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.

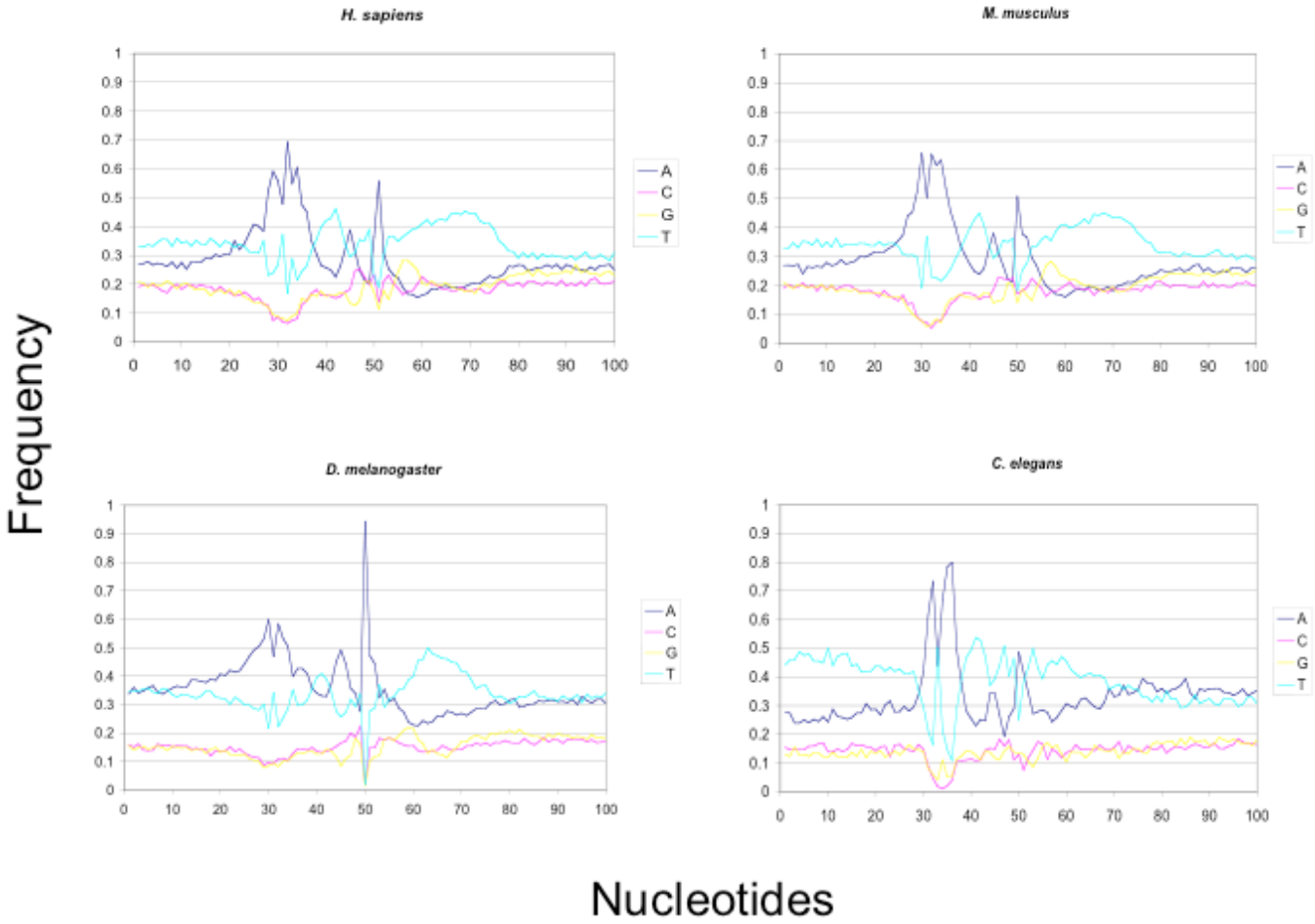

Figure 15. Graphs showing the nucleotide distribution around the cleavage sites of H. sapiens, M. musculus, D. melanogaster, and C. elegans. The maximum likelihood cleavage site occurs at 50 nt in each case.

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

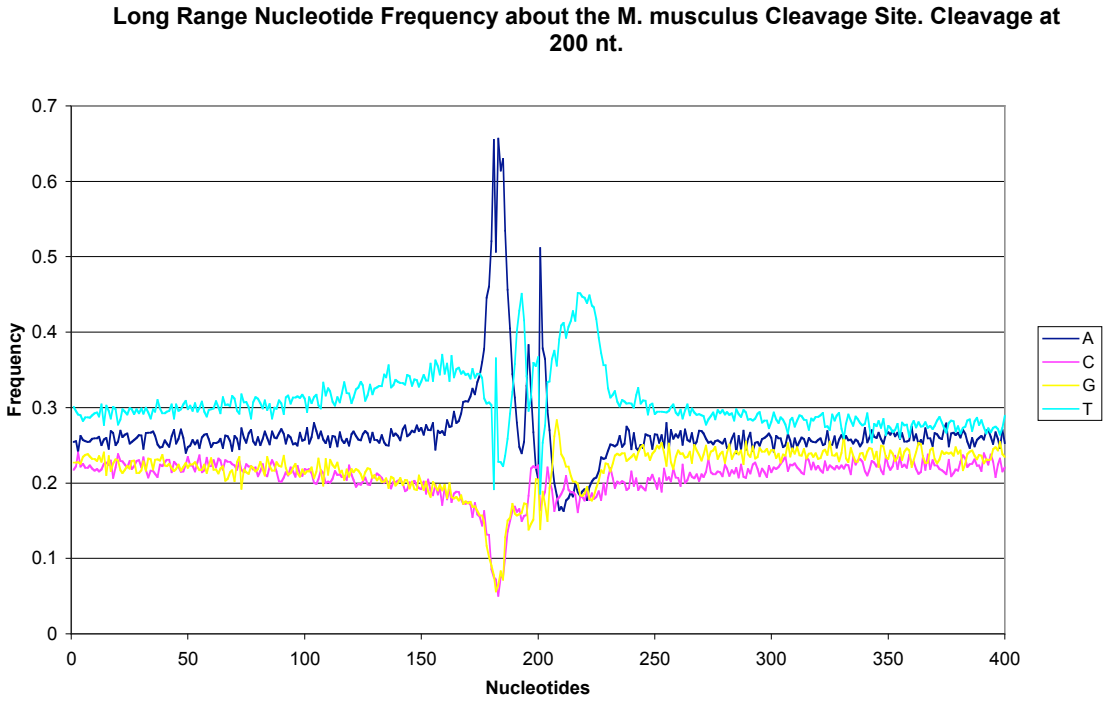

Figure 16. M. musculus graph showing how nucleotide frequency varies over longer ranges.

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.

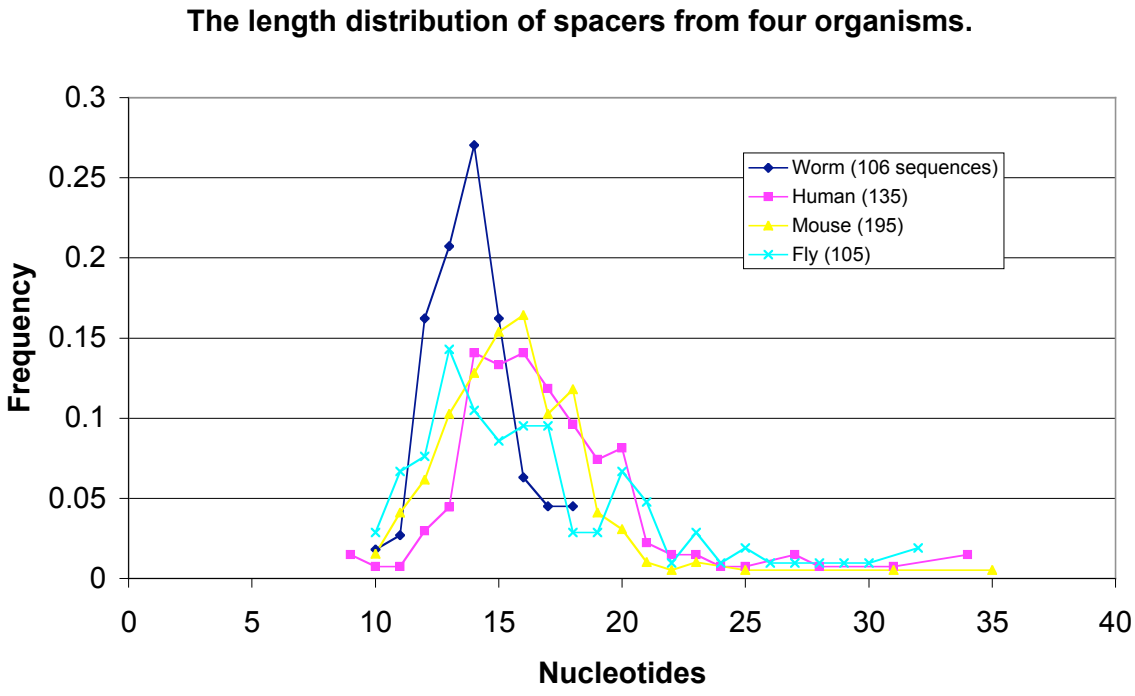

Figure 17. A frequency distribution of the lengths of sequence between unambiguous matches to AATAAA and precisely locatable cleavage sites.

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.

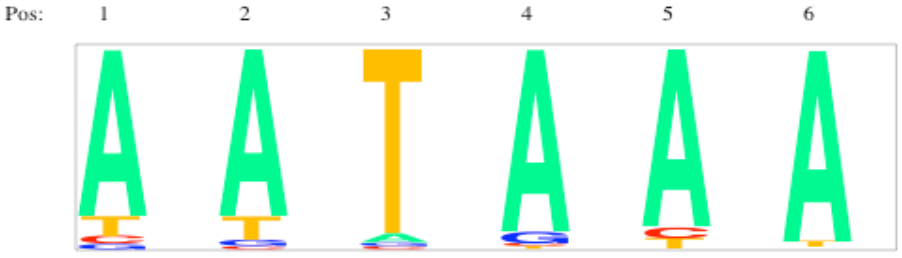

Figure 20. D. melanogaster AATAAA motif.

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).

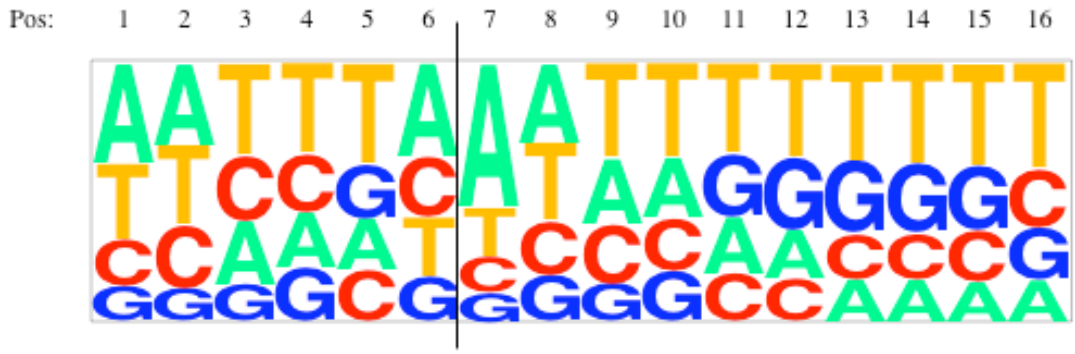

Figure 22. H. sapiens cleavage site weight matrix. Cleavage between positions 6 and 7.

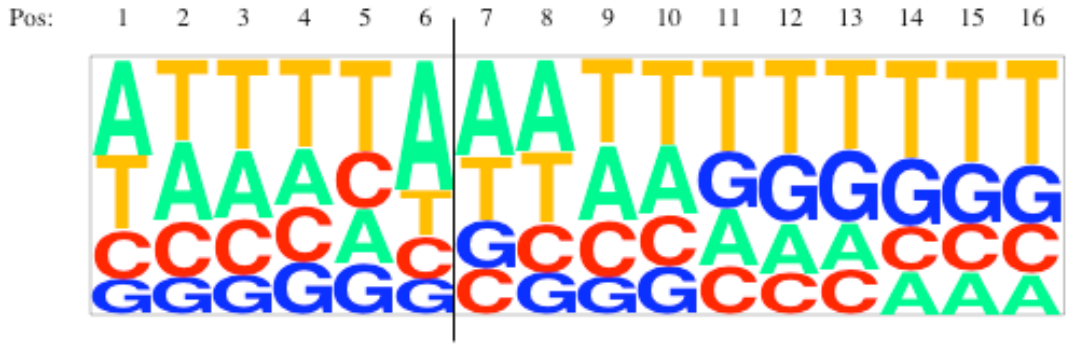

Figure 23. M. musculus cleavage site weight matrix. Cleavage between positions 6 and 7.

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.

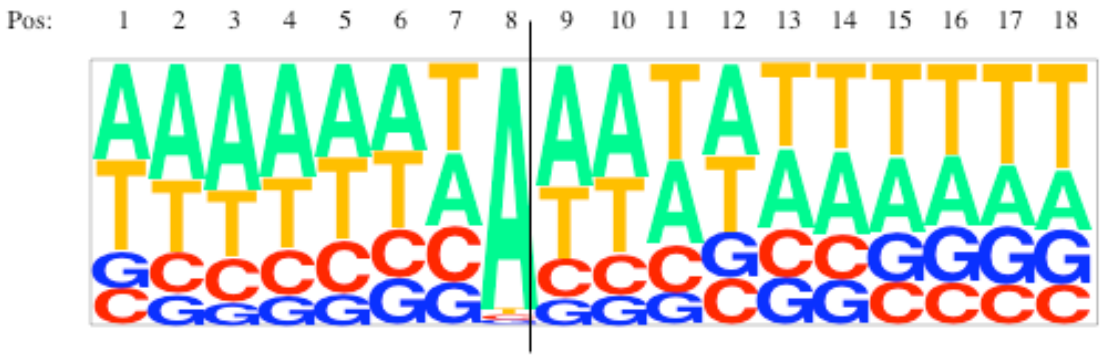

Figure 24. D. melanogaster cleavage site weight matrix. Cleavage between positions 8 and 9

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.