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.

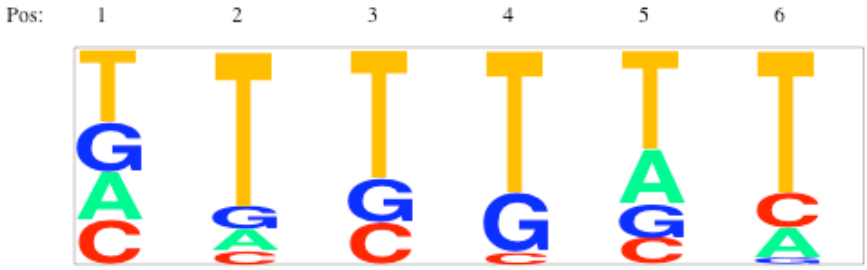

Figure 6. A pictogram of the nucleotide frequencies in the three nucleotides either side of the cleavage site. The cleavage site from 156 sequences occurs between columns 3 and 4. There are no As directly flanking the cleavage site, as it is their absence that defines this class of cleavage. A clear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

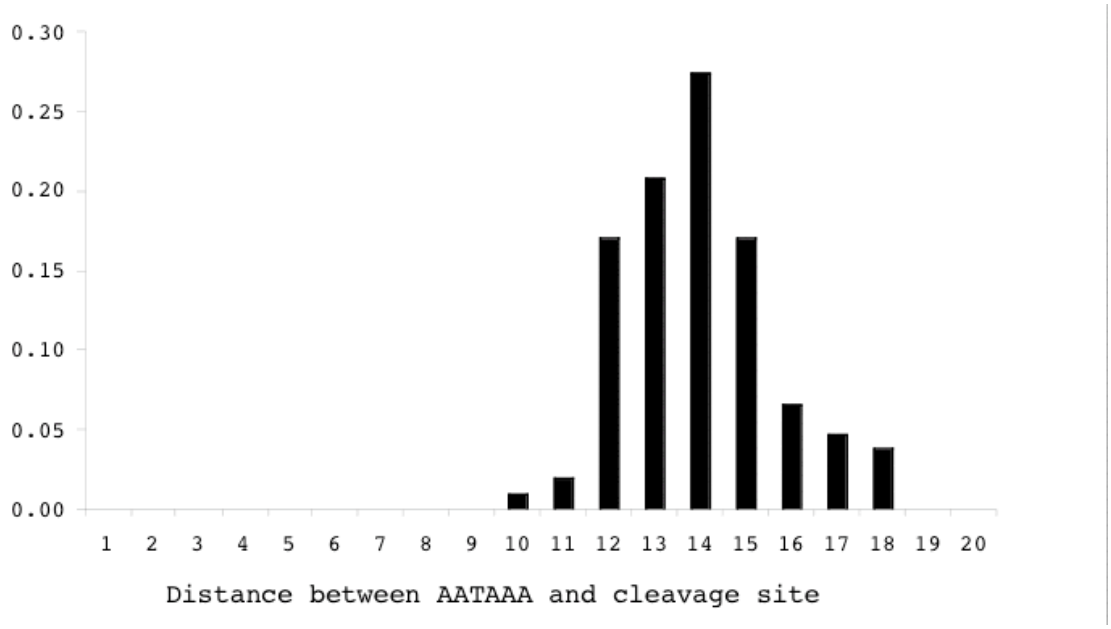

Figure 7. The length distribution of the spacer sequence separating the polyadenylation signal (exact match to AATAAA) and the unambiguously defined cleavage site from 106 sequences.

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.

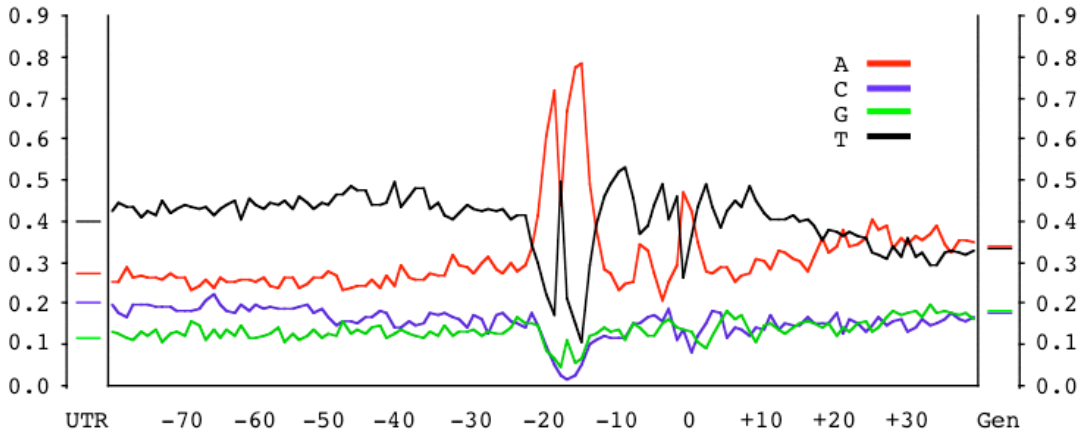

Figure 8. The nucleotide frequencies in 3' UTR, the region -80 to +30 about the cleavage site, and the genomic nucleotide distribution in C. elegans. Each sequence in the training and test sets was annotated into states according to the demarcated zones of distinctive nucleotide frequency.

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)

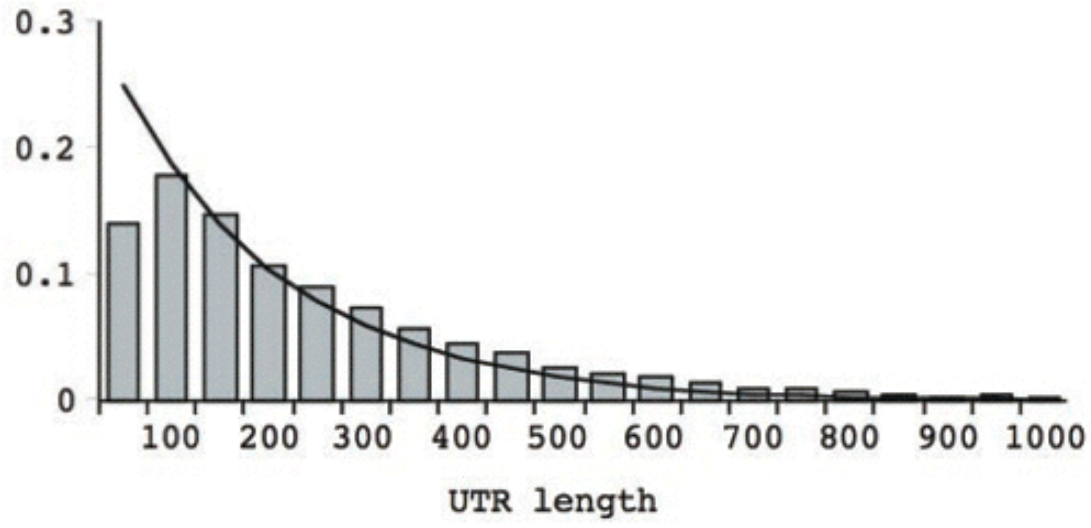

Figure 9. Bars - The length distribution of the 3' UTR sequences in our set of 940. Line - the expectation from a geometric distribution with a mean of 200. The sub-50 nt bar is truncated as a result of our length requirements when building

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.

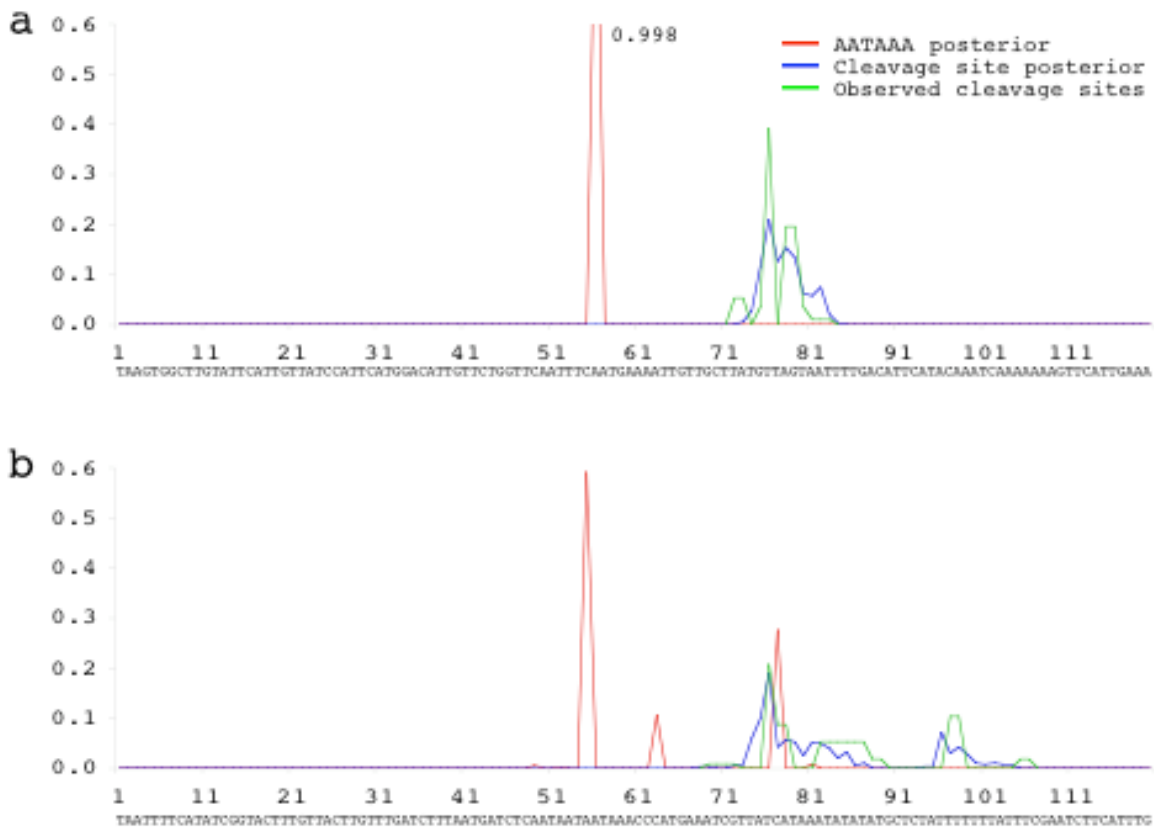

Figure 11. The posterior probability of the AATAAA motif and cleavage site are shown in red and blue lines respectively. The observed frequency of cleavage sites is indicated by a green line. When the cleavage site is ambiguous, the frequency is averaged over the ambiguous positions, which gives the 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.

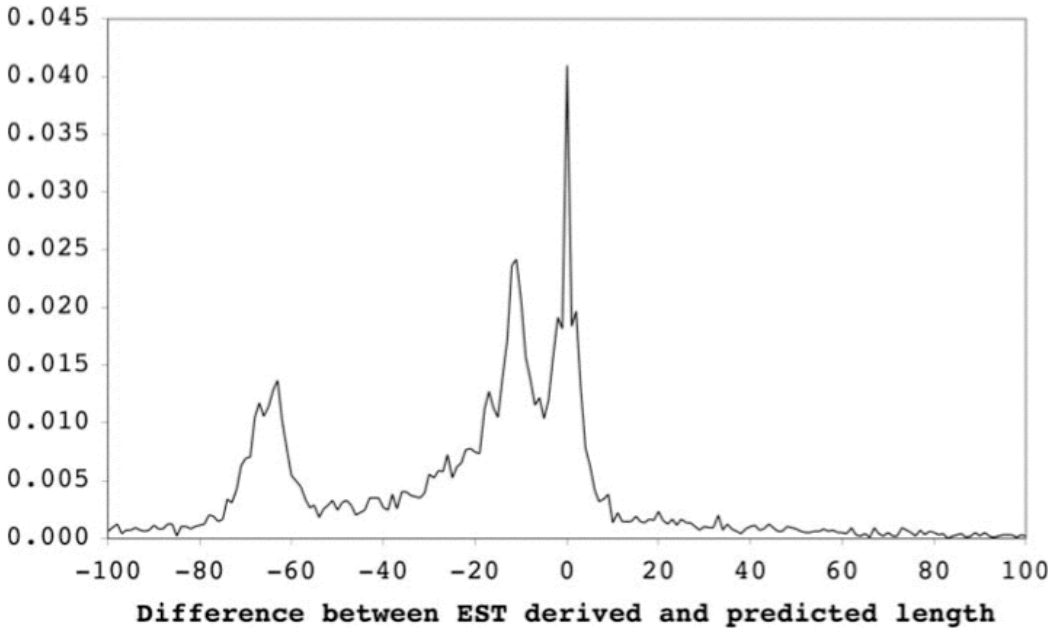

Figure 12. Frequency distribution of the difference between length of 3' UTR as determined by EST alignment and our model.

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.

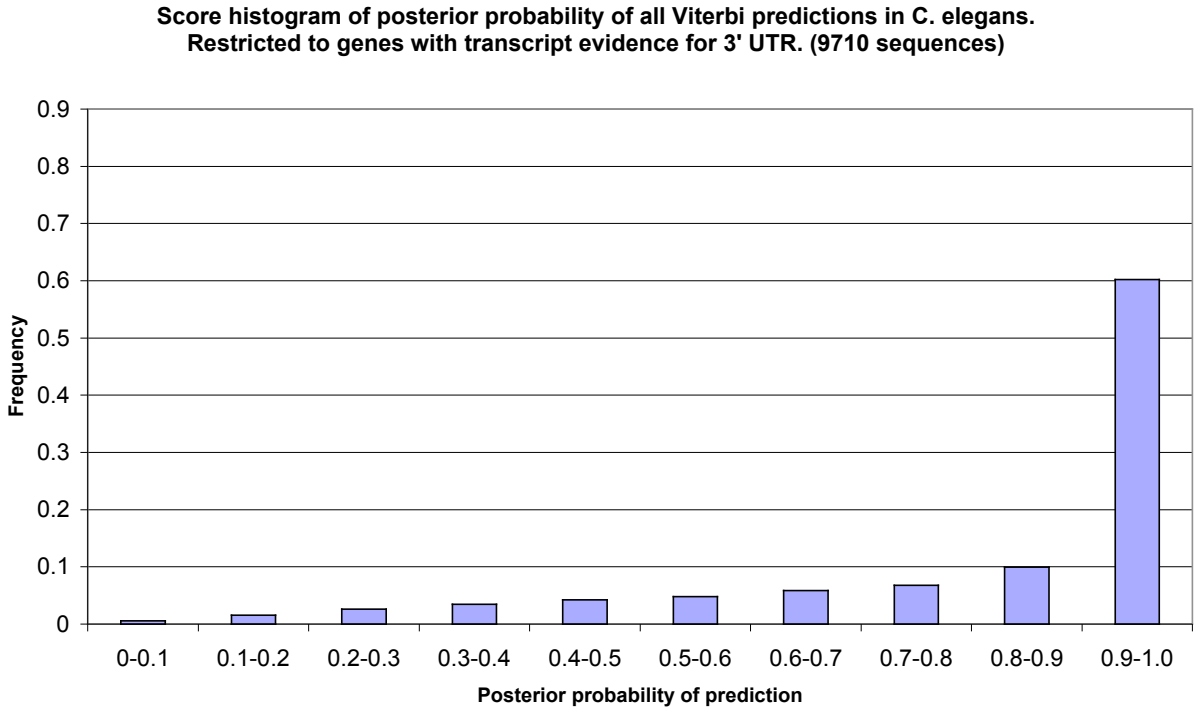

Figure 13. Posterior probability distribution for the Viterbi polyadenylation signal predictions in 9710 3' UTR sequences where a given prediction could be verified by transcript evidence, as ESTs were available.

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.

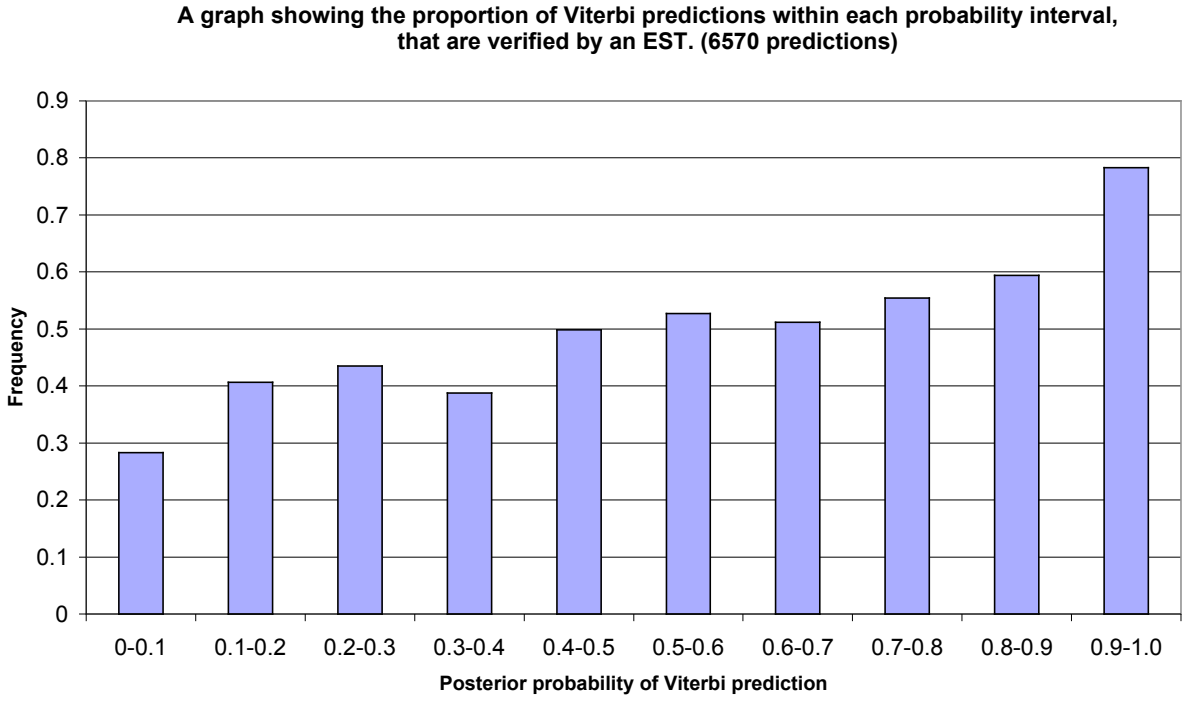

Figure 14. For the 6570 sequences where the position of the Viterbi polyadenylation signal prediction was verified by an EST, this histogram shows

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.