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.