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

5.1. Introduction

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

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

5.1.1. Caenorhabditis briggsae

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

The neutral substitution rate measured at non-synonymous sites is estimated to be about 1.75, which is three times the distance between human and mouse. The two worms, which are similar at the level of ecology and morphology, show extensive identity at the level of genome organisation. The difference in size between the two genomes is accounted for almost entirely by repeat regions. Of the c. 19,500 predicted protein coding genes in *C. briggsae,* about 62% have strong one-to-one orthologues in *C. elegans*. The availability of a large orthologous gene set allows us to study how 3' UTRs and in particular, polyadenylation signals change during evolution.

5.1.2. C. elegans – C. briggsae orthologues

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

5.2. Conservation of absolute position

5.2.1. Introduction

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

5.2.2. Results

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

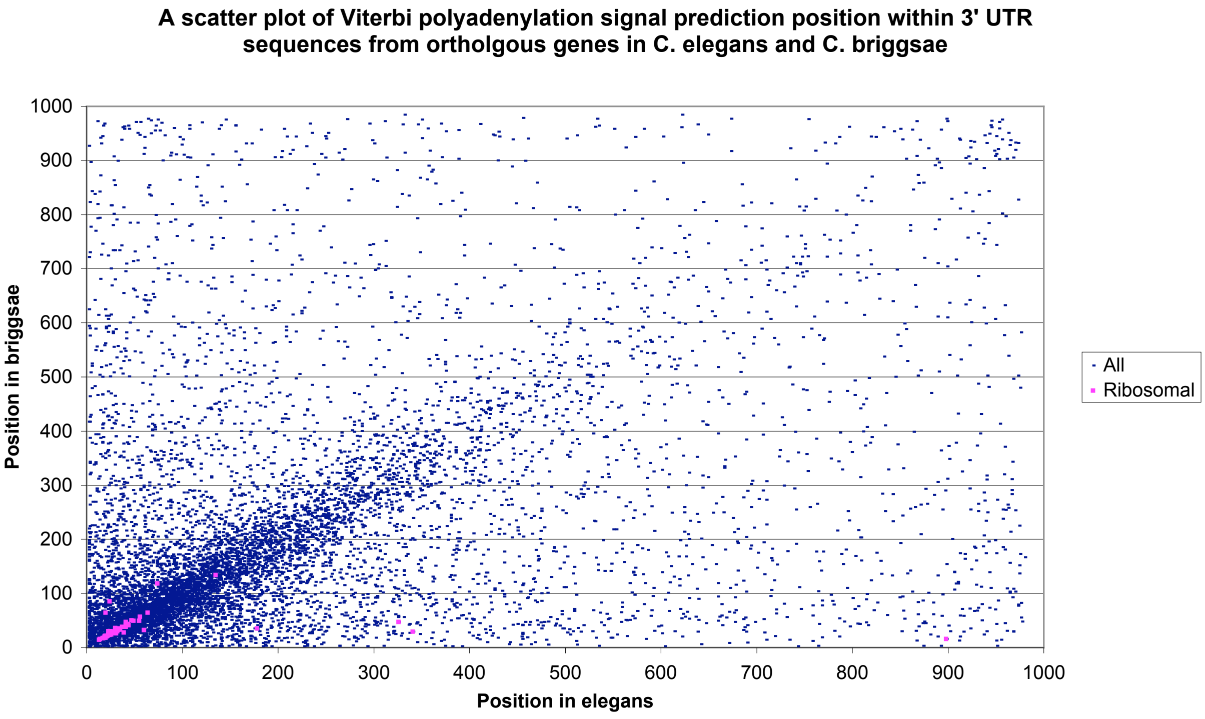


Figure 25. A scatter plot showing the absolute positions of Viterbi polyadenylation signal predictions within the 3' UTRs of 8901 pairs of orthologous elegans and briggsae genes. The 65 ribosomal protein 3'UTRs in the orthologous pair set are shown in pink.

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

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

5.3. Polyadenylation signals in aligned orthologues

5.3.1. Introduction

As mentioned in chapter 3, we have 6570 *C. elegans* 3’ UTRs, in which we have high confidence, on account of there being EST evidence for the predicted polyadenylation signal. Using this high confidence set, we can look at cases where orthologous pairs of worm 3’ UTRs can be aligned by BLAST such that the *C. elegans* polyadenylation signal is within the alignment. In these cases, we are interested in seeing whether the corresponding position in *C. briggsae* is also a likely polyadenylation signal, and if so, whether some signal variants are conserved at a higher rate than others. For example it might be that genes having the AATGAA variant do so for a specific reason, and perhaps are less likely to allow mutations which, although not knocking out the function of the signal, would change which hexamer is used.

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

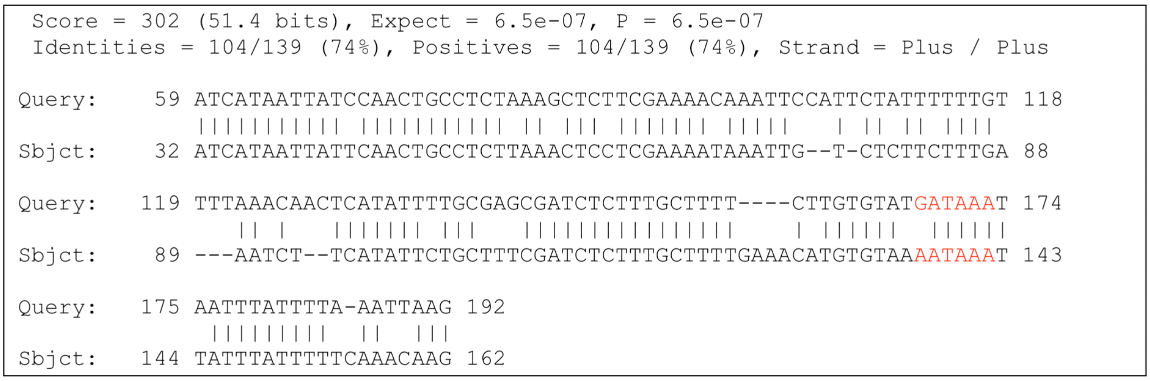
5.3.2. Alignment

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

5.3.3. Results – aligned Viterbi predictions

5.3.3.1. Position of aligned Viterbi signals

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



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

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

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

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

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

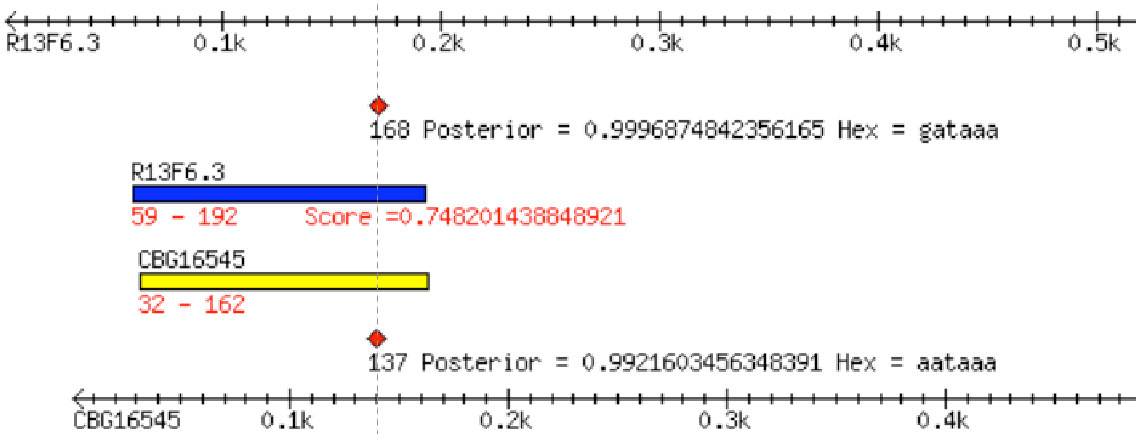


Figure 27. A representation of the sequence of a pair of orthologous 3' UTRs. From the top, the scale bar, red diamond, and blue bar refer to the C. elegans sequence, whilst the yellow bar, red diamond, and scale bar refer to C. briggsae sequence. The positions and coordinates on the blue and yellow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

5.3.3.2. Species distribution of aligned signals

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

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

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

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

5.3.3.3. Pattern of hexanucleotide mutation in aligned signals

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

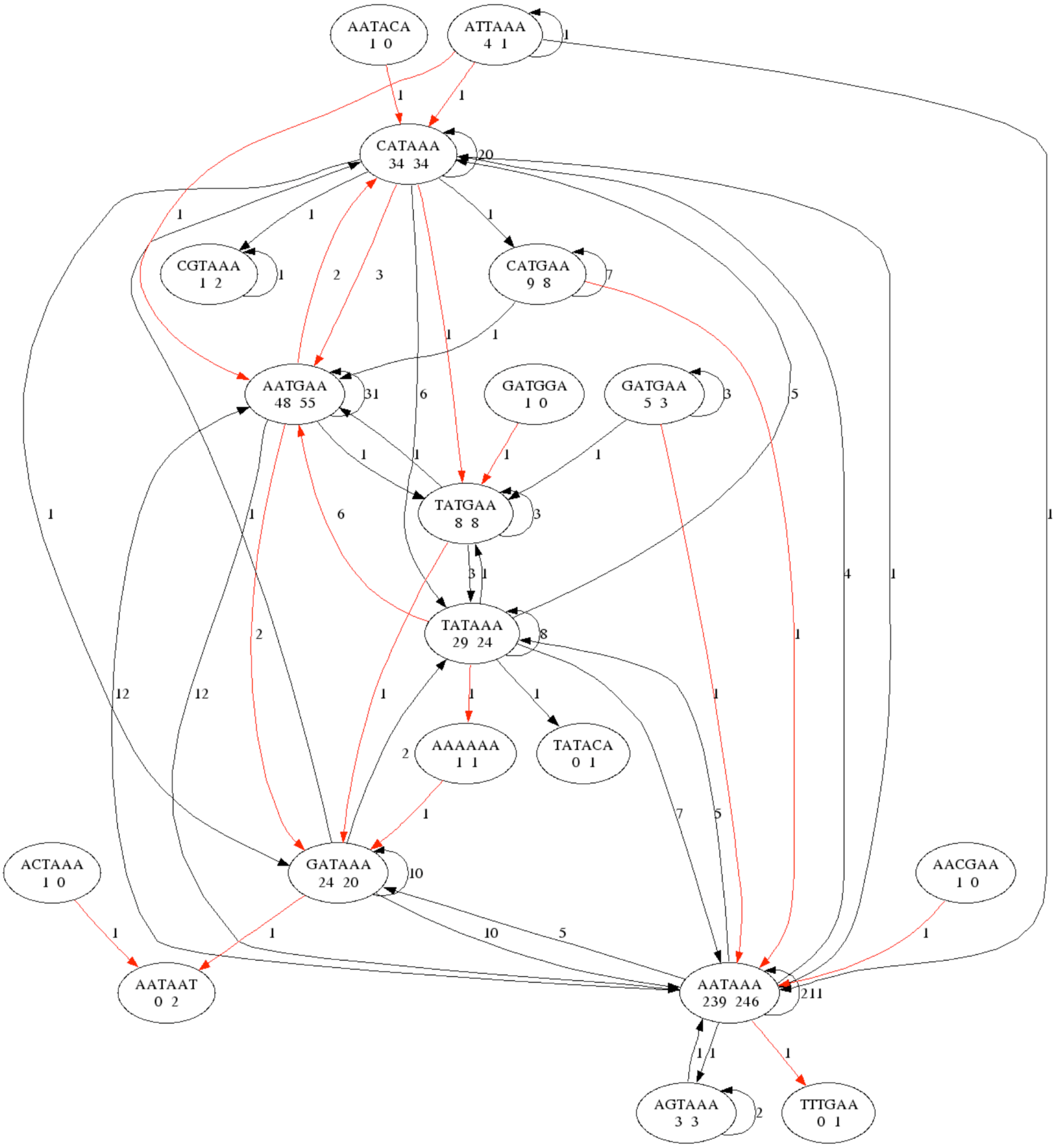


Figure 28. A graph showing the mutations between aligned AATAAA motifs in 411 orthologous worm gene pairs. Nodes represent a particular AATAAA motif. The number of signals appearing in C. elegans and C. briggsae are shown below the AATAAA motif on the left and right respectively. The number of C. elegans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

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

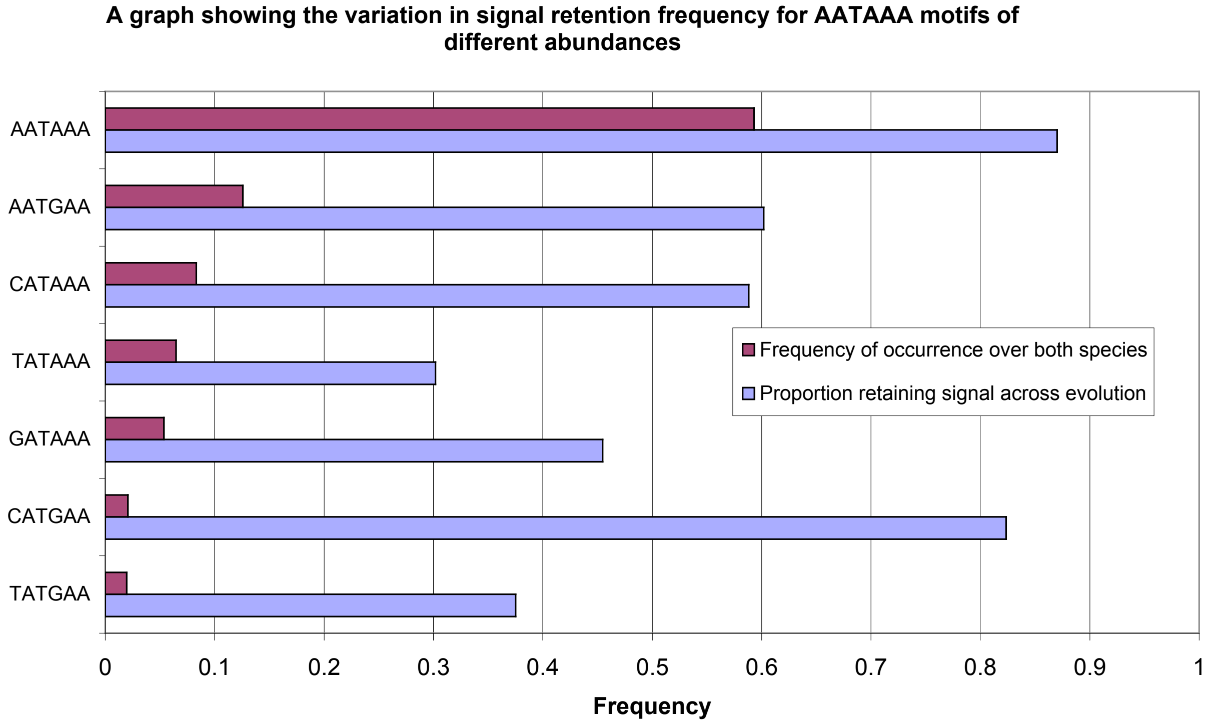


Figure 29. A graph showing the proportion of each AATAAA motif in both worms, and the proportion of each aligned signal being conserved between C. elegans and C. briggsae orthologues.

For the four most commonly used signals, the proportion of orthologous genes conserving that signal decreases with the abundance of the motif variant. 59% of the genes in the orthologous pair set use AATAAA as the polyadenylation signal. 89% of these genes' orthologues also use AATAAA. If we look instead at the 6.5 % of genes that use a TATAAA, the proportion conserving this across evolution drops to 30%. Again, just considering the top four signals, it seems that the more commonly used the motif, the more likely it is to be conserved between species. It could be that some signals, such as AATAAA, are required for some genes, whereas any variant might be tolerable for others.

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

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

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

5.3.4. Evolutionary turnover of polyadenylation signals

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

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

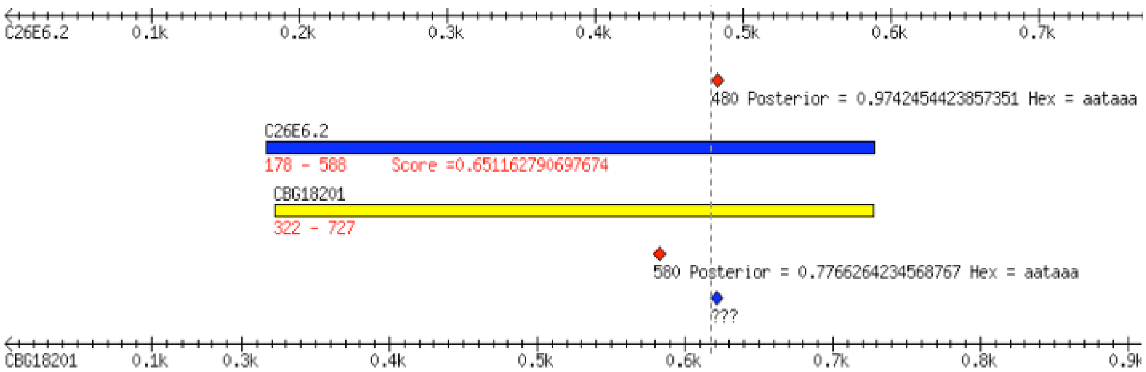


Figure 30. An alignment of two orthologous 3' UTR sequences. The extent of C. elegans and C. briggsae sequence aligned is shown by the blue and yellow bars respectively. Viterbi polyadenylation signal predictions are shown as red diamonds. The blue diamond is that part of the C. briggsae sequence, which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

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

Similarly, when comparing *C. elegans* posterior decoding ($p > 0.1$) predictions to *C. briggsae* Viterbi predictions in these 642 misaligning pairs, there are 50 cases (26 identical, 24 variant) of a *briggsae* Viterbi prediction aligning to a *C. elegans* sequence that resembles an AATAAA motif. 12 of these similar motifs are *C. elegans* posterior decoding hits. This number may be less than the 24 observed when analysing *C. briggsae* posterior decoding on account of our earlier selection against *C. elegans* genes with multiple polyadenylation sites (Chapter 3). These observations show that about half of all orthologous 3' UTRs that contain an alignment have polyadenylation signals that are either aligned to an orthologous signal (409 gene pairs), or to a sequence that may well be a real polyadenylation signal (a further 102 gene pairs, of which 36 had posterior decoding support). For the remaining 511 out of 1052 aligned 3' UTR pairs, there seems to be no sign of signal position conservation, even within the context of an alignment. These sequences have diverged so far, that meaningful polyadenylation signals have been lost.

5.4. Discussion – On the evolution of polyadenylation signals

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

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

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

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