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.

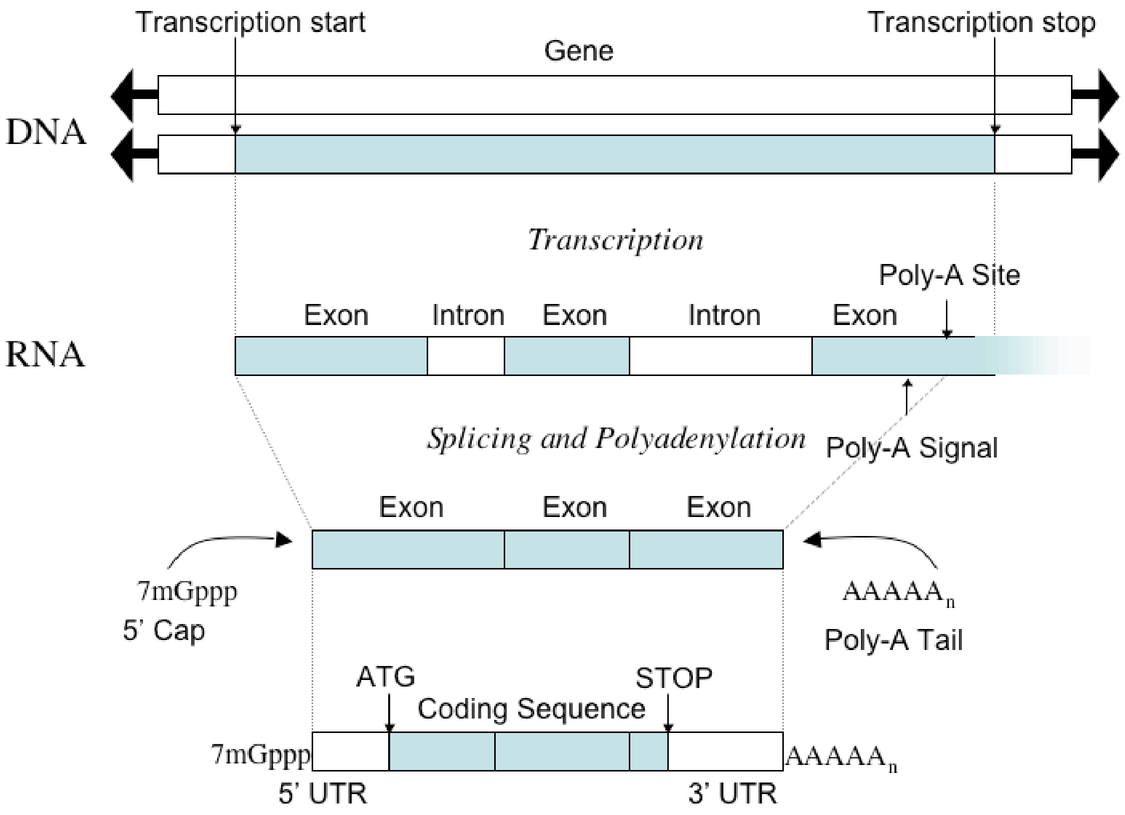

Figure 1. Main steps in the expression of a typical eukaryotic protein coding gene, showing transcription, splicing, and processing.

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$_{\mathrm{Sec}}$. 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.

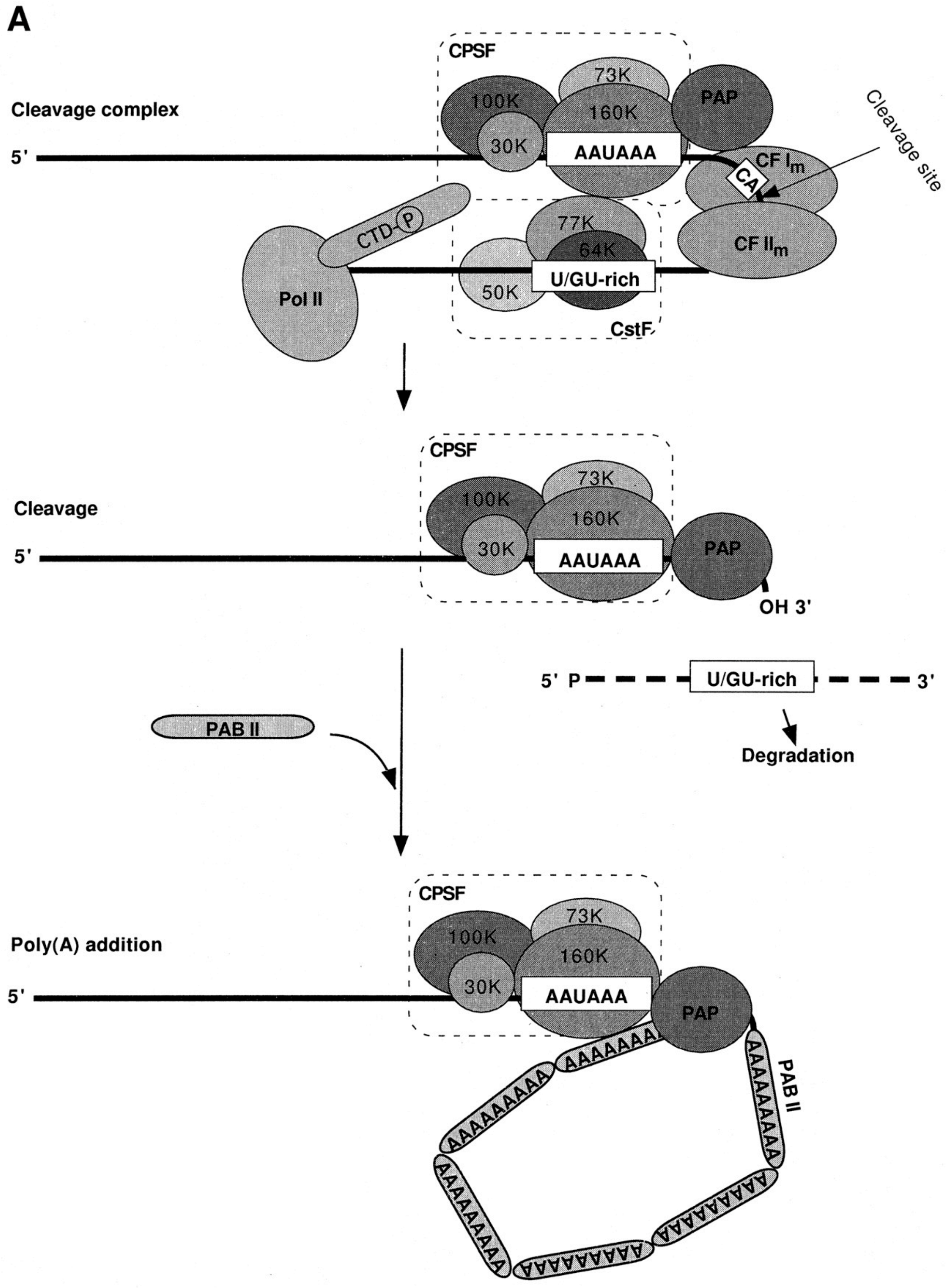

Figure 2. An overview of some of the proteins involved in mammalian 3' end processing. We can see the four subunits of the Cleavage and Polyadenylation Specificity Factor (CPSF), Poly-A Polymerase (PAP), Cleavage Factors I and II (CFI, II), Cleavage Stimulation Factor (CstF), RNA Polymerase II (RNAPol 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).