# 6. Concerning a Sequence Element Detected in Ribosomal mRNAs

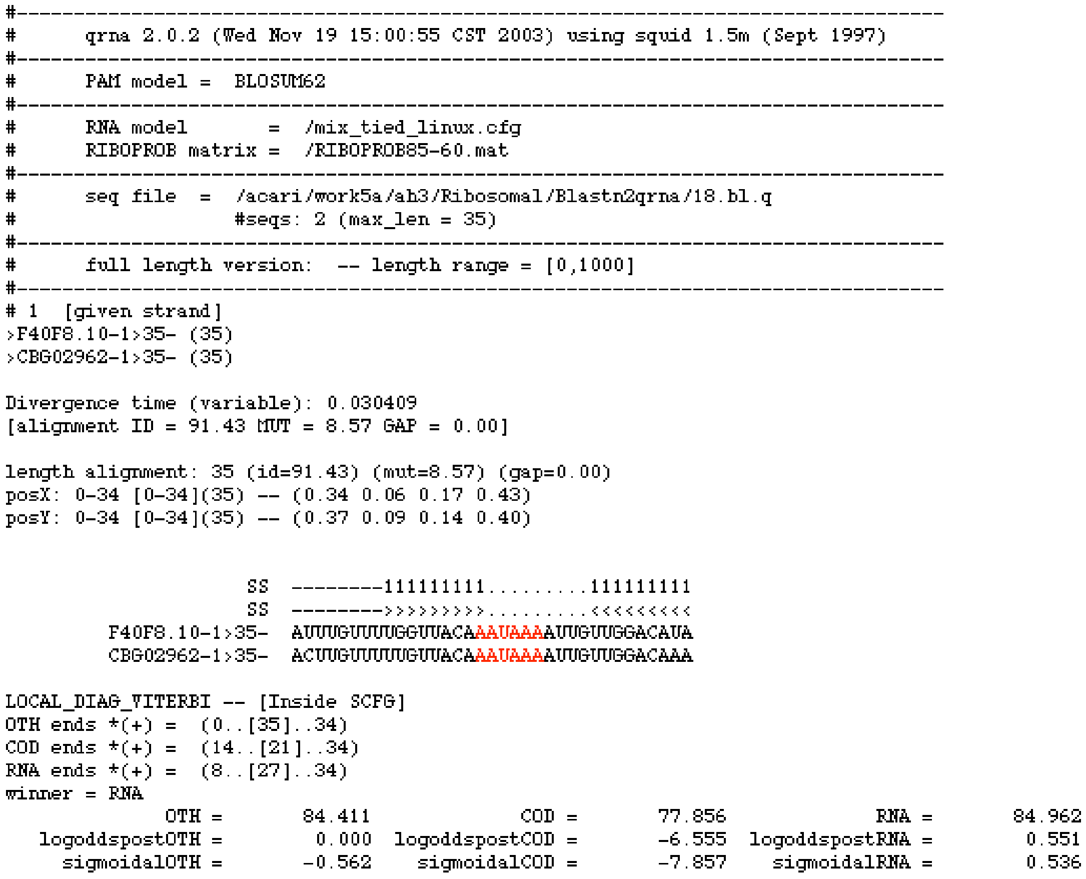
## 6.1. Introduction

Until now, this thesis has focussed on the identification of the polyadenylation signal and the end of the 3’ UTR. In this chapter, we change focus to look for other conserved signals within the 3’ UTR. In particular, we identify a region around the polyadenylation signal in many ribosomal protein mRNAs in *C. elegans* and *C. briggsae* that contains a conserved sequence motif. Building a statistical model of this motif and searching a database of *C. elegans* 3’ UTRs reveals that this motif is also present in the 3’ UTR of some other genes involved in ribosome maturation and translation.

## 6.2. Background

An initial approach that we took to identifying 3’ UTR regulatory elements was to look for conserved secondary structure components in *C. elegans* and *C. briggsae*. We took the 3’ UTRs from about 9000 *C. elegans* genes that were confirmed by ESTs and aligned them to the same length of sequence downstream of the STOP codon of the *C. briggsae* one-to-one orthologue. 6000 of these pairs generated a BLAST alignment according to our alignment parameters. The BLAST alignments were then submitted to QRNA (Rivas et al. 2001), to see if the mutations between a pair of aligned ‘orthologous 3’UTRs’ were co-varying; that is, to discover whether the sequences were evolving in such a way as to conserve a potential RNA secondary structure in an area of relatively lower primary sequence conservation.

125 of these aligned orthologous 3’ UTR pairs were considered by QRNA to contain conserved secondary structures. Of these 125, 14 alignments were from the 3’ UTRs of ribosomal proteins, an example of which is shown in Figure 31. This represents a significant overrepresentation. Further examination of the secondary structure alignments of these UTRs showed that it was unlikely that there was a single secondary structure element common to our set of ribosomal 3’ UTRs. Closer observation of the alignments suggested that in this case, there might be a conserved primary sequence which had some potential to fold into a secondary structure, though the hairpin structure itself was not being specifically conserved. Additionally, building each aligned pair into a covariance model (Eddy 2002) and searching nucleotide databanks did not indicate the presence of different, functionally conserved secondary structures.



```
#---------------------------------------------------------------------------------
#      qrna 2.0.2 (Wed Nov 19 15:00:55 CST 2003) using squid 1.5m (Sept 1997)
#---------------------------------------------------------------------------------
#      PAM model =  BLOSUM62
#---------------------------------------------------------------------------------
#      RNA model       =  /mix_tied_linux.cfg
#      RIBOPROB matrix =  /RIBOPROB85-60.mat
#---------------------------------------------------------------------------------
#      seq file  =  /acari/work5a/ah3/Ribosomal/Blastn2qrna/18.bl.q
#                   #seqs: 2 (max_len = 35)
#---------------------------------------------------------------------------------
#      full length version:  -- length range = [0,1000]
#---------------------------------------------------------------------------------
# 1  [given strand]
>F40F8.10-1>35- (35)
>CBG02962-1>35- (35)

Divergence time (variable): 0.030409
[alignment ID = 91.43 MUT = 8.57 GAP = 0.00]

length alignment: 35 (id=91.43) (mut=8.57) (gap=0.00)
posX: 0-34 [0-34](35) -- (0.34 0.06 0.17 0.43)
posY: 0-34 [0-34](35) -- (0.37 0.09 0.14 0.40)


                     SS  --------111111111.........111111111
                     SS  -------->>>>>>>>>.........<<<<<<<<<
         F40F8.10-1>35-  AUUUGUUUUGGUUACAAAUAAAAUUGUUGGACAUA
         CBG02962-1>35-  ACUUGUUUUUGUUACAAAUAAAAUUGUUGGACAAA

LOCAL_DIAG_VITERBI -- [Inside SCFG]
OTH ends *(+) =  (0..[35]..34)
COD ends *(+) =  (14..[21]..34)
RNA ends *(+) =  (8..[27]..34)
winner = RNA
              OTH =       84.411             COD =       77.856             RNA =       84.962
   logoddspostOTH =        0.000  logoddspostCOD =       -6.555  logoddspostRNA =        0.551
     sigmoidalOTH =       -0.562    sigmoidalCOD =       -7.857    sigmoidalRNA =        0.536
```

**Figure 31. Example output from QRNA when run on a 3' UTR alignment between a C. elegans ribosomal protein gene and its C. briggsae orthologue. The polyadenylation signals for the two genes are shown in red. One co-variant position is seen within the predicted secondary structure.**

### 6.2.1. Polyadenylation Signals

The area of sequence conservation was consistently situated around the polyadenylation signal as detected in our previous study on *C. elegans* polyadenylation signals (Chapter 3).

One approach to finding an unknown, but overrepresented motif or area of homology is to use expectation maximisation. Submitting *C. elegans* and *C. briggsae* ribosomal mRNA 3’ UTR sequences to the MEME program (Bailey et al. 1994) both discovered similar motifs, again based around the polyadenylation signal. Figure

32(a) shows the motif found by MEME by submitting 68 *C. elegans* ribosomal protein 3’ UTRs. The AATAAA in the centre represents a real polyadenylation signal. Figure 32(b) shows the same for 68 *C. briggsae* 3’ UTRs, whose genes are the best one-to-one orthologues of the 68 *C. elegans* genes. In contrast Figure 32(c) shows the expected base composition about 940 experimentally confirmed *C. elegans* polyadenylation signals.

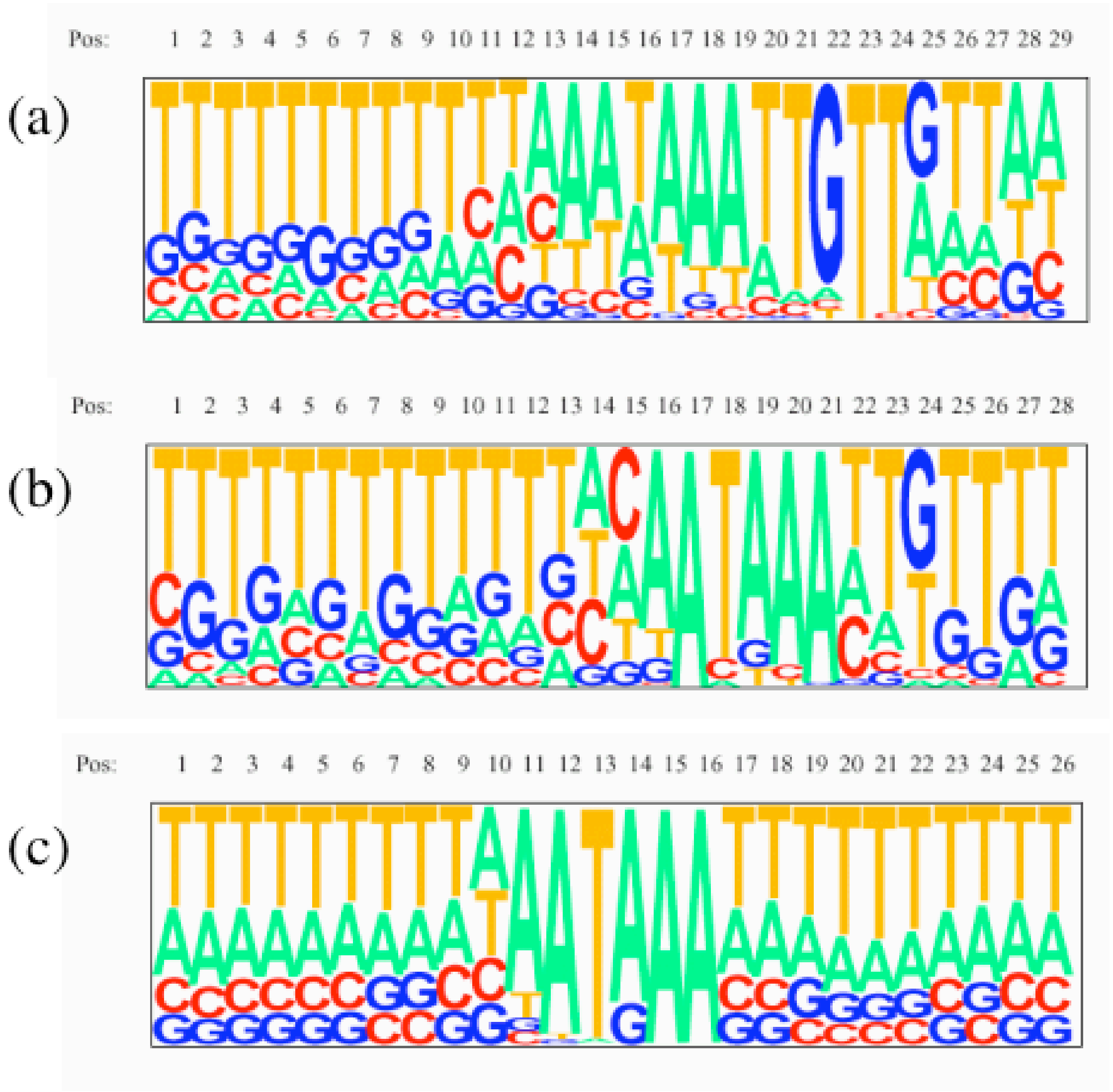


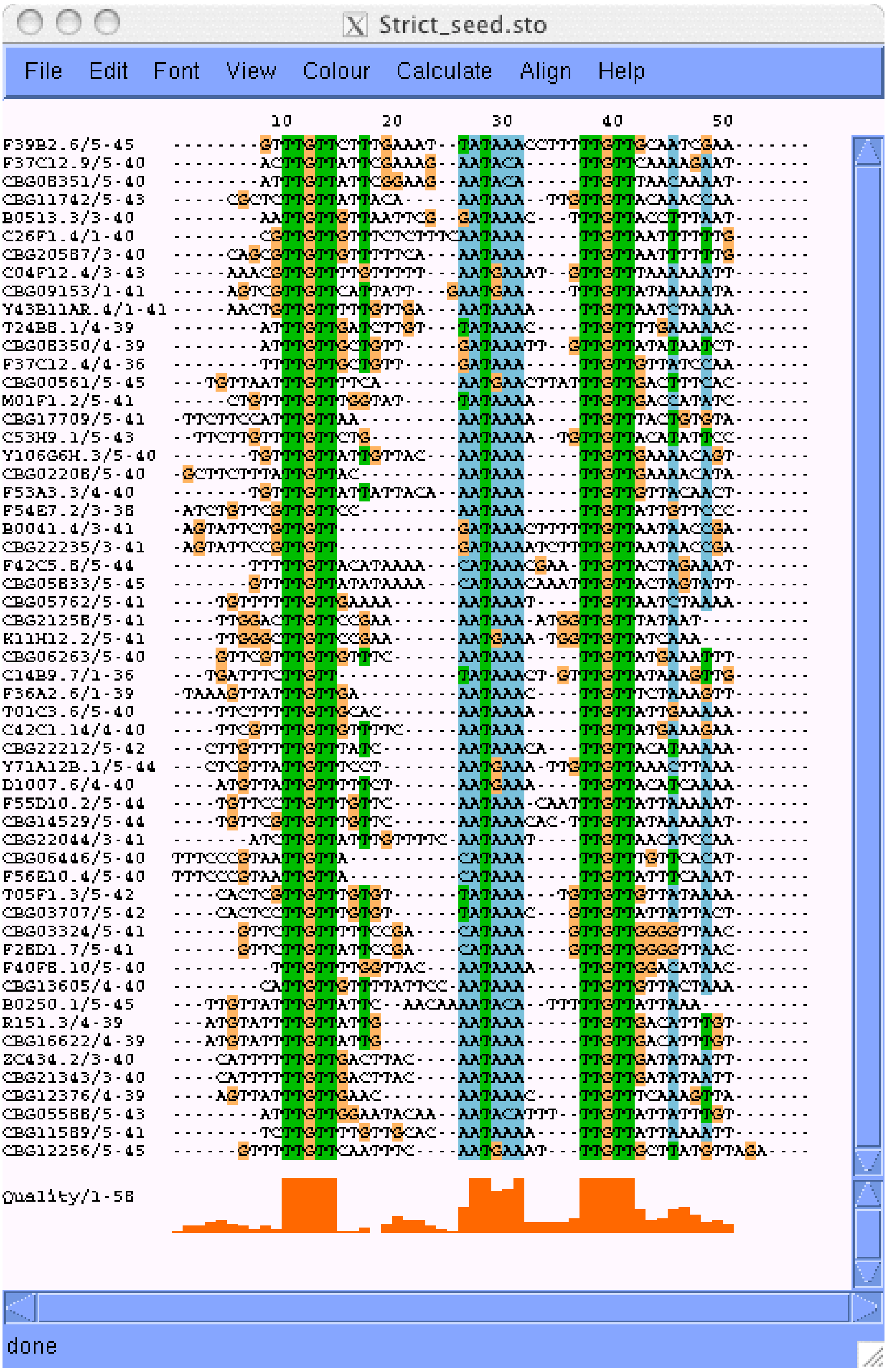
**Figure 32. The nucleotide distribution observed in the region around the polyadenylation signal in (a) 68 C. elegans ribosomal protein genes, (b) 68 C. briggsae one-to-one orthologues of the genes in (a), and (c) 940 experimentally confirmed polyadenylation signals from C. elegans.**

By observation, the sequence directly after the AATAAA motif appears different in the ribosomal mRNAs, with consensus TTGTT. The ribosomal sequences also appear to show higher than typical levels of G bases upstream of the signal, and indeed many have TTGTT, but at variable distances upstream, so the pattern is not visible in a simple alignment. We therefore conjecture that TTGTT sequences in the near neighbourhood of the polyadenylation signal may be important for ribosomal genes. We therefore decided to analyse a large set of aligned ribosomal protein 3' UTRs, anchored on the polyadenylation signal.

## 6.3. Model building

### 6.3.1. Data acquisition

One kilobase sequences representing possible 3' UTRs from 84 ribosomal proteins were extracted from WormBase (http://www.wormbase.org/). 68 of these had putative one-to-one orthologues in *C. briggsae*. Polyadenylation signal predictions (Chapter 3) were run on each sequence, and an alignment of the signal and the 20 nt flanking it on each side was forced by anchoring on the polyadenylation signal. There were 136 sequences in the alignment. The Jalview alignment viewer (Clamp et al. 2004) was used to hand-edit the alignment (Figure 33) so that TTGTT motifs either side of the polyadenylation signal were aligned. Any sequences without TTGTT in a position where it could fit in the alignment were removed. Most sequences had at least one TTGTT, but not on both sides. Some contained TTATT instead. This strict removal process left 57 sequences.



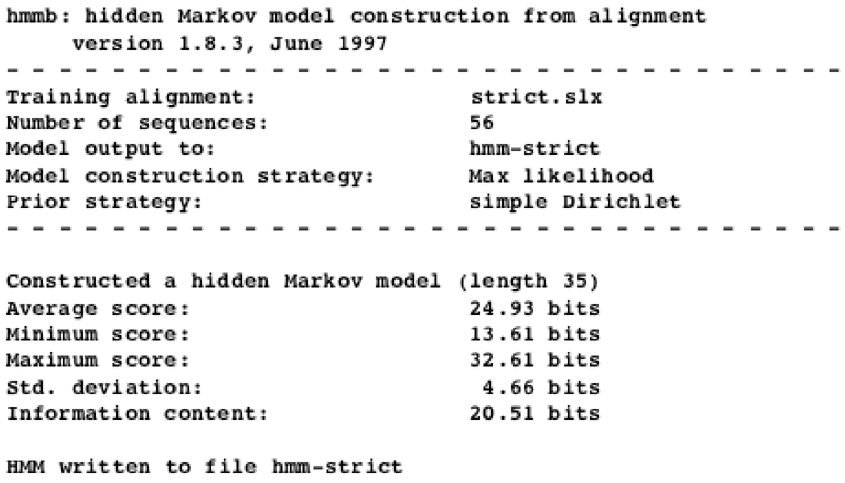
```
                        10        20        30        40        50
F39B2.6/5-45    --------GTTTGTTCTTTGAAAT--TATAAACCTTTTTGTTGCAATCGAA-------
F37C12.9/5-40   --------ACTTGTTATTCGAAAG--AATACA-----TTGTTCAAAAGAAT-------
CBG08351/5-40   --------ATTTGTTATTCGGAAG--AATACA-----TTGTTTAACAAAAT-------
CBG11742/5-43   -----CGCTCTTGTTATTACA-----AATAAA--TTGTTGTTACAAACCAA-------
B0513.3/3-40    --------AATTGTTGTTAATTCG--GATAAAC---TTTGTTACCTTTAAT-------
C26F1.4/1-40    --------CGTTGTTGTTTCTCTTTCAATAAA-----TTGTTAATTTTTTG-------
CBG20587/3-40   -----CAGCGTTGTTGTTTTTCA---AATAAA-----TTGTTAATTTTTTG-------
C04F12.4/3-43   -----AAACGTTGTTTTGTTTTT---AATGAAAT--GTTGTTTAAAAAATT-------
CBG09153/1-41   -----AGTCGTTGTTCATTATT---GAATGAA----TTTGTTATAAAAATA-------
Y43B11AR.4/1-41 -----AACTGTTGTTTTTGTTGA---AATAAAA----TTGTTAATCTAAAA-------
T24B8.1/4-39    --------ATTTGTTGATCTTGT---TATAAAC----TTGTTTTGAAAAAC-------
CBG08350/4-39   --------ATTTGTTGCTGTT-----GATAAATT--GTTGTTATATAATCT-------
F37C12.4/4-36   --------TTTTGTTGCTGTT-----GATAAA-----TTGTTGTTATCCAA-------
CBG00561/5-45   ---TGTTAATTTGTTTTCA-------AATGAACTTATTTGTTGACTTTCAC-------
M01F1.2/5-41    -----CTGTTTTGTTTGGTAT-----TATAAAA----TTGTTGACCATATC-------
CBG17709/5-41   -TTCTTCCATTTGTTAA---------AATAAAA----TTGTTTACTGTGTA-------
C53H9.1/5-43    --TTCTTGTTTTGTTCTG--------AATAAAA--TGTTGTTACATATTCC-------
Y106G6H.3/5-40  -------TGTTTGTTATTGTTAC---AATAAA-----TTGTTGAAAACAGT-------
CBG02208/5-40   -GCTTCTTTATTGTTAC---------AATAAA-----TTGTTGAAAACATA-------
F53A3.3/4-40    -------TGTTTGTTATTATTACA--AATAAA-----TTGTTGTTACAACT-------
F54E7.2/3-38    -ATCTGTTCGTTGTTCC---------AATAAA-----TTGTTATTGTTCCC-------
B0041.4/3-41    -AGTATTCTGTTGTT-----------GATAAACTTTTTTGTTAATAACCGA-------
CBG22235/3-41   -AGTATTCCGTTGTT-----------GATAAAATCTTTTGTTAATAACCGA-------
F42C5.8/5-44    -------TTTTTGTTACATAAAA---CATAAACGAA-TTGTTACTAGAAAT-------
CBG05833/5-45   -------GTTTTGTTATATAAAA---CATAAACAAATTTGTTACTAGTATT-------
CBG05762/5-41   ----TGTTTTTTGTTGAAAA------AATAAAT----TTGTTAATCTAAAA-------
CBG21258/5-41   ----TTGGACTTGTTCCGAA------AATAAA-ATGGTTGTTTATAAT----------
K11H12.2/5-41   ----TTGGGCTTGTTCCGAA------AATGAAA-TGGTTGTTATCAAA----------
CBG06263/5-40   ----GTTCGTTTGTTGTTTC------AATAAA-----TTGTTATGAAATTT-------
C14B9.7/1-36    ---TGATTTCTTGTT-----------TATAAACT-GTTTGTTATAAAGTTG-------
F36A2.6/1-39    -TAAAGTTATTTGTTGA---------AATAAAC----TTGTTTCTAAAGTT-------
T01C3.6/5-40    ----TTCTTTTTGTTGCAC-------AATAAAA----TTGTTATTGAAAAA-------
C42C1.14/4-40   ----TTCGTTTTGTTGTTTTC-----AATAAA-----TTGTTATGAAAGAA-------
CBG22212/5-42   ---CTTGTTTTTGTTTATC-------AATAAACA---TTGTTACATAAAAA-------
Y71A12B.1/5-44  ---CTCGTTATTGTTTCCT-------AATGAAA-TTGTTGTTAAACTTAAA-------
D1007.6/4-40    ----ATGTTATTGTTTTTCT------AATGAAA----TTGTTACATCAAAA-------
F55D10.2/5-44   ----TGTTCCTTGTTTGTTC------AATAAA-CAATTTGTTATTAAAAAT-------
CBG14529/5-44   ----TGTTCGTTGTTTGTTC------AATAAACAC-TTTGTTATAAAAAAT-------
CBG22044/3-41   -------ATCTTGTTATTTGTTTTC-AATAAAT----TTGTTAACATCCAA-------
CBG06446/5-40   TTTCCCGTAATTGTTA----------CATAAA-----TTGTTTGTTCACAT-------
F56E10.4/5-40   TTTCCCGTAATTGTTA----------CATAAA-----TTGTTATTTCAAAT-------
T05F1.3/5-42    ----CACTCGTTGTTTGTGT------TATAAA---TGTTGTTGTTATAAAA-------
CBG03707/5-42   ----CACTCCTTGTTTGTGT------TATAAAC---GTTGTTATTATTACT-------
CBG03324/5-41   ------GTTCTTGTTTTTCCGA----CATAAA----GTTGTTGGGGTTAAC-------
F28D1.7/5-41    ------GTTCTTGTTATTCCGA----CATAAA----GTTGTTGGGGTTAAC-------
F40F8.10/5-40   ---------TTTGTTTTGGTTAC---AATAAAA----TTGTTGGACATAAC-------
CBG13605/4-40   --------CATTGTTGTTTTATTCC-AATAAA-----TTGTTGTTACTAAA-------
B0250.1/5-45    ---TTGTTATTTGTTATTC--AACAAAATACA--TTTTTGTTATTAAA----------
R151.3/4-39     ---ATGTATTTTGTTATTG-------AATAAA-----TTGTTGACATTTGT-------
CBG16622/4-39   ---ATGTATTTTGTTATTG-------AATAAA-----TTGTTGACATTTGT-------
ZC434.2/3-40    ----CATTTTTTGTTGACTTAC----AATAAA-----TTGTTGATATAATT-------
CBG21343/3-40   ----CATTTTTTGTTGACTTAC----AATAAA-----TTGTTGATATAATT-------
CBG12376/4-39   ----AGTTATTTGTTGAAC-------AATAAAC----TTGTTTCAAAGTTA-------
CBG05588/5-43   --------ATTTGTTGGAATACAA--AATACATTT--TTGTTATTATTTGT-------
CBG11589/5-41   --------TCTTGTTTTGTTGCAC--AATAAAA----TTGTTATTAAAATT-------
CBG12256/5-45   ------GTTTTTGTTCAATTTC----AATGAAAT---TTGTTGCTTATGTTAGA----

Quality/1-58
```

**Figure 33. A hand edited alignment of the region around polyadenylation signal predictions from 57 C. elegans and C. briggsae ribosomal protein mRNAs.**

### 6.3.2. Model building with HMMER

A motif is conveniently modelled by a hidden Markov model, as it represents the region as a network of interconnected states, each with characteristic nucleotide frequencies. Variable insertion probabilities can model different motif spacings. The alignment in Figure 33 was built automatically into a hidden Markov model (Figure 34), which can capture sequence motif profiles using HMMER (http://hmmer.wustl.edu). This model was used to search a set of 22156 3' UTR candidates from *C. elegans* (that is, the 1000 bases 3' of each predicted gene's STOP codon, or the longest length up to 1000 nt before overlapping into the 3' gene.) Hits above 20 bits were reported. This generated 470 hits, of which 300 flanked a predicted polyadenylation signal. These 300 hits could be split into two groups of 150, the first containing an exact TTGTT…PolyA_Signal…TTGTT, with the other set containing at least one mismatch to one or more of the TTGTT motifs.



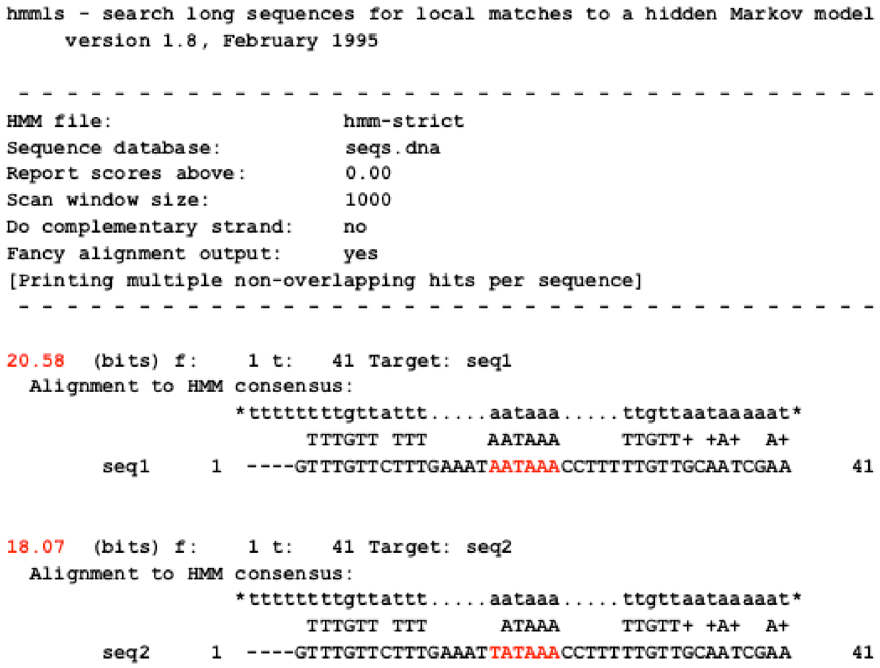
```
hmmb: hidden Markov model construction from alignment
     version 1.8.3, June 1997
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Training alignment:              strict.slx
Number of sequences:             56
Model output to:                 hmm-strict
Model construction strategy:     Max likelihood
Prior strategy:                  simple Dirichlet
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Constructed a hidden Markov model (length 35)
Average score:                   24.93 bits
Minimum score:                   13.61 bits
Maximum score:                   32.61 bits
Std. deviation:                   4.66 bits
Information content:             20.51 bits

HMM written to file hmm-strict
```

**Figure 34. Summary of the HMM built from the alignment of 57 ribosomal mRNA polyadenylation signals.**

Both sets contain hits to non-training set ribosomal protein genes, along with other genes. However, there are two potential disadvantages to this method. One is that hits containing non-canonical AATAAA polyadenylation signals are penalised, as the signal forms part of the overall motif pattern. (Figure 35) shows HMMLS hits on two sequences, which are identical apart from seq1 containing AATAAA, and seq2 TATAAA. The seq2 score is 2.5 bits lower than the seq1 score, and using a cutoff of 20 bits, seq2, which comes from WormBase CDS F39B2.6 (40S ribosomal protein S26), would be missed as a false negative.



```
hmmls - search long sequences for local matches to a hidden Markov model
     version 1.8, February 1995

 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
HMM file:                   hmm-strict
Sequence database:          seqs.dna
Report scores above:        0.00
Scan window size:           1000
Do complementary strand:    no
Fancy alignment output:     yes
[Printing multiple non-overlapping hits per sequence]
 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

20.58  (bits) f:     1 t:    41 Target: seq1
  Alignment to HMM consensus:
                   *ttttttttgttattt.....aataaa.....ttgttaataaaaat*
                         TTTGTT TTT     AATAAA     TTGTT+ +A+  A+
        seq1     1  ----GTTTGTTCTTTGAAATAATAAACCTTTTTGTTGCAATCGAA     41

18.07  (bits) f:     1 t:    41 Target: seq2
  Alignment to HMM consensus:
                   *ttttttttgttattt.....aataaa.....ttgttaataaaaat*
                         TTTGTT TTT      ATAAA     TTGTT+ +A+  A+
        seq2     1  ----GTTTGTTCTTTGAAATTATAAACCTTTTTGTTGCAATCGAA     41
```

**Figure 35. Output from HMMLS searching two sequences for hits to the HMM constructed from an alignment of ribosomal mRNA polyadenylation signals**

The other problem is that the separation of the TTGTT to the polyadenylation signal has a distinctive length distribution, as does the separation on the 3' side of the signal. HMMER does not model these two different length distributions explicitly, but rather allows hits to contain gap symbols with a penalty score, corresponding to a negative exponential distribution of gap length. The observed gap length distributions in our alignment are more flat upstream of the signal, and have a definite length preference downstream.

### 6.3.3. A more specific model

Both of the problems described above can be solved by incorporating the ribosomal motif information into a PAjHMMA model for the whole region. This 'ribosomal' model can be compared to our standard 'background' polyadenylation model to find cases which closest resemble how the TTGTT motifs flank the polyadenylation signal in the ribosomal protein mRNAs. The benefits of using PAjHMMA are that the models are polyadenylation signal-aware, unlike HMMER, and can explicitly model the observed separation between TTGTT and AATAAA motifs.

The ribosomal polyadenylation signal PAjHMMA model (Figure 36) is derived from the standard polyadenylation signal model. There are 12 additional states. TTGTT motifs (each with a state for each of the 5 columns) are inserted either side of the AATAAA motif states. The separations (from the AATAAA motif) between the upstream and downstream TTGTT motifs, are each modelled with distinctive lengths, corresponding to two more states, U and D. The ribosomal model forces each sequence to pass through both TTGTT motifs, though the two separator states can be bypassed with a probability reflecting the occurrences of upstream or downstream separator length being zero. The TTGTT motifs themselves are built empirically, scoring 1/100 for a mismatch and 97/100 for a match. In the third column, the occurrence of A is penalised to a slightly lesser degree than the others, scoring 5/100.

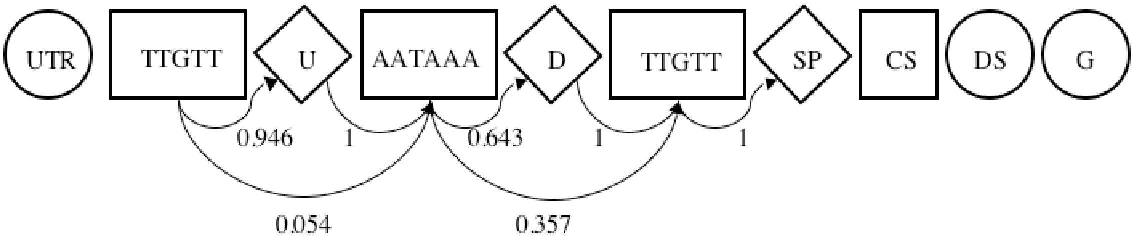


**Figure 36. State transition diagram for ribosomal polyadenylation signal model. Circular states have geometric length distributions, boxes represent weight matrices, and diamond states have explicitly modelled lengths. Where transition probabilities are not given, they are set to the same values as in the standard model given in Chapter 3.**

As discussed in Chapter 2, one of the by-products of the forward and backward algorithms is *P(x)*, the probability of the sequence given the model, or the probability that the sequence was generated by the given model. For any given sequence, we find this value given the extended ribosomal polyadenylation signal model, and the standard *C. elegans* polyadenylation signal model. The difference in the logarithms of the probabilities is a bit score measuring how well the sequence fits the ribosomal model relative to the background.

The observed length distributions upstream (Table 9) and downstream (Table 10) of the AATAAA motif are given below.

**Table 9. The length distribution observed between the upstream TTGTT motif and the polyadenylation signal from 57 ribosomal protein mRNA sequences.**

| Length i | u(i) |
|---|---|
| 0 | 0.054 |
| 1 | 0.071 |
| 2 | 0.125 |
| 3 | 0.071 |
| 4 | 0.143 |
| 5 | 0.125 |
| 6 | 0.107 |
| 7 | 0.089 |
| 8 | 0.107 |
| 9 | 0.107 |

**Table 10. The length distribution observed between the polyadenylation signal and the downstream TTGTT motif from 57 ribosomal protein mRNA sequences.**

| Length i | d(i) |
|---|---|
| 0 | 0.357 |
| 1 | 0.285 |
| 2 | 0.089 |
| 3 | 0.071 |
| 4 | 0.107 |
| 5 | 0.089 |

## 6.4. Model testing

To test whether the ribosomal model is able to differentiate ribosomal protein 3’ UTRs, bit scores relative to the standard model were found for sequences from four different sets. The four sets were:

(1) Predictions made over 22069 *C. elegans* non-ribosomal protein 3’ UTRs.

(2) Predictions from 54 *C. elegans* ribosomal sequences, that were not included in model training.

(3) Predictions made on 104 sequences of 3’ UTRs from *C. elegans*. The proteins of these genes represent the best BLASTP hit for 165 proteins from *S. cerevisiae*, that are implicated in pre-ribosomal complex formation in yeast (Fromont-Racine et al. 2003), but the set includes few ribosomal proteins.

(4) Predictions made on 63 *C. briggsae* orthologues of the 100 genes from set (1) that had the highest bit score under the ribosomal model.

## 6.5. Results

Figure 37 shows that the distributions of score for ribosomal and non-ribosomal proteins do appear to be different. The peaks in the 0 and 5 bit regions are caused by single and double mismatches respectively to TTGTT, either upstream or downstream of the polyadenylation signal. The *C. elegans* orthologues of the yeast proteins involved in ribosome assembly have a similar score distribution as the non-ribosomal protein set, but does contain a ‘shoulder’ of higher bit scores. It could be that the motif confers some function or fate involving ribosomal protein mRNAs that is distinct from ribosome assembly, and a subset of the ribosomal assembly complex have strong matches to the motif.

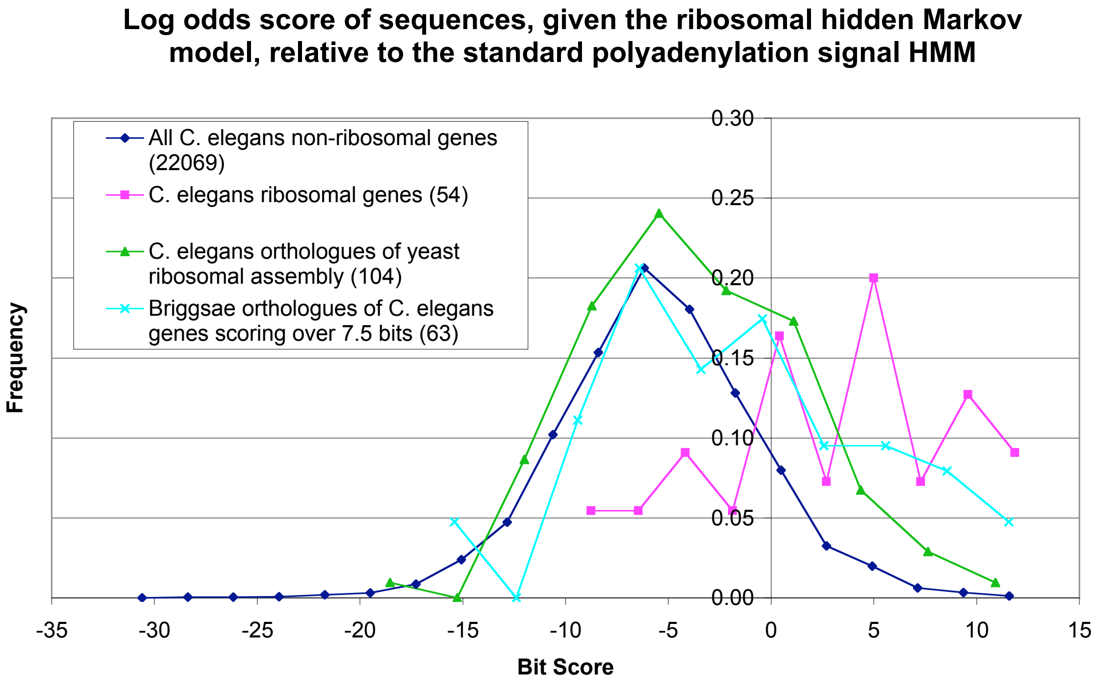


**Figure 37. The bit score histogram resulting from finding the log(2) probability of various 3' UTR sequence sets under the ribosomal model minus the log(2) probability under the standard model. Dark blue: All C. elegans non-ribosomal protein genes - this is the background distribution. Pink: C. elegans ribosomal protein genes. Green: C. elegans orthologues of yeast ribosomal assembly complex. Light blue: C. briggsae orthologues of C. elegans genes scoring over 7.5 bits.**

One hundred *C. elegans* non-ribosomal protein 3’UTRs have a bit score greater than 7.5 bits. Looking at the 63 *C. briggsae* orthologues of these high scoring *C.elegans* genes shows that the appearance of high scoring motifs in the 3’ UTR are not necessarily completely conserved between species. However, a subset (13%) of these *C. briggsae* sequences do appear to have high scores (> 5.5 bits) which are conserved. A cutoff of 5.5 bits still allows significant overlap with the score distribution of the ribosomal protein genes.

The highest scoring 100 of the non-ribosomal predictions (~0.5% of the total) all score over 7.5 bits. These 100 predictions come from genes which may therefore have some function related to that of the ribosomal protein genes. Most of these (77) have some annotation evidence, either from WormBase, or from analysis of protein domains and BLASTP homologies to better annotated proteins. Appendix I shows the set of 77 *C. elegans* polyadenylation signal predictions where the motif score was greater than 7.5 bits. Those genes thought to have some role related to that of the ribosomal proteins are marked with an asterisk. For those *C. elegans* genes with a putative *C. briggsae* orthologue, ribosomal motif log odds scores were also found for the orthologue's 3' UTR. The ranks of the log odds scores are also provided in the Appendix.

There are 22 (29% of the 77 having annotation) whose annotations confirm likely function in translation. Genes in this annotated set include 3 genes related to eukaryotic translation factors, 5 involved in tRNA synthesis and processing, and 11 contributing to ribosomal and rRNA maturation. These can be seen in Table 11.

**Table 11. A subset of the C. elegans genes having polyadenylation signals closest resembling those seen in ribosomal proteins. These have a log odds score that is within the top 0.5% of scores. These are the 22 (of 77) whose annotation suggests involvement in translation.**

| Elegans CDS | Elegans log odds score | Briggsae CDS | Briggsae log odds score | Description |
|---|---|---|---|---|
| Y48A6B.3 | 10.909 | CBG18231 | 10.620 | Contains Protein domains known in Ribosomal proteins. Similarity to L7. COG suggests Box H/ACA snoRNP component, involved in ribosomal RNA pseudouridinylation |
| F10E9.11 | 10.878 | CBG16573 | -3.314 | Similarity to elegans helicase, but also similar to Rat splicing factor and Yeast rRNA processing protein |
| F10E7.5 | 10.711 | CBG13068 | 2.064 | Similar to Ribosomal protein L-10 (may be L-10?) Similar to non-elegans ribosomal proteins. Cog suggests involved in mRNA turnover |
| W06H3.2 | 10.681 | CBG23897 | -5.100 | pus-1 encodes a putative tRNA pseudouridine synthase |
| C28H8.11a | 10.228 | no_briggsae | - | Trp dioxygenase - trp Metabolism |
| Y105E8B.7 | 9.827 | CBG19797 | -7.843 | YEATS family domain - cog suggests similarity to eukaryotic transcription factor IIF |
| ZK524.3b | 9.65 | no_briggsae | - | lrs-2 Leucyl tRNA synthetase - probably mitochondrial |
| C50F2.1 | 9.265 | no_briggsae | - | Contains ARM fold often found in RNA binding.Tranlation initiation proteins |
| T01C3.7 | 9.196 | CBG11588 | 11.559 | fib-1 Fibrillarin - nucleolar rRNA processing |
| Y45F10D.7 | 9.079 | CBG22378 | 3.040 | WD40 repeats - thought to be involved in 18S rRNA maturation |
| Y56A3A.11 | 8.807 | no_briggsae | - | tRNA splicing endonuclease |
| K07E8.7 | 8.688 | CBG19546 | 1.800 | Mitochondrial pseudouridylate synthase (RNA) |
| C01B10.8 | 8.577 | CBG05389 | 4.274 | Spermine/spermidine synthase has S-adenosyl-methione dependent methyltransferase activity |
| F28D1.8 | 8.413 | no_briggsae | - | Possible peptide-prolyl cis-trans isomerase |
| W02A11.1 | 8.096 | CBG13601 | 2.567 | Cog suggests tRNA(1-methyladenosine) methyltransferase, subunit GCD14 [KOG2915] |
| Y24D9A.4c | 7.995 | no_briggsae | - | Ribosomal protein rpl-7A/rpl-8 |
| F18A11.6 | 7.758 | no_briggsae | - | SNAP50 - Small nuclear RNA activating protein complex - 50kD subunit (SNAP50) |
| T23D8.7 | 7.734 | CBG03777 | 5.666 | High similarity to eif-2C/argonaute |
| T03F1.7 | 7.347 | CBG11970 | 1.548 | rRNA methyltransferase |
| F36A2.2 | 7.337 | CBG12371 | 8.207 | tRNA modification |
| C07E3.2 | 7.268 | CBG02729 | -4.740 | Predicted protein involved in nuclear export of pre-ribosomes |
| W04B5.4 | 7.148 | CBG15659 | -6.639 | Mitochondrial rpl-30 |

Of these 22, 15 have a *C. briggsae* orthologue. 4 of these contain a motif score of greater than 5 bits (giving significant overlap with the distribution of ribosomal genes), of which two are greater than 10 bits. The signal appears to be conserved across species in only a small number of genes. Bearing in mind the width of the distribution of the bit scores of ribosomal protein 3' UTRs (Figure 37) and the observation that many ribosomal sequences were discarded from the 136 total during model building to arrive at 57, the function, if any, provided by this motif may be highly specialised within translation.

## 6.6. Discussion

It has been observed that the regulation of synthesis of the translational apparatus is at the translational level (Meyuhas 2000). Ribosomal protein mRNAs commonly contain a 5' terminal oligopyrimidine tract (TOP) (Levy et al. 1991), which is thought to bind to La protein (Cardinali et al. 1993) with Cellular Nucleic Acid Binding Protein binding downstream (Pellizzoni et al. 1997). Subsequently, other genes involved in translation and its regulation have been found to have TOP mRNAs (Meyuhas 2000). The studies carried out in vertebrates suggest that there is a precedent for searching for some form of class-specific regulation at the mRNA level in the nematodes.

An important aspect of nematode molecular biology is the phenomenon of *trans*-splicing (Blumenthal and Steward, 1997). Approximately 70% of *C. elegans* genes are *trans*-spliced, including all but two of the ribosomal proteins. The efficiency of the *trans*-splicing reaction and the introduction of the conserved *trans*-splice leader sequence means that these genes have a very short 5’ UTR, often of just a few bases. There are only two ribosomal protein genes that do have long 5’ UTR sequences as determined by EST (Expressed Sequence Tag) alignment. A large number of the supporting ESTs start with ACTTTT, which is pyrimidine rich, and potentially a good match to the TOP sequence.

Given the lack of 5’ UTRs in many nematode ribosomal protein mRNAs, it could be that the element allowing their common control is in the 3’ UTR. Of the high-scoring set of genes observed above, it seems quite plausible that genes such as fibrillarin, which is involved in rRNA processing, should be under common control with the ribosomal protein genes. It is additionally promising that fibrillarin has the highest bit score in *C. briggsae*. The appearance in this set of some genes, which are unlikely to be involved in translation however, suggests that the motif alone may not be specific for this function.

## 6.7. Conclusions

We have seen in this chapter that some ribosomal protein genes from both *C. elegans* and *C. briggsae* contain a distinctive sequence motif around the polyadenylation signal. This motif is also found around the polyadenylation signals from other genes, some of which are known to be involved in translation.

There may be other regulatory sequence motifs related to other functions. The motif described here was found by the coordinated analysis of ribosomal protein genes; similar functional clustering has been used previously to find novel regulatory motifs, such as in histones (Dominski et al. 1999).

One suggestion for future work would be to see if this sequence motif is specific to nematodes or whether it is found in a wider range of other species. If it is only required in a subset of ribosomal protein mRNAs, it would be interesting to rationalise why this subset in particular might need some sequence motif. Another approach would be to obtain direct experimental evidence for its function.

## 6.8. Collaboration – the analysis of another 3’ UTR binding motif

I was involved in collaboration with David Bernstein from Professor Marvin Wickens’ lab at the University of Wisconsin-Madison. The work concerned an example of an evolutionarily and functionally conserved 3’ UTR motif. This is that found in genes regulated by the PUF proteins (Wickens et al. 2002). These proteins are thought to bind to the 3’ UTR of target genes, and thus repress expression by the separate mechanisms of promoting mRNA degradation or interfering with the formation of the mRNA-protein particle (mRNP). Repression by PUF proteins is particularly important during development; they are thought to maintain stem cells by preventing premature differentiation, and to repress the *C. elegans* feminine-repressor fem-3, thus permitting switching from spermatogenesis to oogenesis in hermaphrodites.

Looking for 3’ UTRs containing binding sites for PUF proteins can give an insight into the timing and targets of regulatory events in development. Although methods for identifying protein-RNA binding exist (Bernstein et al. 2002), it would be prohibitively onerous to carry out such an analysis on a whole-genome scale. Accurate computational detection of PUF protein binding sites can reduce the search space to a tractable size, and in addition, can provide independent confirmation of *in-vitro*/*vivo* work.

In a collaborative project (see Appendix II), (Bernstein et al. 2005) used mutagenesis to identify nucleotides that are essential for the binding of a *C. elegans* PUF protein, FBF-1, to a target 3' UTR. Several rounds of mutagenesis allowed the development and optimisation of binding consensus. The identification of essential "core" and influential "flanking" bases within the RNA sequence enabled us to build binding site consensus models (Dsouza et al. 1997), that constrain core residues whilst allowing for degeneracy outside the region. These were used to search against the set of *C. elegans* 3' UTRs. This computational search enabled the establishment of a set of 150 possible targets for FBF-1. In the collaborative paper, yeast three-hybrid analysis confirmed the formation of mRNA-FBF-1 complexes by 70% of a representative set of sequences from this candidate set. This shows that the computational model is a reasonable. The further analysis of the 3' UTR sequence from those genes found experimentally to have FBF-1 binding sites could be used to refine the model. This way, a combination of computational and laboratory techniques has furthered our knowledge of developmental biology. It serves also as a good example as to how genes can be co-regulated at the post-transcriptional level by a sequence motif in the 3' UTR.