents a probable path through the dynamic programming matrix. This is one reason why sensitivity under posterior decoding is consistently better than under Viterbi predictions. However, this method is only suitable when the probabilities of the two paths both pass some threshold (0.1 in our case). Another way of modelling multiple polyadenylation would be to allow our PAjHMMA model to loop into an AATAAA motif state at will, predicting multiple sites in a single pass, though this

would require building of more complex data sets to train emission and transition parameters. Another factor that could be added for sequences with multiple polyadenylation signals is to use all cDNAs from a single library, so that if one site had many polyadenylated mRNAs and another had fewer, some kind of weighting strategy for the nucleotide frequency distributions at each site could build a more realistic model.

4.5.2. Specificity

Table 7 shows that no method reaches 50% specificity, apart from in the worm. This is because of the large number of false positives, caused especially by the methods which can predict multiple polyadenylation signals in a single sequence, and by the fact that 3' UTRs are longer in mammals and flies. Our datasets were built specifically with sequences containing only one confirmed cleavage and polyadenylation site. If an algorithm predicts a signal in the test set where there is none annotated, this is marked as a false positive. However, it is not fair to say that this predicted site is not a real site, simply because there is no (as yet) polyadenylated cDNA evidence for it. There is no way to prove that a sequence is not a polyadenylation signal. Many such false positives in *C. elegans* were subsequently found to have EST evidence, so the specificity value obtained represents a lower bound for some actual value.

4.6. Conclusions

We have shown in this chapter that the method used in chapter 3 can be extended to build polyadenylation signal models for other species, and that the software developed for this purpose is robust and flexible. Although it performs best on the species for which it was developed, there are some interesting results in other species. On our test data the human PAjHMMA HMM is the best performer compared to previously published methods.

5. On the Evolution of 3'UTRs and Polyadenylation Signals

5.1. Introduction

In this chapter, we look at how polyadenylation signals evolve between *C. elegans* and *C. briggsae*. This was done by using a set of 3' UTR sequences from a set of genes that are considered to be 1:1 orthologues at the protein level.

We analyse a set of orthologous pairs where the polyadenylation signals are part of a BLAST alignment and align to each other. We observe an interesting pattern of mutation and conservation between aligned polyadenylation signals of orthologous pairs. We also consider cases when it appears that non-homologous signals are used in the two species, i.e., when they do not derive from the same signal in the common ancestor. This may occur via multiple polyadenylation signals at some point during evolution.

5.1.1. Caenorhabditis briggsae

C. briggsae is another soil nematode, whose sequence was published in 2003 (Stein et al. 2003). It is thought that *C. briggsae* and *C. elegans* diverged from a common ancestor roughly 100 million years ago.

The neutral substitution rate measured at non-synonymous sites is estimated to be about 1.75, which is three times the distance between human and mouse. The two worms, which are similar at the level of ecology and morphology, show extensive identity at the level of genome organisation. The difference in size between the two genomes is accounted for almost entirely by repeat regions. Of the c. 19,500 predicted

protein coding genes in *C. briggsae,* about 62% have strong one-to-one orthologues in *C. elegans*. The availability of a large orthologous gene set allows us to study how 3' UTRs and in particular, polyadenylation signals change during evolution.

5.1.2. C. elegans – C. briggsae orthologues

The set of orthologous *elegans-briggsae* pairs on which all the analyses in this chapter are based come from a hybrid reciprocal best 1:1 BLASTP hit and synteny analysis (Stein et al. 2003). 12155 pairs exist at the protein level. For each *C. elegans* gene with a pair, an orthologous 3' UTR pair was made by extracting the final coding exon of the *briggsae* orthologue, checking that the gene prediction ended at a stop codon (which was not the case for 3254 genes), and extending from the stop codon the same length as the *elegans* non-overlapping 3' UTR candidate, as discussed in chapter 3 (1000nt or up to the next gene). This leaves us with 8901 orthologous 3' UTR pairs. Polyadenylation signals were predicted on all sequences using the *C. elegans* PAjHMMA model looking for all signals with a posterior probability greater than 0.1. Viterbi predictions were also carried out.

5.2. Conservation of absolute position

5.2.1. Introduction

We first examined how 3' UTR length correlates between orthologous pairs. For the purposes of this experiment, we define the length of the 3' UTR as the distance from the stop codon to the start of the AATAAA motif.

5.2.2. Results

Figure 25 shows the weak correlation (r=0.45) between absolute positions of orthologous Viterbi polyadenylation signal predictions. The distribution of signal positions in both species is very close to the observed length distribution of 3' UTRs, and thus the vast majority of the data falls into the bottom-left quadrant. This in itself shows the relative specificity of the prediction method.

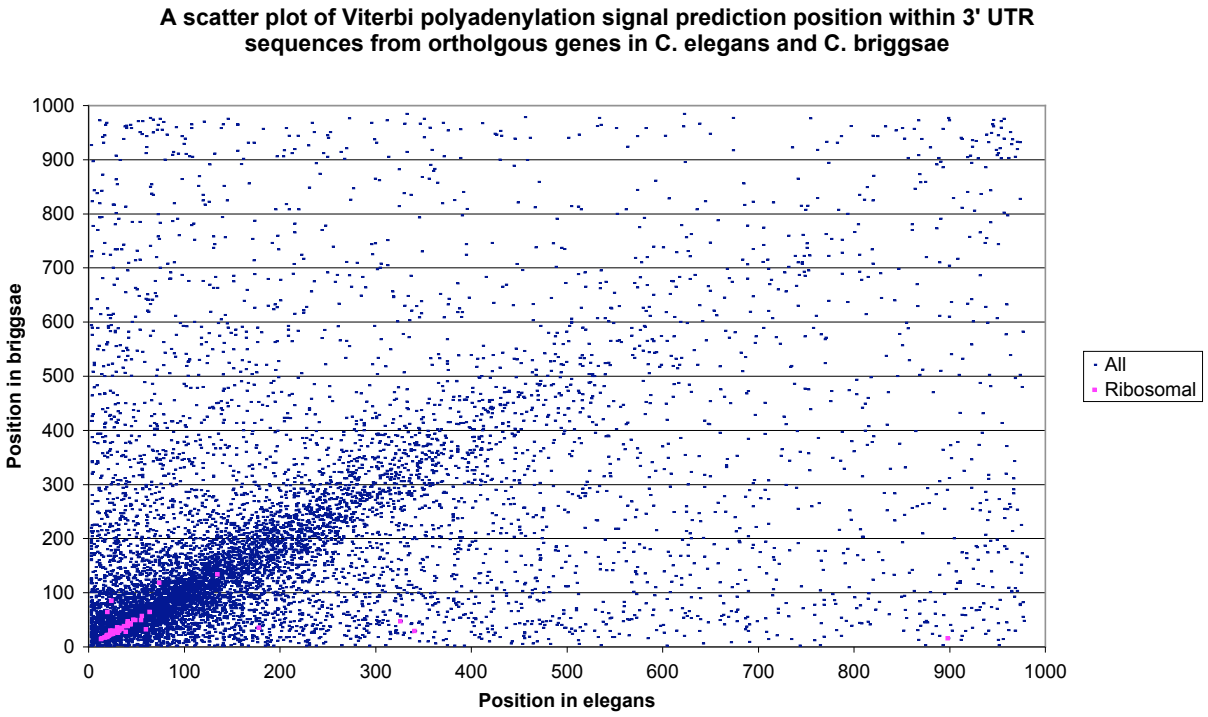

Figure 25. A scatter plot showing the absolute positions of Viterbi polyadenylation signal predictions within the 3' UTRs of 8901 pairs of orthologous elegans and briggsae genes. The 65 ribosomal protein 3'UTRs in the orthologous pair set are shown in pink.

3' UTR lengths between orthologous genes are rarely conserved exactly, on account of the sequences readily accepting indels and often containing repeat regions (Jareborg et al. 1999; Larizza et al. 2002). Of the 8901 pairs plotted, 120 are found to have perfectly conserved 3' UTR lengths.

One striking observation is that 24 of the 120 paired 3' UTRs with conserved length are from ribosomal protein mRNAs, out of 65 ribosomal genes that are included in our set of pairs. The proportion of non-ribosomal 3' UTRs that are under 100 nt and have an orthologous pair of the same length is 2%. Hence, the 24 ribosomal 3' UTRs appearing with pair having the same length represents a significant overrepresentation. It is likely that regulatory conservation has restricted mutation in the 3' UTR of these genes. As we shall discover in chapter 6, there is a putative conserved regulatory motif, which spans the polyadenylation signal of ribosomal protein genes and is also found in other genes implicated in translation. However, analysis of the non-ribosomal genes having conserved 3' UTR lengths did not reveal any functional bias.

5.3. Polyadenylation signals in aligned orthologues

5.3.1. Introduction

As mentioned in chapter 3, we have 6570 *C. elegans* 3' UTRs, in which we have high confidence, on account of there being EST evidence for the predicted polyadenylation signal. Using this high confidence set, we can look at cases where orthologous pairs of worm 3' UTRs can be aligned by BLAST such that the *C. elegans* polyadenylation signal is within the alignment. In these cases, we are

interested in seeing whether the corresponding position in *C. briggsae* is also a likely polyadenylation signal, and if so, whether some signal variants are conserved at a higher rate than others. For example it might be that genes having the AATGAA variant do so for a specific reason, and perhaps are less likely to allow mutations which, although not knocking out the function of the signal, would change which hexamer is used.

There are also cases when the 3' UTRs of orthologous genes do align, but have polyadenylation signals that are in different parts of the alignment. This may give some insight into signal gain and loss over evolution.

5.3.2. Alignment

3400 of the 6570 *C. elegans* high confidence sequences had orthologous *C. briggsae* predictions. Each of these 3400 paired sequences were BLASTed (W=3, E> 0.01, --top) against each other to find regions of sequence homology. 1840 of these pairs contained a BLAST alignment. There were 545 cases in which orthologous pairs had signal predictions, but neither of them fell in the alignment. There were 1238 cases where the *elegans* Viterbi polyadenylation signal was contained in the alignment. Of these, 1052 had one *C. briggsae* signal (determined by Viterbi) also in the alignment.

5.3.3. Results – aligned Viterbi predictions

5.3.3.1. Position of aligned Viterbi signals

The BLAST output from this alignment is shown in Figure 26. Here we can see how the GATAAA in *C. elegans* is aligned with the AATAAA in *C. briggsae*. A graphical representation of this case is shown in Figure 27. For the set of 1052 pairs where the AATAAA motifs from both species were contained in the alignment, 409 (39%) are in corresponding positions as shown in Figure 26, and 643 (61%) are in non-homologous positions. There is no preference for any offset in the alignment if the signals themselves are not aligned.

```
 Score = 302 (51.4 bits), Expect = 6.5e-07, P = 6.5e-07
 Identities = 104/139 (74%), Positives = 104/139 (74%), Strand = Plus / Plus

Query:    59 ATCATAATTATCCAACTGCCTCTAAAGCTCTTCGAAAACAAATTCCATTCTATTTTTTGT 118
             |||||||||||| |||||||||| || ||| ||||||| |||||    | || || ||||
Sbjct:    32 ATCATAATTATTCAACTGCCTCTTAAACTCCTCGAAAATAAATTG--T-CTCTTCTTTGA 88

Query:   119 TTTAAACAACTCATATTTTGCGAGCGATCTCTTTGCTTTT----CTTGTGTATGATAAAT 174
                || |   ||||||| |||   ||||||||||||||     | ||||||  ||||||
Sbjct:    89 ---AATCT--TCATATTCTGCTTTCGATCTCTTTGCTTTTGAAACATGTGTAAATAAAT 143

Query:   175 AATTTATTTTA-AATTAAG 192
              |||||||||  ||  |||
Sbjct:   144 TATTTATTTTTCAAACAAG 162
```

Figure 26. A BLAST alignment showing aligned polyadenylation signals. AATAAA motif is in red.

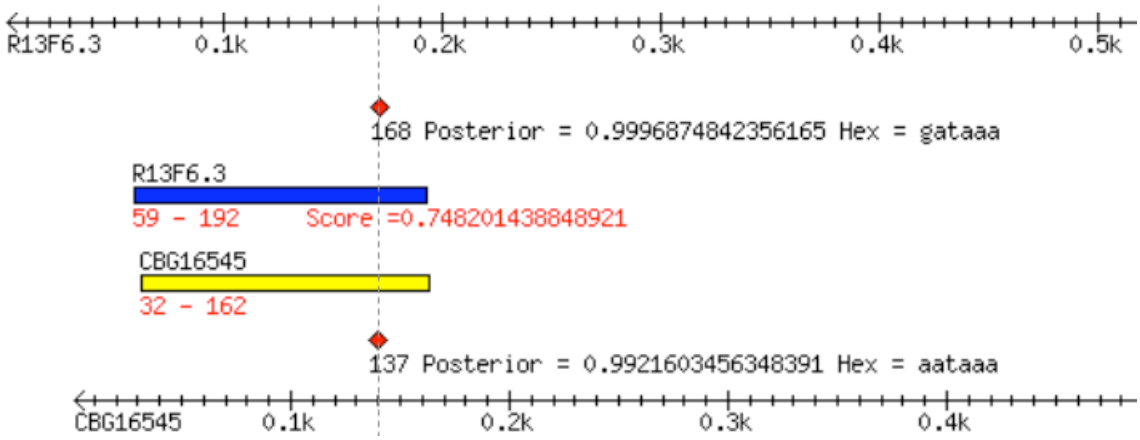

Figure 27. A representation of the sequence of a pair of orthologous 3' UTRs. From the top, the scale bar, red diamond, and blue bar refer to the C. elegans sequence, whilst the yellow bar, red diamond, and scale bar refer to C. briggsae sequence. The positions and coordinates on the blue and yellow bars show details of the BLAST alignment relative to elegans and briggsae respectively. The polyadenylation signals indicated by diamonds are in corresponding positions in the alignment.

5.3.3.2. Species distribution of aligned signals

If we isolate the ~40% of gene pairs showing aligned polyadenylation signals, we can see whether different AATAAA motif variants are freely interchangeable between orthologous genes, or whether genes tend to conserve them. First, though, it is necessary to ascertain whether there is any difference in the frequency distribution of AATAAA motifs in the two species. Table 8 shows the overall distributions of signals in *C. elegans* and *C. briggsae*. These distributions are not significantly different.

Table 8. A table of hexanucleotide frequencies of aligned polyadenylation signals of orthologous worm genes.

	Elegans		Briggsae	
	Count	Freq	Count	Freq
AATAAA	239	0.583	246	0.600
AATGAA	48	0.117	55	0.134
CATAAA	34	0.083	34	0.083
TATAAA	29	0.071	24	0.059
GATAAA	24	0.059	20	0.049
CATGAA	9	0.022	8	0.020
TATGAA	8	0.020	8	0.020
GATGAA	5	0.012	3	0.007
ATTAAA	4	0.010	1	0.002
AGTAAA	3	0.007	3	0.007
GATGGA	1	0.002	0	0.000
CGTAAA	1	0.002	2	0.005
ACTAAA	1	0.002	0	0.000
AATACA	1	0.002	0	0.000
AACGAA	1	0.002	0	0.000
AAAAAA	1	0.002	1	0.002
AATAAT	0	0.000	2	0.005
TTTGAA	0	0.000	1	0.002
TATACA	0	0.000	1	0.002

As the two species have apparently similar distributions of AATAAA motifs, an analysis of whether they are conserved between orthologous genes will not be skewed by a genome-wide flux away from or towards particular signals.

5.3.3.3. Pattern of hexanucleotide mutation in aligned signals

In Figure 26 we see how the polyadenylation signal from two orthologous genes differ in which AATAAA motif is used. Figure 28 is a full graphical representation of the AATAAA motif transitions observed in 409 pairs of orthologous *elegans* and *briggsae* aligned polyadenylation signals. It shows that on average, 74% of pairs conserve the particular variant of AATAAA being used.

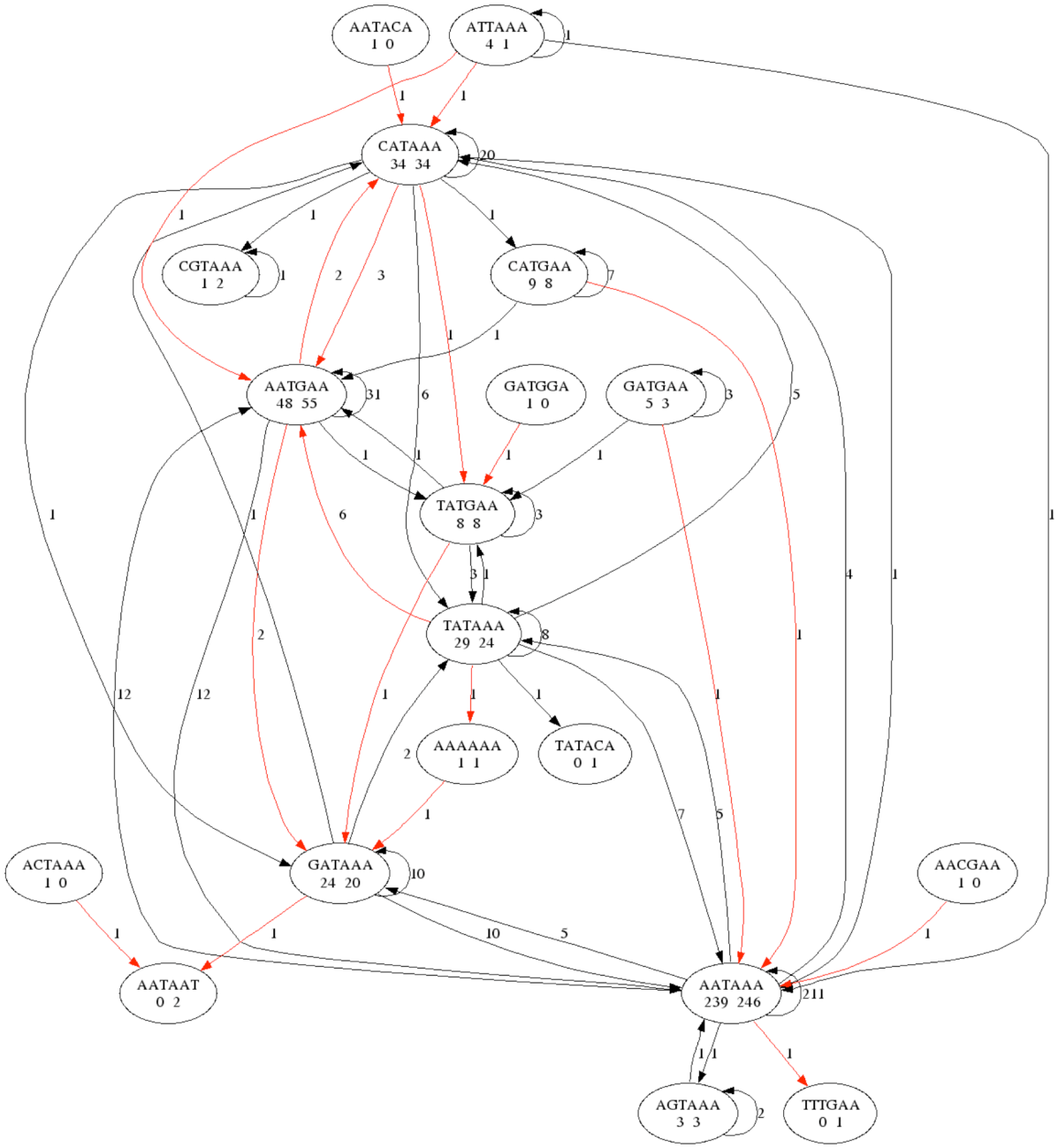

Figure 28. A graph showing the mutations between aligned AATAAA motifs in 411 orthologous worm gene pairs. Nodes represent a particular AATAAA motif. The number of signals appearing in C. elegans and C. briggsae are shown below the AATAAA motif on the left and right respectively. The number of C. elegans genes with AATAAA aligning to a GATAAA in C. briggsae is 5. Of the 20 genes in C. briggsae having a GATAAA, 10 have C. elegans orthologues where the polyadenylation signal is an AATAAA. Red arrows denote AATAAA motif transitions involving the mutation of more than one base pair, such as AATGAA->TATAAA.

This raises an interesting point regarding conservation of AATAAA motif. It is obvious that there is some flux in this system; that is, there are mutations between aligned orthologous polyadenylation signals. However, there are many nodes where a large proportion of self-cycling occurs. Figure 29 shows the how the proportion of a particular AATAAA motif variant that is conserved between species varies with that motif's frequency of occurrence.

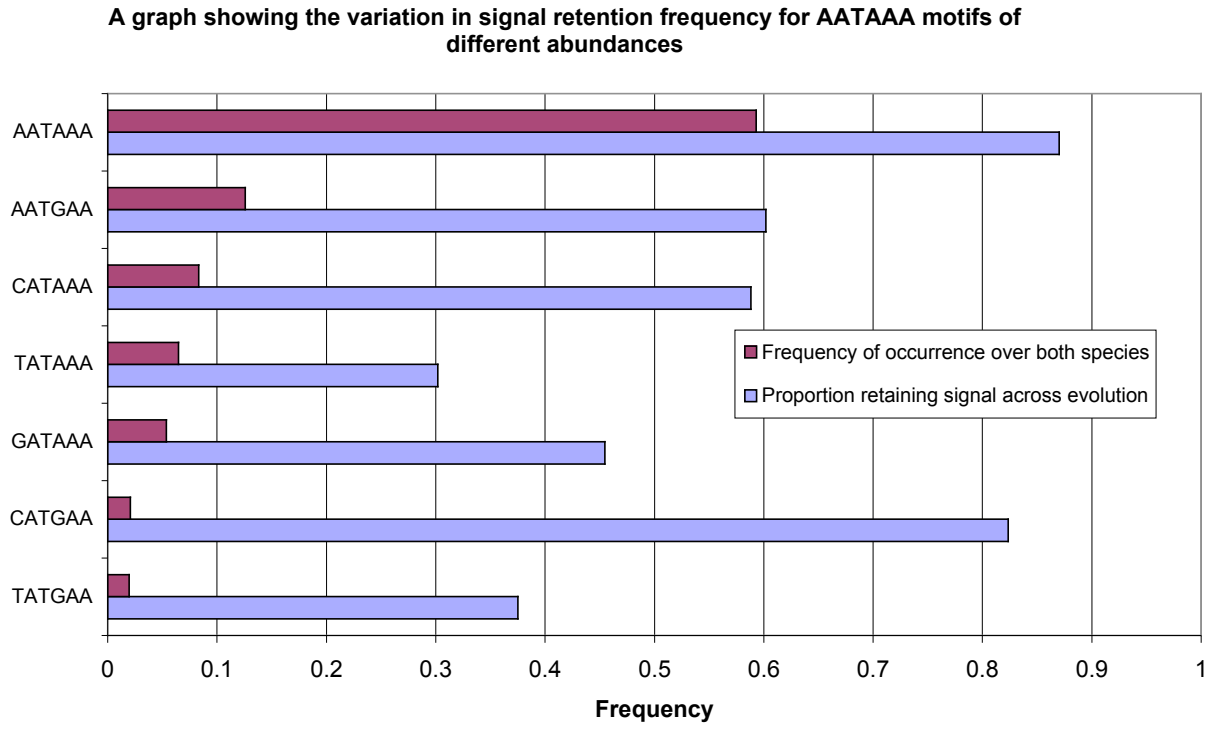

Figure 29. A graph showing the proportion of each AATAAA motif in both worms, and the proportion of each aligned signal being conserved between C. elegans and C. briggsae orthologues.

For the four most commonly used signals, the proportion of orthologous genes conserving that signal decreases with the abundance of the motif variant. 59% of the genes in the orthologous pair set use AATAAA as the polyadenylation signal. 89% of

these genes' orthologues also use AATAAA. If we look instead at the 6.5 % of genes that use a TATAAA, the proportion conserving this across evolution drops to 30%. Again, just considering the top four signals, it seems that the more commonly used the motif, the more likely it is to be conserved between species. It could be that some signals, such as AATAAA, are required for some genes, whereas any variant might be tolerable for others.

For the other three signals, although the proportion retaining values may be distorted by small sample size, we notice that 7 of the 9 recorded CATGAA in *C. elegans* are retained as CATGAA in *C. briggsae*. This represents a much larger proportion of retention than, say GATAAA, where fewer than half (10 out of 24) of the *elegans* genes retain this signal. It could be that certain genes' signals are constrained between orthologues for functional reasons. However, no distinctive functional characteristics were found empirically by looking at the functions of the 7 genes conserving CATGAA.

There are also interesting patterns of mutation, if we consider those aligned signals which differ between the two worms. For example, Figure 28 shows that of the 24 *C. elegans* genes with GATAAA, 14 mutate, of which 10 mutate to AATAAA in *C. briggsae*. However, of the 34 genes with CATAAA, again 14 mutate, but only one changes to AATAAA. The majority of those changing mutate to TATAAA instead. These observations can perhaps be explained by the difference in rates of transition vs. transversion, but this is inconsistent with the observation that of the 21 genes mutating away from TATAAA, 5 mutate to CATAAA (transition) versus 7 mutating to AATAAA (transversion). It seems in these circumstances that a transversion event, which should be rarer than transition, is favoured as it introduces a more commonly used AATAAA motif.

The second most commonly used AATAAA motif in the worms is AATGAA. As we see from the red lines in Figure 28, many of which involve AATGAA, there are several cases where aligned signals have two mutations between species. Although this should be a relatively rare event, there are similar numbers of changes from TATAAA to AATGAA, CATAAA, and AATAAA. We have been unable to find an explanation for this behaviour, which appears not to show the expected mutation parameters favouring the A/T-richness of the *Caenorhabditis* genomes nor the expected proportion of transition to transversion. Perhaps a larger set of aligned orthologous polyadenylation signals with mutations might show that the weights on the mutation graph split the AATAAA motifs into cliques, where mutations within cliques are more favoured than those between cliques.

5.3.4. Evolutionary turnover of polyadenylation signals

We have mentioned earlier that of the orthologous gene pairs in which the 3' UTRs can be aligned across both species Viterbi polyadenylation signals, 61% of orthologous gene pairs have the signal predictions in non-corresponding positions. This suggests a relatively high level of turnover of polyadenylation signals. We wish to explain possible ways in which this can happen. One way to imagine how a new cleavage and polyadenylation site evolves is via an intermediate state in which both sites are active. On divergence, we might expect this to leave a trace of another potential site at the aligning position.

Figure 30 shows an example in which Viterbi polyadenylation signals do not align between species.

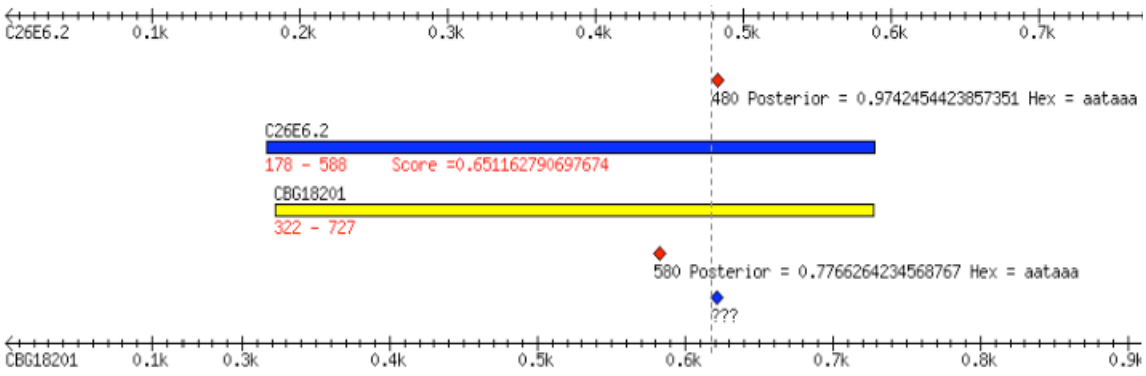

Figure 30. An alignment of two orthologous 3' UTR sequences. The extent of C. elegans and C. briggsae sequence aligned is shown by the blue and yellow bars respectively. Viterbi polyadenylation signal predictions are shown as red diamonds. The blue diamond is that part of the C. briggsae sequence, which aligns to the C. elegans polyadenylation signal.

For 642 such aligned pairs, we analysed the *C. briggsae* sequences aligned to the *C. elegans* Viterbi prediction. Of the whole set, there are only 31 occurrences where the *C. briggsae* hexamer sequence, represented by the blue diamond in Figure 30, is exactly the same as the *C. elegans* AATAAA motif. A further 20 mutate to one of the other top 8 occurring AATAAA motifs. Hence, 51 *C. briggsae* genes have a sequence resembling an AATAAA motif in the same place in the alignment as the *C. elegans* Viterbi hit. Only 24 (8% of the 642) of these *C. briggsae* hits have posterior probabilities of greater than 10%.

Similarly, when comparing *C. elegans* posterior decoding ($p > 0.1$) predictions to *C. briggsae* Viterbi predictions in these 642 misaligning pairs, there are 50 cases (26 identical, 24 variant) of a *briggsae* Viterbi prediction aligning to a *C. elegans* sequence that resembles an AATAAA motif. 12 of these similar motifs are *C. elegans* posterior decoding hits. This number may be less than the 24 observed when analysing *C. briggsae* posterior decoding on account of our earlier selection against *C. elegans* genes with multiple polyadenylation sites (Chapter 3). These observations

show that about half of all orthologous 3' UTRs that contain an alignment have polyadenylation signals that are either aligned to an orthologous signal (409 gene pairs), or to a sequence that may well be a real polyadenylation signal (a further 102 gene pairs, of which 36 had posterior decoding support). For the remaining 511 out of 1052 aligned 3' UTR pairs, there seems to be no sign of signal position conservation, even within the context of an alignment. These sequences have diverged so far, that meaningful polyadenylation signals have been lost.

5.4. Discussion – On the evolution of polyadenylation signals

We have shown here that weakly constrained mutation of 3' UTRs mean orthologous genes only weakly conserve the absolute position of polyadenylation signals. Analysis of those cases where 3' UTRs can be aligned and where polyadenylation signal predictions align with each other show an unusual pattern of substitution. This seems not to match what might be expected from a simple model of nucleotide mutation, and there may be functional constraints as to the choice of AATAAA motif that is required for a particular gene. Within alignments, if the most likely polyadenylation signals themselves do not align, we investigated whether there are signs that a given gene from the common ancestor to *C. elegans* and *C. briggsae* may have had two equal polyadenylation signals, with the two different species favouring different signals following the evolutionary split. However, in only one sixth of these cases could we see a residual aligned site with posterior probability >10%.

The study of evolution of regulatory regions has been made possible by comparative studies on recently sequenced genomes, and has focussed mainly on

enhancers and promoters, such as (Dermitzakis et al. 2003). Whilst it is expected that sequence with regulatory function should be conserved beyond the background of non-functional sequence, it is not understood how selection operates on regulatory regions, and it is surprising that the turnover of polyadenylation signals is so high. Although there is scant previous work on evolution of polyadenylation signals in particular, there have been studies on evolutionary dynamics of *cis*-regulatory regions (Johnson et al. 2004; Ludwig et al. 2005).

Of particular relevance to this study is the paper by (Ludwig et al. 2000), which concerns the enhancer element of *even-skipped* mRNA in *Drosophila melanogaster* and related flies. They show that the elements occurring in *D. melanogaster* and *D. pseudoobscura* can be aligned (504 nt in *melanogaster* vs. 691nt in *pseudoobscura)*, though the enhancer differs in certain places between the two species. Constructs containing the whole element from each of the flies give identical patterns of gene expression in the reporter system, despite the differences. However, splitting the enhancer in half, and building two chimaeric constructs, each containing either the first half from one species, and the second from the other both give a mutant phenotype. Crucially, the two mutant phenotypes are not identical. It is proposed that stabilising selection is maintaining phenotypic identity in the region, but has allowed mutational turnover of important regulatory sites.

Although the group do not mention whether evolution has left any trace of an ancestral site in either species (as was the case for some 17% of the polyadenylation signals in aligned regions), the work sets a precedent that fast turnover of regulatory motifs in the context of an aligned background can be expected between species.

6. Concerning a Sequence Element Detected in Ribosomal mRNAs

6.1. Introduction

Until now, this thesis has focussed on the identification of the polyadenylation signal and the end of the 3' UTR. In this chapter, we change focus to look for other conserved signals within the 3' UTR. In particular, we identify a region around the polyadenylation signal in many ribosomal protein mRNAs in *C. elegans* and *C. briggsae* that contains a conserved sequence motif. Building a statistical model of this motif and searching a database of *C. elegans* 3' UTRs reveals that this motif is also present in the 3' UTR of some other genes involved in ribosome maturation and translation.

6.2. Background

An initial approach that we took to identifying 3' UTR regulatory elements was to look for conserved secondary structure components in *C. elegans* and *C. briggsae*. We took the 3' UTRs from about 9000 *C. elegans* genes that were confirmed by ESTs and aligned them to the same length of sequence downstream of the STOP codon of the *C. briggsae* one-to-one orthologue. 6000 of these pairs generated a BLAST alignment according to our alignment parameters. The BLAST alignments were then submitted to QRNA (Rivas et al. 2001), to see if the mutations between a pair of aligned 'orthologous 3'UTRs' were co-varying; that is, to discover

whether the sequences were evolving in such a way as to conserve a potential RNA secondary structure in an area of relatively lower primary sequence conservation.

125 of these aligned orthologous 3' UTR pairs were considered by QRNA to contain conserved secondary structures. Of these 125, 14 alignments were from the 3' UTRs of ribosomal proteins, an example of which is shown in Figure 31. This represents a significant overrepresentation. Further examination of the secondary structure alignments of these UTRs showed that it was unlikely that there was a single secondary structure element common to our set of ribosomal 3' UTRs. Closer observation of the alignments suggested that in this case, there might be a conserved primary sequence which had some potential to fold into a secondary structure, though the hairpin structure itself was not being specifically conserved. Additionally, building each aligned pair into a covariance model (Eddy 2002) and searching nucleotide databanks did not indicate the presence of different, functionally conserved secondary structures.

```
#-------------------------------------------------------------------------------
#      qrna 2.0.2 (Wed Nov 19 15:00:55 CST 2003) using squid 1.5m (Sept 1997)
#-------------------------------------------------------------------------------
#      PAM model =  BLOSUM62
#-------------------------------------------------------------------------------
#      RNA model       =  /mix_tied_linux.cfg
#      RIBOPROB matrix =  /RIBOPROB85-60.mat
#-------------------------------------------------------------------------------
#      seq file  =  /acari/work5a/ah3/Ribosomal/Blastn2qrna/18.bl.q
#                   #seqs: 2 (max_len = 35)
#-------------------------------------------------------------------------------
#      full length version:  -- length range = [0,1000]
#-------------------------------------------------------------------------------
# 1  [given strand]
>F40F8.10-1>35- (35)
>CBG02962-1>35- (35)

Divergence time (variable): 0.030409
[alignment ID = 91.43 MUT = 8.57 GAP = 0.00]

length alignment: 35 (id=91.43) (mut=8.57) (gap=0.00)
posX: 0-34 [0-34](35) -- (0.34 0.06 0.17 0.43)
posY: 0-34 [0-34](35) -- (0.37 0.09 0.14 0.40)


                    SS  --------111111111.........111111111
                    SS  -------->>>>>>>>>.........<<<<<<<<<
         F40F8.10-1>35-  AUUUGUUUUGGUUACAAAUAAAAUUGUUGGACAUA
         CBG02962-1>35-  ACUUGUUUUUGUUACAAAUAAAAUUGUUGGACAAA

LOCAL_DIAG_VITERBI -- [Inside SCFG]
OTH ends *(+) =  (0..[35]..34)
COD ends *(+) =  (14..[21]..34)
RNA ends *(+) =  (8..[27]..34)
winner = RNA
              OTH =       84.411             COD =       77.856             RNA =       84.962
   logoddspostOTH =        0.000  logoddspostCOD =       -6.555  logoddspostRNA =        0.551
     sigmoidalOTH =       -0.562    sigmoidalCOD =       -7.857    sigmoidalRNA =        0.536
```

Figure 31. Example output from QRNA when run on a 3' UTR alignment between a C. elegans ribosomal protein gene and its C. briggsae orthologue. The polyadenylation signals for the two genes are shown in red. One co-variant position is seen within the predicted secondary structure.

6.2.1. Polyadenylation Signals

The area of sequence conservation was consistently situated around the polyadenylation signal as detected in our previous study on *C. elegans* polyadenylation signals (Chapter 3).

One approach to finding an unknown, but overrepresented motif or area of homology is to use expectation maximisation. Submitting *C. elegans* and *C. briggsae* ribosomal mRNA 3' UTR sequences to the MEME program (Bailey et al. 1994) both discovered similar motifs, again based around the polyadenylation signal. Figure

32(a) shows the motif found by MEME by submitting 68 *C. elegans* ribosomal protein 3' UTRs. The AATAAA in the centre represents a real polyadenylation signal. Figure 32(b) shows the same for 68 *C. briggsae* 3' UTRs, whose genes are the best one-to-one orthologues of the 68 *C. elegans* genes. In contrast Figure 32(c) shows the expected base composition about 940 experimentally confirmed *C. elegans* polyadenylation signals.

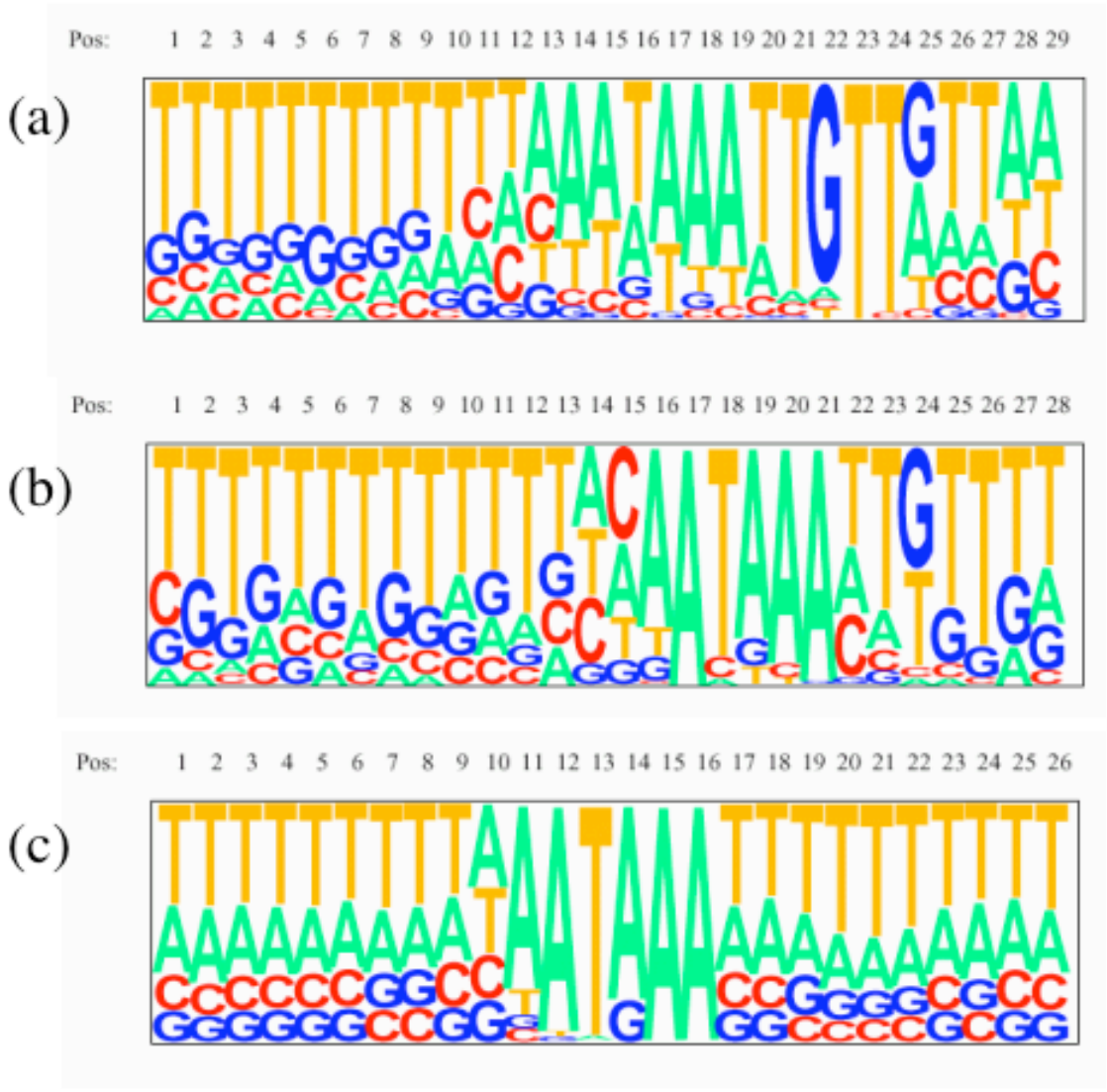

Figure 32. The nucleotide distribution observed in the region around the polyadenylation signal in (a) 68 C. elegans ribosomal protein genes, (b) 68 C. briggsae one-to-one orthologues of the genes in (a), and (c) 940 experimentally confirmed polyadenylation signals from C. elegans.

By observation, the sequence directly after the AATAAA motif appears different in the ribosomal mRNAs, with consensus TTGTT. The ribosomal sequences also appear to show higher than typical levels of G bases upstream of the signal, and indeed many have TTGTT, but at variable distances upstream, so the pattern is not visible in a simple alignment. We therefore conjecture that TTGTT sequences in the near neighbourhood of the polyadenylation signal may be important for ribosomal genes. We therefore decided to analyse a large set of aligned ribosomal protein 3' UTRs, anchored on the polyadenylation signal.

6.3. Model building

6.3.1. Data acquisition

One kilobase sequences representing possible 3' UTRs from 84 ribosomal proteins were extracted from WormBase (http://www.wormbase.org/). 68 of these had putative one-to-one orthologues in *C. briggsae*. Polyadenylation signal predictions (Chapter 3) were run on each sequence, and an alignment of the signal and the 20 nt flanking it on each side was forced by anchoring on the polyadenylation signal. There were 136 sequences in the alignment. The Jalview alignment viewer (Clamp et al. 2004) was used to hand-edit the alignment (Figure 33) so that TTGTT motifs either side of the polyadenylation signal were aligned. Any sequences without TTGTT in a position where it could fit in the alignment were removed. Most sequences had at least one TTGTT, but not on both sides. Some contained TTATT instead. This strict removal process left 57 sequences.

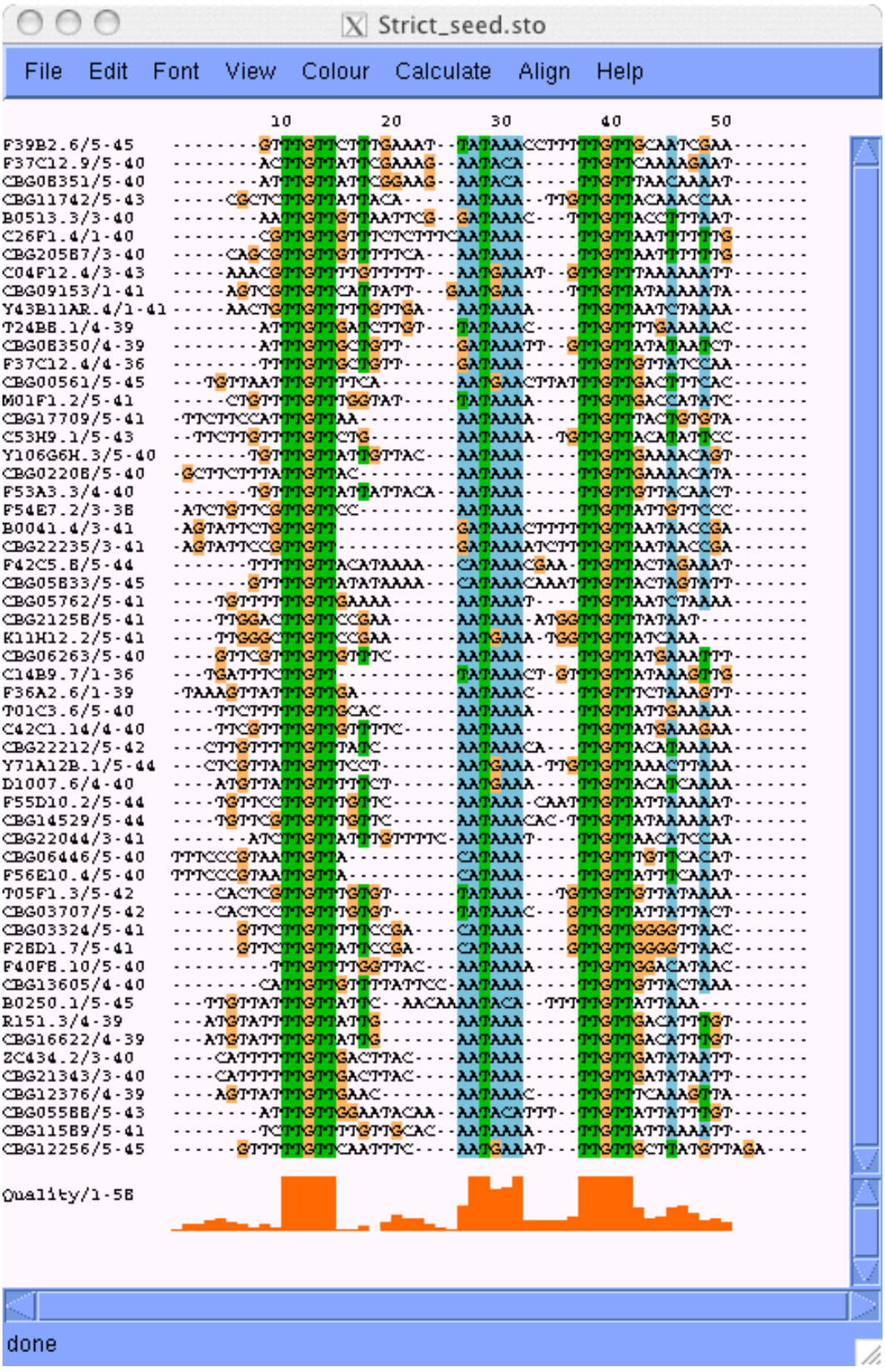

Figure 33. A hand edited alignment of the region around polyadenylation signal predictions from 57 C. elegans and C. briggsae ribosomal protein mRNAs.

6.3.2. Model building with HMMER

A motif is conveniently modelled by a hidden Markov model, as it represents the region as a network of interconnected states, each with characteristic nucleotide frequencies. Variable insertion probabilities can model different motif spacings. The alignment in Figure 33 was built automatically into a hidden Markov model (Figure 34), which can capture sequence motif profiles using HMMER (http://hmmer.wustl.edu). This model was used to search a set of 22156 3' UTR candidates from *C. elegans* (that is, the 1000 bases 3' of each predicted gene's STOP codon, or the longest length up to 1000 nt before overlapping into the 3' gene.) Hits above 20 bits were reported. This generated 470 hits, of which 300 flanked a predicted polyadenylation signal. These 300 hits could be split into two groups of 150, the first containing an exact TTGTT…PolyA_Signal…TTGTT, with the other set containing at least one mismatch to one or more of the TTGTT motifs.

```
hmmb: hidden Markov model construction from alignment
     version 1.8.3, June 1997
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Training alignment:               strict.slx
Number of sequences:              56
Model output to:                  hmm-strict
Model construction strategy:      Max likelihood
Prior strategy:                   simple Dirichlet
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Constructed a hidden Markov model (length 35)
Average score:                    24.93 bits
Minimum score:                    13.61 bits
Maximum score:                    32.61 bits
Std. deviation:                    4.66 bits
Information content:              20.51 bits

HMM written to file hmm-strict
```

Figure 34. Summary of the HMM built from the alignment of 57 ribosomal mRNA polyadenylation signals.

Both sets contain hits to non-training set ribosomal protein genes, along with other genes. However, there are two potential disadvantages to this method. One is that hits containing non-canonical AATAAA polyadenylation signals are penalised, as the signal forms part of the overall motif pattern. (Figure 35) shows HMMLS hits on two sequences, which are identical apart from seq1 containing AATAAA, and seq2 TATAAA. The seq2 score is 2.5 bits lower than the seq1 score, and using a cutoff of 20 bits, seq2, which comes from WormBase CDS F39B2.6 (40S ribosomal protein S26), would be missed as a false negative.

```
hmmls - search long sequences for local matches to a hidden Markov model
     version 1.8, February 1995

 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
HMM file:                 hmm-strict
Sequence database:        seqs.dna
Report scores above:      0.00
Scan window size:         1000
Do complementary strand:  no
Fancy alignment output:   yes
[Printing multiple non-overlapping hits per sequence]
 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

20.58  (bits) f:    1 t:   41 Target: seq1
  Alignment to HMM consensus:
                  *ttttttttgttattt.....aataaa.....ttgttaataaaaat*
                        TTTGTT TTT     AATAAA     TTGTT+ +A+  A+
        seq1     1  ----GTTTGTTCTTTGAAATAATAAACCTTTTTGTTGCAATCGAA     41


18.07  (bits) f:    1 t:   41 Target: seq2
  Alignment to HMM consensus:
                  *ttttttttgttattt.....aataaa.....ttgttaataaaaat*
                        TTTGTT TTT      ATAAA     TTGTT+ +A+  A+
        seq2     1  ----GTTTGTTCTTTGAAATTATAAACCTTTTTGTTGCAATCGAA     41
```

Figure 35. Output from HMMLS searching two sequences for hits to the HMM constructed from an alignment of ribosomal mRNA polyadenylation signals

The other problem is that the separation of the TTGTT to the polyadenylation signal has a distinctive length distribution, as does the separation on the 3' side of the signal. HMMER does not model these two different length distributions explicitly, but rather allows hits to contain gap symbols with a penalty score, corresponding to a negative exponential distribution of gap length. The observed gap length distributions in our alignment are more flat upstream of the signal, and have a definite length preference downstream.

6.3.3. A more specific model

Both of the problems described above can be solved by incorporating the ribosomal motif information into a PAjHMMA model for the whole region. This 'ribosomal' model can be compared to our standard 'background' polyadenylation model to find cases which closest resemble how the TTGTT motifs flank the polyadenylation signal in the ribosomal protein mRNAs. The benefits of using PAjHMMA are that the models are polyadenylation signal-aware, unlike HMMER, and can explicitly model the observed separation between TTGTT and AATAAA motifs.

The ribosomal polyadenylation signal PAjHMMA model (Figure 36) is derived from the standard polyadenylation signal model. There are 12 additional states. TTGTT motifs (each with a state for each of the 5 columns) are inserted either side of the AATAAA motif states. The separations (from the AATAAA motif) between the upstream and downstream TTGTT motifs, are each modelled with distinctive lengths, corresponding to two more states, U and D. The ribosomal model

forces each sequence to pass through both TTGTT motifs, though the two separator states can be bypassed with a probability reflecting the occurrences of upstream or downstream separator length being zero. The TTGTT motifs themselves are built empirically, scoring 1/100 for a mismatch and 97/100 for a match. In the third column, the occurrence of A is penalised to a slightly lesser degree than the others, scoring 5/100.

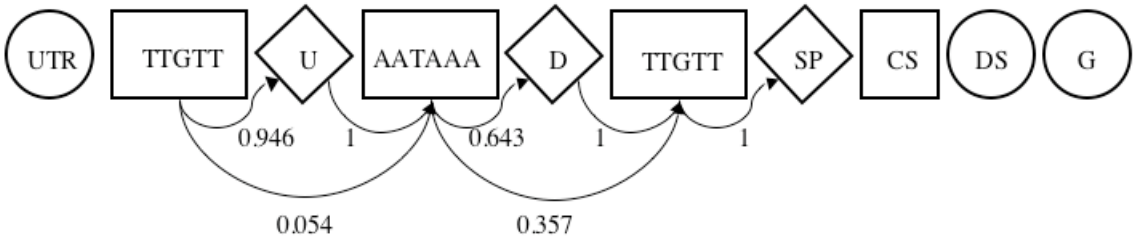

Figure 36. State transition diagram for ribosomal polyadenylation signal model. Circular states have geometric length distributions, boxes represent weight matrices, and diamond states have explicitly modelled lengths. Where transition probabilities are not given, they are set to the same values as in the standard model given in Chapter 3.

As discussed in Chapter 2, one of the by-products of the forward and backward algorithms is *P(x)*, the probability of the sequence given the model, or the probability that the sequence was generated by the given model. For any given sequence, we find this value given the extended ribosomal polyadenylation signal model, and the standard *C. elegans* polyadenylation signal model. The difference in the logarithms of the probabilities is a bit score measuring how well the sequence fits the ribosomal model relative to the background.

The observed length distributions upstream (Table 9) and downstream (Table 10) of the AATAAA motif are given below.

Table 9. The length distribution observed between the upstream TTGTT motif and the polyadenylation signal from 57 ribosomal protein mRNA sequences.

Length i	u(i)
0	0.054
1	0.071
2	0.125
3	0.071
4	0.143
5	0.125
6	0.107
7	0.089
8	0.107
9	0.107

Table 10. The length distribution observed between the polyadenylation signal and the downstream TTGTT motif from 57 ribosomal protein mRNA sequences.

Length i	d(i)
0	0.357
1	0.285
2	0.089
3	0.071
4	0.107
5	0.089

6.4. Model testing

To test whether the ribosomal model is able to differentiate ribosomal protein 3' UTRs, bit scores relative to the standard model were found for sequences from four different sets. The four sets were:

(1) Predictions made over 22069 *C. elegans* non-ribosomal protein 3' UTRs.

(2) Predictions from 54 *C. elegans* ribosomal sequences, that were not included in model training.

(3) Predictions made on 104 sequences of 3' UTRs from *C. elegans*. The proteins of these genes represent the best BLASTP hit for 165 proteins from *S. cerevisiae*, that are implicated in pre-ribosomal complex formation in yeast (Fromont-Racine et al. 2003), but the set includes few ribosomal proteins.

(4) Predictions made on 63 *C. briggsae* orthologues of the 100 genes from set (1) that had the highest bit score under the ribosomal model.

6.5. Results

Figure 37 shows that the distributions of score for ribosomal and non-ribosomal proteins do appear to be different. The peaks in the 0 and 5 bit regions are caused by single and double mismatches respectively to TTGTT, either upstream or downstream of the polyadenylation signal. The *C. elegans* orthologues of the yeast proteins involved in ribosome assembly have a similar score distribution as the non-

ribosomal protein set, but does contain a 'shoulder' of higher bit scores. It could be that the motif confers some function or fate involving ribosomal protein mRNAs that is distinct from ribosome assembly, and a subset of the ribosomal assembly complex have strong matches to the motif.

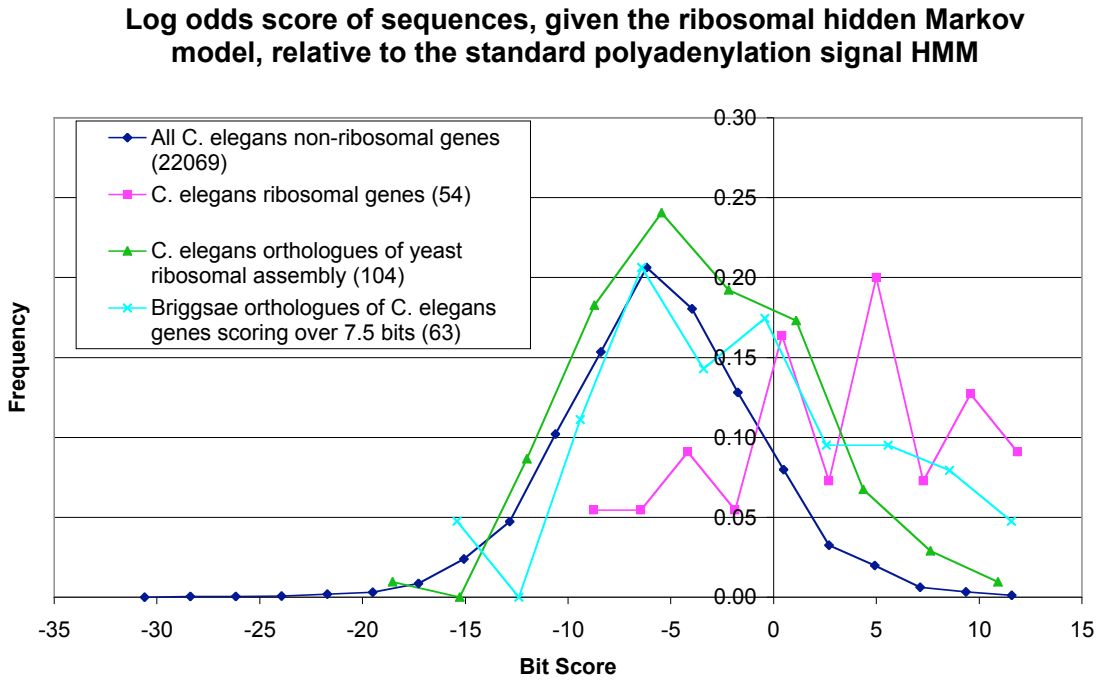

Figure 37. The bit score histogram resulting from finding the log(2) probability of various 3' UTR sequence sets under the ribosomal model minus the log(2) probability under the standard model. Dark blue: All C. elegans non-ribosomal protein genes - this is the background distribution. Pink: C. elegans ribosomal protein genes. Green: C. elegans orthologues of yeast ribosomal assembly complex. Light blue: C. briggsae orthologues of C. elegans genes scoring over 7.5 bits.

One hundred *C. elegans* non-ribosomal protein 3'UTRs have a bit score greater than 7.5 bits. Looking at the 63 *C. briggsae* orthologues of these high scoring *C.elegans* genes shows that the appearance of high scoring motifs in the 3' UTR are

not necessarily completely conserved between species. However, a subset (13%) of these *C. briggsae* sequences do appear to have high scores (> 5.5 bits) which are conserved. A cutoff of 5.5 bits still allows significant overlap with the score distribution of the ribosomal protein genes.

The highest scoring 100 of the non-ribosomal predictions (~0.5% of the total) all score over 7.5 bits. These 100 predictions come from genes which may therefore have some function related to that of the ribosomal protein genes. Most of these (77) have some annotation evidence, either from WormBase, or from analysis of protein domains and BLASTP homologies to better annotated proteins. Appendix I shows the set of 77 *C. elegans* polyadenylation signal predictions where the motif score was greater than 7.5 bits. Those genes thought to have some role related to that of the ribosomal proteins are marked with an asterisk. For those *C. elegans* genes with a putative *C. briggsae* orthologue, ribosomal motif log odds scores were also found for the orthologue's 3' UTR. The ranks of the log odds scores are also provided in the Appendix.

There are 22 (29% of the 77 having annotation) whose annotations confirm likely function in translation. Genes in this annotated set include 3 genes related to eukaryotic translation factors, 5 involved in tRNA synthesis and processing, and 11 contributing to ribosomal and rRNA maturation. These can be seen in Table 11.

Table 11. A subset of the C. elegans genes having polyadenylation signals closest resembling those seen in ribosomal proteins. These have a log odds score that is within the top 0.5% of scores. These are the 22 (of 77) whose annotation suggests involvement in translation.

Elegans CDS	Elegans log odds score	Briggsae CDS	Briggsae log odds score	Description
Y48A6B.3	10.909	CBG18231	10.620	Contains Protein domains known in Ribosomal proteins. Similarity to L7. COG suggests Box H/ACA snoRNP component, involved in ribosomal RNA pseudouridinylation
F10E9.11	10.878	CBG16573	-3.314	Similarity to elegans helicase, but also similar to Rat splicing factor and Yeast rRNA processing protein
F10E7.5	10.711	CBG13068	2.064	Similar to Ribosomal protein L-10 (may be L-10?) Similar to non-elegans ribosomal proteins. Cog suggests involved in mRNA turnover
W06H3.2	10.681	CBG23897	-5.100	pus-1 encodes a putative tRNA pseudouridine synthase
C28H8.11a	10.228	no_briggsae	-	Trp dioxygenase - trp Metabolism
Y105E8B.7	9.827	CBG19797	-7.843	YEATS family domain - cog suggests similarity to eukaryotic transcription factor IIF
ZK524.3b	9.65	no_briggsae	-	lrs-2 Leucyl tRNA synthetase - probably mitochondrial
C50F2.1	9.265	no_briggsae	-	Contains ARM fold often found in RNA binding.Tranlation initiation proteins
T01C3.7	9.196	CBG11588	11.559	fib-1 Fibrillarin - nucleolar rRNA processing
Y45F10D.7	9.079	CBG22378	3.040	WD40 repeats - thought to be involved in 18S rRNA maturation
Y56A3A.11	8.807	no_briggsae	-	tRNA splicing endonuclease
K07E8.7	8.688	CBG19546	1.800	Mitochondrial pseudouridylate synthase (RNA)
C01B10.8	8.577	CBG05389	4.274	Spermine/spermidine synthase has S-adenosyl-methione dependent methyltransferase activity
F28D1.8	8.413	no_briggsae	-	Possible peptide-prolyl cis-trans isomerase
W02A11.1	8.096	CBG13601	2.567	Cog suggests tRNA(1-methyladenosine) methyltransferase, subunit GCD14 [KOG2915]
Y24D9A.4c	7.995	no_briggsae	-	Ribosomal protein rpl-7A/rpl-8
F18A11.6	7.758	no_briggsae	-	SNAP50 - Small nuclear RNA activating protein complex - 50kD subunit (SNAP50)
T23D8.7	7.734	CBG03777	5.666	High similarity to eif-2C/argonaute
T03F1.7	7.347	CBG11970	1.548	rRNA methyltransferase
F36A2.2	7.337	CBG12371	8.207	tRNA modification
C07E3.2	7.268	CBG02729	-4.740	Predicted protein involved in nuclear export of pre-ribosomes
W04B5.4	7.148	CBG15659	-6.639	Mitochondrial rpl-30

Of these 22, 15 have a *C. briggsae* orthologue. 4 of these contain a motif score of greater than 5 bits (giving significant overlap with the distribution of ribosomal genes), of which two are greater than 10 bits. The signal appears to be conserved across species in only a small number of genes. Bearing in mind the width of the distribution of the bit scores of ribosomal protein 3' UTRs (Figure 37) and the observation that many ribosomal sequences were discarded from the 136 total during model building to arrive at 57, the function, if any, provided by this motif may be highly specialised within translation.

6.6. Discussion

It has been observed that the regulation of synthesis of the translational apparatus is at the translational level (Meyuhas 2000). Ribosomal protein mRNAs commonly contain a 5' terminal oligopyrimidine tract (TOP) (Levy et al. 1991), which is thought to bind to La protein (Cardinali et al. 1993) with Cellular Nucleic Acid Binding Protein binding downstream (Pellizzoni et al. 1997). Subsequently, other genes involved in translation and its regulation have been found to have TOP mRNAs (Meyuhas 2000). The studies carried out in vertebrates suggest that there is a precedent for searching for some form of class-specific regulation at the mRNA level in the nematodes.

An important aspect of nematode molecular biology is the phenomenon of *trans*-splicing (Blumenthal and Steward, 1997). Approximately 70% of *C. elegans* genes are *trans*-spliced, including all but two of the ribosomal proteins. The efficiency of the *trans*-splicing reaction and the introduction of the conserved *trans*-splice leader

sequence means that these genes have a very short 5' UTR, often of just a few bases. There are only two ribosomal protein genes that do have long 5' UTR sequences as determined by EST (Expressed Sequence Tag) alignment. A large number of the supporting ESTs start with ACTTTT, which is pyrimidine rich, and potentially a good match to the TOP sequence.

Given the lack of 5' UTRs in many nematode ribosomal protein mRNAs, it could be that the element allowing their common control is in the 3' UTR. Of the high-scoring set of genes observed above, it seems quite plausible that genes such as fibrillarin, which is involved in rRNA processing, should be under common control with the ribosomal protein genes. It is additionally promising that fibrillarin has the highest bit score in *C. briggsae*. The appearance in this set of some genes, which are unlikely to be involved in translation however, suggests that the motif alone may not be specific for this function.

6.7. Conclusions

We have seen in this chapter that some ribosomal protein genes from both *C. elegans* and *C. briggsae* contain a distinctive sequence motif around the polyadenylation signal. This motif is also found around the polyadenylation signals from other genes, some of which are known to be involved in translation.

There may be other regulatory sequence motifs related to other functions. The motif described here was found by the coordinated analysis of ribosomal protein genes; similar functional clustering has been used previously to find novel regulatory motifs, such as in histones (Dominski et al. 1999).

One suggestion for future work would be to see if this sequence motif is specific to nematodes or whether it is found in a wider range of other species. If it is only required in a subset of ribosomal protein mRNAs, it would be interesting to rationalise why this subset in particular might need some sequence motif. Another approach would be to obtain direct experimental evidence for its function.

6.8. Collaboration – the analysis of another 3' UTR binding motif

I was involved in collaboration with David Bernstein from Professor Marvin Wickens' lab at the University of Wisconsin-Madison. The work concerned an example of an evolutionarily and functionally conserved 3' UTR motif. This is that found in genes regulated by the PUF proteins (Wickens et al. 2002). These proteins are thought to bind to the 3' UTR of target genes, and thus repress expression by the separate mechanisms of promoting mRNA degradation or interfering with the formation of the mRNA-protein particle (mRNP). Repression by PUF proteins is particularly important during development; they are thought to maintain stem cells by preventing premature differentiation, and to repress the *C. elegans* feminine-repressor fem-3, thus permitting switching from spermatogenesis to oogenesis in hermaphrodites.

Looking for 3' UTRs containing binding sites for PUF proteins can give an insight into the timing and targets of regulatory events in development. Although methods for identifying protein-RNA binding exist (Bernstein et al. 2002), it would be prohibitively onerous to carry out such an analysis on a whole-genome scale. Accurate computational detection of PUF protein binding sites can reduce the search

space to a tractable size, and in addition, can provide independent confirmation of *in-vitro/vivo* work.

In a collaborative project (see Appendix II), (Bernstein et al. 2005) used mutagenesis to identify nucleotides that are essential for the binding of a *C. elegans* PUF protein, FBF-1, to a target 3' UTR. Several rounds of mutagenesis allowed the development and optimisation of binding consensus. The identification of essential "core" and influential "flanking" bases within the RNA sequence enabled us to build binding site consensus models (Dsouza et al. 1997), that constrain core residues whilst allowing for degeneracy outside the region. These were used to search against the set of *C. elegans* 3' UTRs. This computational search enabled the establishment of a set of 150 possible targets for FBF-1. In the collaborative paper, yeast three-hybrid analysis confirmed the formation of mRNA-FBF-1 complexes by 70% of a representative set of sequences from this candidate set. This shows that the computational model is a reasonable. The further analysis of the 3' UTR sequence from those genes found experimentally to have FBF-1 binding sites could be used to refine the model. This way, a combination of computational and laboratory techniques has furthered our knowledge of developmental biology. It serves also as a good example as to how genes can be co-regulated at the post-transcriptional level by a sequence motif in the 3' UTR.

7. Conclusions

Whole genome sequences are now being made available at a rate, the order of which the early pioneers of DNA sequencing could only have dreamed. However, in order to achieve a commensurate understanding of systems and molecular biology, it is necessary to annotate these genomes accurately and to develop new computational tools to help us. Each genome must be interpreted, both in itself and (now increasingly importantly) in the context of others, so that functional, regulatory, and evolutionary information can be found. Without annotation, a genome sequence is of little use.

In this thesis, the main motivation has been the problem of polyadenylation signal prediction. Polyadenylation signal prediction can serve as an alternative method to transcript alignment for annotating 3' UTRs. Some evidence for alternative polyadenylation can also be found. Although it does not provide all the information that we gain by having full-length transcripts, computational polyadenylation signal detection is fast and easy by comparison, and complements data found in the laboratory.

In Chapter 3, by the assembly and functional alignment of large sets of experimentally confirmed cleavage and polyadenylation sites, I have shown that the information specifying this important signal is encoded within nucleotide frequencies in the vicinity. I have shown that a hidden Markov model approach is appropriate for detection of such signals.

The first models built were for the detection of polyadenylation signals in *C. elegans*. Sensitivity in this organism may approach 90%, and the model appears to be

able to simulate observed cleavage site frequencies in deep alignments with large amounts of cDNA evidence.

I have provided a set of high confidence 3' UTR sequences that are extended to a cleavage site, rather than some end defined by the 3' end of a clipped EST. Data from this analysis is already being found useful by the scientific community (Hieronymus et al. 2004; Porter et al. 2005; Zhang et al. 2005).

The parameters (such as emission frequencies and number of states) required for signal detection in other species such as mouse, human, and fruitfly vary from those developed for *C. elegans*. However, the core algorithms required for annotating a sequence with an HMM remain the same. The problem of how to implement these algorithms, coupled with the need to quickly modify a model (such as by the addition of a new state) led to the development of PAjHMMA (Chapter 2). This is a flexible framework for decoding a generalised hidden Markov model against a DNA sequence. Changing model parameters require no changes to the code, but simply to a text file containing a representation of the model, the states, and their properties.

One other key feature of PAjHMMA is its ability to decode generalised HMMs. This does not lose encoded length information, thus improving over the (ab)use of generic protein profile HMM software.

In chapter 4, I extend the work carried out in *C. elegans,* and show that distinctive nucleotide biases are a feature of polyadenylation signals in other species. The flexible framework shows itself to be robust and adjustable for use in species other than the one for which it was originally developed. For *D. melanogaster,* this work represents the only example of polyadenylation signal prediction specific for this species that I am aware of. In mouse and human, the performance of my software is slightly greater than existing methods at the sensitivity level. On the data set given,

the HMM also has a slightly higher lower bound for specificity. All three methods tested detect about 50% of all signals. An annotation pipeline could possibly use all three groups' software to generate a set of high confidence predictions if all three predict at the same site.

Chapter 5 concentrates on the change, gain and loss of polyadenylation signals over the course of nematode evolution. By comparing orthologous genes in *C. elegans* and *C. briggsae*, over 60% of sites are not conserved, even when the relevant 3' UTR sequence can be aligned. This demonstrates a high turnover of cleavage and polyadenylation sites. High turnover of transcription factor binding sites have been observed in other organisms' enhancers (Ludwig et al. 2000), and thus it appears that our observations are another case where there is high turnover of protein-binding sites.

In about 40% of aligned orthologous 3' UTR pairs, polyadenylation signals are aligned. About a quarter of these aligned hexamer pairs show a mutation, such that different variants of the AATAAA motif are used. The pattern of mutation is striking.

I have previously mentioned the importance of 3' UTR regulatory motifs. In chapter 6, I show that clustering a set of genes known to function together reveals the conservation of a sequence motif either side of the polyadenylation signal. This signal is conserved in *C. briggsae*. In this case, the motif is found initially in the 3' UTR of ribosomal protein genes. Searching for matches to the motif in other *C. elegans* genes shows that it appears in some other genes having some function in translation. It is extremely likely that ribosomal genes are co-expressed, alongside the other genes containing the motif such as translation elongation factors. The appearance of this motif in *C. elegans* genes implicated in translation, coupled with its conservation

between the two nematodes (within ribosomal protein genes), suggests that it has some regulatory function.

This thesis has focussed on the detection of polyadenylation signals by HMM methods. Other sequence features could also be modelled. Future projects that could benefit by utilising the PAjHMMA framework are limited only by researchers' imaginations. Although beyond the scope of this work, it would be quite possible to provide parameters for an entire gene model incorporating a full 3' end model. In particular this may aid in increasing accuracy of terminal exon prediction.

My work has concerned detection and analysis of a sequence feature that is required for mRNA processing; its accurate detection will give access to 3' UTR sequences in many species. Their coordinated analysis should facilitate the discovery of conserved regulatory regions. It is this form of annotation, coupled with breakthroughs in detecting and analysing other sequence features such as promoters and non-coding RNA genes, that will best complement our current use of protein coding gene annotation and thus fuel our further understanding of systems biology.

References

(1998). "The C. elegans Sequencing Consortium. (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology." Science **282**(5396): 2012-8.

Alberts, B., D. Bray, et al. (2002). Molecular Biology of the Cell, Garland.

Alt, F. W., V. Enea, et al. (1980). "Activity of multiple light chain genes in murine myeloma cells producing a single, functional light chain." Cell **21**(1): 1-12.

Awasthi, S. and J. C. Alwine (2003). "Association of polyadenylation cleavage factor I with U1 snRNP." Rna **9**(11): 1400-9.

Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol **2**: 28-36.

Bateman, A., L. Coin, et al. (2004). "The Pfam protein families database." Nucleic Acids Res **32 Database issue**: D138-41.

Beaudoing, E., S. Freier, et al. (2000). "Patterns of variant polyadenylation signal usage in human genes." Genome Res **10**(7): 1001-10.

Berget, S. M. (1995). "Exon recognition in vertebrate splicing." J Biol Chem **270**(6): 2411-4.

Bernstein, D. S., N. Buter, et al. (2002). "Analyzing mRNA-protein complexes using a yeast three-hybrid system." Methods **26**(2): 123-41.

Bernstein, D. S., B. Hook, et al. (2005). "Binding specificity and mRNA targets of a *C. elegans* PUF protein, FBF-1." RNA **11**(4): 447-58.

Bienroth, S., W. Keller, et al. (1993). "Assembly of a processive messenger RNA polyadenylation complex." Embo J **12**(2): 585-94.

Blumenthal, T. (1995). "Trans-splicing and polycistronic transcription in Caenorhabditis elegans." Trends Genet **11**(4): 132-6.

Blumenthal, T. and K. Steward (1997). RNA processing and Gene Structure. C. elegans II. D. Riddle, T. Blumenthal, B. Meyer and J. R. Preiss. Cold Spring Harbour, Cold Spring Harbour Laboratory Press**:** 117-145.

Bousquet-Antonelli, C., C. Presutti, et al. (2000). "Identification of a regulated pathway for nuclear pre-mRNA turnover." Cell **102**(6): 765-75.

Brukner, I., R. Sanchez, et al. (1995). "Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides." Embo J **14**(8): 1812-8.

Burge, C. and S. Karlin (1997). "Prediction of complete gene structures in human genomic DNA." J Mol Biol **268**(1): 78-94.

Cardinali, B., M. Di Cristina, et al. (1993). "Interaction of proteins with the mRNA for ribosomal protein L1 in Xenopus: structural characterization of in vivo complexes and identification of proteins that bind in vitro to its 5'UTR." Nucleic Acids Res **21**(10): 2301-8.

Castelo-Branco, P., A. Furger, et al. (2004). "Polypyrimidine tract binding protein modulates efficiency of polyadenylation." Mol Cell Biol **24**(10): 4174-83.

Chen, C. Y. and A. B. Shyu (2003). "Rapid deadenylation triggered by a nonsense codon precedes decay of the RNA body in a mammalian cytoplasmic nonsense-mediated decay pathway." Mol Cell Biol **23**(14): 4805-13.

Chen, F., C. C. MacDonald, et al. (1995). "Cleavage site determinants in the mammalian polyadenylation signal." Nucleic Acids Res **23**(14): 2614-20.
Chen, F. and J. Wilusz (1998). "Auxiliary downstream elements are required for efficient polyadenylation of mammalian pre-mRNAs." Nucleic Acids Res **26**(12): 2891-8.
Chen, N., T. W. Harris, et al. (2005). "WormBase: a comprehensive data resource for Caenorhabditis biology and genomics." Nucleic Acids Res **33 Database Issue**: D383-9.
Clamp, M., J. Cuff, et al. (2004). "The Jalview Java alignment editor." Bioinformatics **20**(3): 426-7.
Colgan, D. F. and J. L. Manley (1997). "Mechanism and regulation of mRNA polyadenylation." Genes Dev **11**(21): 2755-66.
Cooke, C. and J. C. Alwine (1996). "The cap and the 3' splice site similarly affect polyadenylation efficiency." Mol Cell Biol **16**(6): 2579-84.
Dantonel, J. C., K. G. Murthy, et al. (1997). "Transcription factor TFIID recruits factor CPSF for formation of 3' end of mRNA." Nature **389**(6649): 399-402.
Darnell, J. E., R. Wall, et al. (1971). "An adenylic acid-rich sequence in messenger RNA of HeLa cells and its possible relationship to reiterated sites in DNA." Proc Natl Acad Sci U S A **68**(6): 1321-5.
Dermitzakis, E. T., C. M. Bergman, et al. (2003). "Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites." Mol Biol Evol **20**(5): 703-14.
Dichtl, B. and W. Keller (2001). "Recognition of polyadenylation sites in yeast pre-mRNAs by cleavage and polyadenylation factor." Embo J **20**(12): 3197-209.
Dominski, Z. and W. F. Marzluff (1999). "Formation of the 3' end of histone mRNA." Gene **239**(1): 1-14.
Down, T. A. and T. J. Hubbard (2002). "Computational detection and location of transcription start sites in mammalian genomic DNA." Genome Res **12**(3): 458-61.
Dsouza, M., N. Larsen, et al. (1997). "Searching for patterns in genomic data." Trends Genet **13**(12): 497-8.
Durbin, R., S. R. Eddy, et al. (1998). Biological Sequence Analysis. Cambridge, Cambridge University Press.
Eddy, S. R. (2001). "Non-coding RNA genes and the modern RNA world." Nat Rev Genet **2**(12): 919-29.
Eddy, S. R. (2002). "A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure." BMC Bioinformatics **3**(1): 18.
Edmonds, M., M. H. Vaughan, Jr., et al. (1971). "Polyadenylic acid sequences in the heterogeneous nuclear RNA and rapidly-labeled polyribosomal RNA of HeLa cells: possible evidence for a precursor relationship." Proc Natl Acad Sci U S A **68**(6): 1336-40.
Edwalds-Gilbert, G., K. L. Veraldi, et al. (1997). "Alternative poly(A) site selection in complex transcription units: means to an end?" Nucleic Acids Res **25**(13): 2547-61.
Evans, D., I. Perez, et al. (2001). "A complex containing CstF-64 and the SL2 snRNP connects mRNA 3' end formation and trans-splicing in C. elegans operons." Genes Dev **15**(19): 2562-71.

Ford, J. P. and M. T. Hsu (1978). "Transcription pattern of in vivo-labeled late simian virus 40 RNA: equimolar transcription beyond the mRNA 3' terminus." J Virol **28**(3): 795-801.

Ford, L. P., P. S. Bagga, et al. (1997). "The poly(A) tail inhibits the assembly of a 3'-to-5' exonuclease in an in vitro RNA stability system." Mol Cell Biol **17**(1): 398-406.

Fromont-Racine, M., B. Senger, et al. (2003). "Ribosome assembly in eukaryotes." Gene **313**: 17-42.

Gardiner-Garden, M. and M. Frommer (1987). "CpG islands in vertebrate genomes." J Mol Biol **196**(2): 261-82.

Gautheret, D. and A. Lambert (2001). "Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles." J Mol Biol **313**(5): 1003-11.

Gavis, E. R., D. Curtis, et al. (1996). "Identification of cis-acting sequences that control nanos RNA localization." Dev Biol **176**(1): 36-50.

Gill, G. (2001). "Regulation of the initiation of eukaryotic transcription." Essays Biochem **37**: 33-43.

Graber, J. H., C. R. Cantor, et al. (1999). "In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species." Proc Natl Acad Sci U S A **96**(24): 14055-60.

Graber, J. H., G. D. McAllister, et al. (2002). "Probabilistic prediction of Saccharomyces cerevisiae mRNA 3'-processing sites." Nucleic Acids Res **30**(8): 1851-8.

Gray, N. K. (1998). "Translational control by repressor proteins binding to the 5'UTR of mRNAs." Methods Mol Biol **77**: 379-97.

Griffiths-Jones, S., S. Moxon, et al. (2005). "Rfam: annotating non-coding RNAs in complete genomes." Nucleic Acids Res **33 Database Issue**: D121-4.

Gross, S. and C. L. Moore (2001). "Rna15 interaction with the A-rich yeast polyadenylation signal is an essential step in mRNA 3'-end formation." Mol Cell Biol **21**(23): 8045-55.

Guo, Z. and F. Sherman (1996). "3'-end-forming signals of yeast mRNA." Trends Biochem Sci **21**(12): 477-81.

Hajarnavis, A., I. Korf, et al. (2004). "A probabilistic model of 3' end formation in Caenorhabditis elegans." Nucleic Acids Res **32**(11): 3392-9.

Hieronymus, H. and P. A. Silver (2004). "A systems view of mRNP biology." Genes Dev **18**(23): 2845-60.

Higgs, D. R., S. E. Goodbourn, et al. (1983). "Alpha-thalassaemia caused by a polyadenylation signal mutation." Nature **306**(5941): 398-400.

Hirose, Y. and J. L. Manley (1998). "RNA polymerase II is an essential mRNA polyadenylation factor." Nature **395**(6697): 93-6.

Howe, K. L., T. Chothia, et al. (2002). "GAZE: a generic framework for the integration of gene-prediction data by dynamic programming." Genome Res **12**(9): 1418-27.

Huang, T., S. Kuersten, et al. (2001). "Intercistronic region required for polycistronic pre-mRNA processing in Caenorhabditis elegans." Mol Cell Biol **21**(4): 1111-20.

Hubbard, T., D. Andrews, et al. (2005). "Ensembl 2005." Nucleic Acids Res **33 Database Issue**: D447-53.

Hubert, N., R. Walczak, et al. (1996). "A protein binds the selenocysteine insertion element in the 3'-UTR of mammalian selenoprotein mRNAs." Nucleic Acids Res **24**(3): 464-9.

Jareborg, N., E. Birney, et al. (1999). "Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs." Genome Res **9**(9): 815-24.

Johnson, D. S., B. Davidson, et al. (2004). "Noncoding regulatory sequences of Ciona exhibit strong correspondence between evolutionary constraint and functional importance." Genome Res **14**(12): 2448-56.

Kasprzyk, A., D. Keefe, et al. (2004). "EnsMart: a generic system for fast and flexible access to biological data." Genome Res **14**(1): 160-9.

Kessler, M. M., M. F. Henry, et al. (1997). "Hrp1, a sequence-specific RNA-binding protein that shuttles between the nucleus and the cytoplasm, is required for mRNA 3'-end formation in yeast." Genes Dev **11**(19): 2545-56.

Larizza, A., W. Makalowski, et al. (2002). "Evolutionary dynamics of mammalian mRNA untranslated regions by comparative analysis of orthologous human, artiodactyl and rodent gene pairs." Comput Chem **26**(5): 479-90.

Lee, S. Y., J. Mendecki, et al. (1971). "A polynucleotide segment rich in adenylic acid in the rapidly-labeled polyribosomal RNA component of mouse sarcoma 180 ascites cells." Proc Natl Acad Sci U S A **68**(6): 1331-5.

Legendre, M. and D. Gautheret (2003). "Sequence determinants in human polyadenylation site selection." BMC Genomics **4**(1): 7.

Levy, S., D. Avni, et al. (1991). "Oligopyrimidine tract at the 5' end of mammalian ribosomal protein mRNAs is required for their translational control." Proc Natl Acad Sci U S A **88**(8): 3319-23.

Lin, C. H. and J. G. Patton (1995). "Regulation of alternative 3' splice site selection by constitutive splicing factors." Rna **1**(3): 234-45.

Ludwig, M. Z., C. Bergman, et al. (2000). "Evidence for stabilizing selection in a eukaryotic enhancer element." Nature **403**(6769): 564-7.

Ludwig, M. Z., A. Palsson, et al. (2005). "Functional Evolution of a cis-Regulatory Module." PLoS Biol **3**(4): e93.

Lund, M. and J. Kjems (2002). "Defining a 5' splice site by functional selection in the presence and absence of U1 snRNA 5' end." Rna **8**(2): 166-79.

Lutz, C. S., K. G. Murthy, et al. (1996). "Interaction between the U1 snRNP-A protein and the 160-kD subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency in vitro." Genes Dev **10**(3): 325-37.

MacDonald, C. C. and J. L. Redondo (2002). "Reexamining the polyadenylation signal: were we wrong about AAUAAA?" Mol Cell Endocrinol **190**(1-2): 1-8.

Makarov, V. (2002). "Computer programs for eukaryotic gene prediction." Brief Bioinform **3**(2): 195-9.

Mathe, C., M. F. Sagot, et al. (2002). "Current methods of gene prediction, their strengths and weaknesses." Nucleic Acids Res **30**(19): 4103-17.

Mattick, J. S. (2001). "Non-coding RNAs: the architects of eukaryotic complexity." EMBO Rep **2**(11): 986-91.

Meyuhas, O. (2000). "Synthesis of the translational apparatus is regulated at the translational level." Eur J Biochem **267**(21): 6321-30.

Mignone, F., G. Grillo, et al. (2005). "UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs." Nucleic Acids Res **33 Database Issue**: D141-6.

Moreira, A., Y. Takagaki, et al. (1998). "The upstream sequence element of the C2 complement poly(A) signal activates mRNA 3' end formation by two distinct mechanisms." Genes Dev **12**(16): 2522-34.

Morimoto, R. I. (1993). "Cells in stress: transcriptional activation of heat shock genes." Science **259**(5100): 1409-10.

Neu-Yilik, G., N. H. Gehring, et al. (2004). "Nonsense-mediated mRNA decay: from vacuum cleaner to Swiss army knife." Genome Biol **5**(4): 218.

Nirenberg, M. (2004). "Historical review: Deciphering the genetic code--a personal account." Trends Biochem Sci **29**(1): 46-54.

Niu, D. K., K. Lin, et al. (2003). "Strand compositional asymmetries of nuclear DNA in eukaryotes." J Mol Evol **57**(3): 325-34.

Niwa, M. and S. M. Berget (1991). "Mutation of the AAUAAA polyadenylation signal depresses in vitro splicing of proximal but not distal introns." Genes Dev **5**(11): 2086-95.

Ohno, M., H. Sakamoto, et al. (1987). "Preferential excision of the 5' proximal intron from mRNA precursors with two introns as mediated by the cap structure." Proc Natl Acad Sci U S A **84**(15): 5187-91.

Olsen, P. H. and V. Ambros (1999). "The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation." Dev Biol **216**(2): 671-80.

Pellizzoni, L., F. Lotti, et al. (1997). "Cellular nucleic acid binding protein binds a conserved region of the 5' UTR of Xenopus laevis ribosomal protein mRNAs." J Mol Biol **267**(2): 264-75.

Porter, M. Y., M. Turmaine, et al. (2005). "Identification and characterization of Caenorhabditis elegans palmitoyl protein thioesterase1." J Neurosci Res **79**(6): 836-48.

Pritsker, M., Y. C. Liu, et al. (2004). "Whole-genome discovery of transcription factor binding sites by network-level conservation." Genome Res **14**(1): 99-108.

Proudfoot, N. (1991). "Poly(A) signals." Cell **64**(4): 671-4.

Proudfoot, N. J. (2001). "Genetic dangers in poly(A) signals." EMBO Rep **2**(10): 891-2.

Proudfoot, N. J., A. Furger, et al. (2002). "Integrating mRNA processing with transcription." Cell **108**(4): 501-12.

Qu, X., Y. Qi, et al. (2002). "Generation of multiple mRNA transcripts from the novel human apoptosis-inducing gene hap by alternative polyadenylation utilization and the translational activation function of 3' untranslated region." Arch Biochem Biophys **400**(2): 233-44.

Rabiner, L. R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition." Proceedings of the IEEE **77**: 257-286.

Raphael, C. (1999). "Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models." IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(4): 360-370.

Rivas, E. and S. R. Eddy (2001). "Noncoding RNA gene detection using comparative sequence analysis." BMC Bioinformatics **2**(1): 8.

Salamov, A. A. and V. V. Solovyev (1997). "Recognition of 3'-processing sites of human mRNA precursors." Comput Appl Biosci **13**(1): 23-8.

Salzberg, S. L., M. Pertea, et al. (1999). "Interpolated Markov models for eukaryotic gene finding." Genomics **59**(1): 24-31.

Scorilas, A. (2002). "Polyadenylate polymerase (PAP) and 3' end pre-mRNA processing: function, assays, and association with disease." Crit Rev Clin Lab Sci **39**(3): 193-224.

Sheets, M. D., S. C. Ogg, et al. (1990). "Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro." Nucleic Acids Res **18**(19): 5799-805.

Simmer, F., C. Moorman, et al. (2003). "Genome-wide RNAi of C. elegans using the hypersensitive rrf-3 strain reveals novel gene functions." PLoS Biol **1**(1): E12.

Stabenau, A., G. McVicker, et al. (2004). "The Ensembl core software libraries." Genome Res **14**(5): 929-33.

Stajich, J. E., D. Block, et al. (2002). "The Bioperl toolkit: Perl modules for the life sciences." Genome Res **12**(10): 1611-8.

Stanke, M. and S. Waack (2003). "Gene prediction with a hidden Markov model and a new intron submodel." Bioinformatics **19 Suppl 2**: II215-II225.

Stein, L. D., Z. Bao, et al. (2003). "The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics." PLoS Biol **1**(2): E45.

Tabaska, J. E. and M. Q. Zhang (1999). "Detection of polyadenylation signals in human DNA sequences." Gene **231**(1-2): 77-86.

Thastrom, A., L. M. Bingham, et al. (2004). "Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning." J Mol Biol **338**(4): 695-709.

Tian, B., J. Hu, et al. (2005). "A large-scale analysis of mRNA polyadenylation of human and mouse genes." Nucleic Acids Res **33**(1): 201-12.

Timchenko, L. T. (1999). "Myotonic dystrophy: the role of RNA CUG triplet repeats." Am J Hum Genet **64**(2): 360-4.

Touchon, M., A. Arneodo, et al. (2004). "Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes." Nucleic Acids Res **32**(17): 4969-78.

Vagner, S., C. Vagner, et al. (2000). "The carboxyl terminus of vertebrate poly(A) polymerase interacts with U2AF 65 to couple 3'-end processing and splicing." Genes Dev **14**(4): 403-13.

Viterbi, A. (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm." IEEE Transactions on Information Theory: 260-269.

Wahle, E., A. Lustig, et al. (1993). "Mammalian poly(A)-binding protein II. Physical properties and binding to polynucleotides." J Biol Chem **268**(4): 2937-45.

Walhout, A. J. and M. Vidal (2001). "Protein interaction maps for model organisms." Nat Rev Mol Cell Biol **2**(1): 55-62.

Wickens, M., D. S. Bernstein, et al. (2002). "A PUF family portrait: 3'UTR regulation as a way of life." Trends Genet **18**(3): 150-7.

Wilkie, G. S., K. S. Dickson, et al. (2003). "Regulation of mRNA translation by 5'- and 3'-UTR-binding factors." Trends Biochem Sci **28**(4): 182-8.

Xu, N., C. Y. Chen, et al. (1997). "Modulation of the fate of cytoplasmic mRNA by AU-rich elements: key sequence features controlling mRNA deadenylation and decay." Mol Cell Biol **17**(8): 4611-21.

Zhang, H., J. Hu, et al. (2005). "PolyA_DB: a database for mammalian mRNA polyadenylation." Nucleic Acids Res **33 Database Issue**: D116-20.

Zhang, M. Q. (2002). "Computational prediction of eukaryotic protein-coding genes." Nat Rev Genet **3**(9): 698-709.

Zhao, J., L. Hyman, et al. (1999). "Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis." Microbiol Mol Biol Rev **63**(2): 405-45.

Appendices

Appendix I: A table of the 77 *C. elegans* genes with the highest fit to the ribosomal HMM (Chapter 6).

Appendix II:
Bernstein D, Hook B, Hajarnavis A, Opperman L, Wickens M.
Binding specificity and mRNA targets of a *C. elegans* PUF protein, FBF-1.
RNA. 2005 Apr;11(4):447-58.

Appendices

Appendix I: A table of the 77 *C. elegans* genes with the highest fit to the ribosomal HMM (Chapter 6).

Appendix II:
Bernstein D, Hook B, Hajarnavis A, Opperman L, Wickens M.
Binding specificity and mRNA targets of a *C. elegans* PUF protein, FBF-1.
RNA. 2005 Apr;11(4):447-58.