Chapter 7

Classification of DNA Transposons in *Caenorhabditis elegans*

7.1 Abstract

7.2 Introduction

Transposable elements - parasitic elements, integrated into the host genome but exhibiting mobility in their genetic locus - constitute a significant fraction of eukaryotic genomes [Jur98, Smi96]. From a practical viewpoint, repetitive DNA has a detrimental effect on database searching as it spoils the assumption of sequence "randomness" on which statistical methods rely. Transposons are also emerging as players in genomic evolution, with occurences of repetitive elements reported in introns [Wes89], promoter regions [OGB95] and coding regions [BHP89]; they have also been hypothesised to trigger meiotic recombination [NCC$^+$92, YWB97, FBT$^+$91]. Transposons provide useful vectors for germline transformation [PvL97]. As model selfish genetic elements, their population dynamics is an interesting topic [YWB97, LC97, HLL97, HLNL97].

Transposons are distinguished from other types of repetitive DNA - such as microsatellite repeats and unique tandem duplications - by their spontaneous re-insertion at new positions in the genome. In doing this they pass through an intermediate phase either as RNA which is then reverse-transcribed back into the genome, or as double-stranded DNA which is excised and re-integrated elsewhere. The two kinds of transposon are described as "class I" or "class II". Both classes can be further categorised according to whether they are autonomous or non-autonomous: transposons in the former category contain genes coding for all the transposase proteins necessary for mobilising transposition, whereas those in the latter, non-autonomous category depend on enzymes provided by the former category and are often nicknamed "hitch-hikers" [Smi96]. Hitch-hikers may be closely related to autonomous elements by mis-sense mutations or may display similarity restricted to the tranposase binding sites [RvLDP97, OGB96].

The presence of parasitic hitch-hikers is posited to be detrimental to the reproductive success of transposons, especially so for DNA tranposons, whose tranposase proteins may have a more difficult job finding the particular sequences from which they were transcribed, leaving them vulnerable to parasitic mimics [HLL97, LC97, HLNL97]. Two mechanisms by which a transposon may avoid becoming overburdened with hitch-hikers include: (1) evolution of new specificity in its transposase-nucleic acid interactions; (2) invasion of fresh host genomes that are free of hitch-hikers. It may be envisaged that these work in tandem, i.e. new genomes provide the spatial heterogeneity necessary for new specificity to evolve. The presence of hitch-hikers is just one factor proposed to restrict the mobility of transposons; others include DNA methylation [YWB97] (though this is absent in *C.elegans*), self-inhibition [LC97] and titration by defective transposase proteins [HLL97].

One of the most widely studied families of DNA-mediated transposable elements is the Tcl/*mariner* family [PvL97, HLL97]. Members of this ubiquitous family typically contain a two-exon gene of around 300-400 codons flanked by short (11-80bp) terminal inverted repeats (invreps). The Tcl transposase, which has been demonstrated to be sufficient to mediate transposition in *C.elegans* [VBP96], catalyses the staggered double-strand endonuclease cleavage of the DNA substrate and re-integration of the transposon into the sequence TA [HLL97, Cra95, LCR96, vLCP94, VBP96]. Tcl excision is followed by double-stranded DNA breakage repair, which can entail a variety of mutations including deletions, insertions and duplications [MKW91]. The putative domain structure of the Tcl transposase is shown in Figure 7.1. Three domains have been proposed [VvLP93]: (i) a specific DNA-binding domain that binds between bases 5 and 26 of the Tcl invrep and shows weak transitive homology to the DNA-binding domain of the Drosophila *paired* gene, a transcription factor involved in embryonic development [FLD$^+$94, GW92]; (ii) a non-specific DNA-binding domain that might be responsible for DNA-protein interactions determining the structure of the transpososome [VvLP93]; and (iii) a catalytic domain that belongs to the D35E superfamily of transposases and retroviral integrases, the struc-

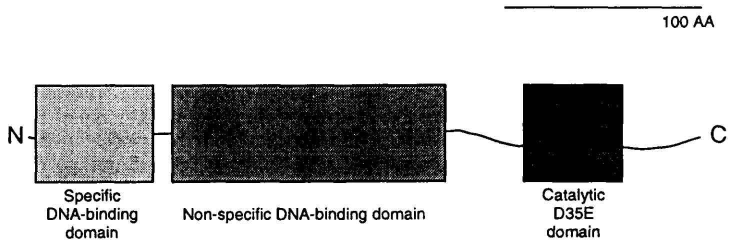

Figure 7.1: The putative domain structure of the 343-amino acid Tcl transposase protein. The 63 N-terminal residues bind specifically between bases 5 and 26 of the Tcl terminal invreps. The corresponding domain in the *Minos* elements from *Drosophila hydei* show weak homology to the *paired* gene in *Drosophila* [FLD+94]. Amino acids 71 to 203 contain a non-specific DNA-binding domain [VvLP93] and amino acids 247 to 296 are the D35E catalytic motif that is highly conserved in a number of transposase and viral integrase proteins [DDJH94].

tures of several members of which have been solved [DDJH94, GL95a]. There may be additional, cryptic DNA-protein interactions affecting transposase activity [VvLP93]. It is thought that the terminal 6 bases of the Tcl invrep are important for catalysis [VP94]. The Tc3 transposase has a similar catalytic mechanism but binds to two regions in the (longer) Tc3 invrep, rather than one [CvLP94].

Computational analyses based on the clustering of inverted repeats have uncovered several putative families of transposable elements in various organisms [OGB95, OGB96, Smi96] including the partially sequenced genome of the Bristol N2 strain of *C.elegans* [SDT+92]. At least one of these families has since been demonstrated to be mobile [RvLDP97]. In this chapter, a comprehensive list of all previously characterised transposons (including those described above) in the *C.elegans* genome is first presented. A computational approach to identifying new members of a DNA transposon family is next described; this approach is used to identify several previously uncharacterised subfamilies of the Tcl/*mariner* group. Phylogenetic evidence suggests that at least one of these families may have been active in the recent past. Profile hidden Markov models [DEKM98] of the inverted repeat sequences characterising the new families are published as part of the *C.elegans* annotation [CSC98] and in the Wormdup release along with annotation files describing the locations of identified elements.

7.3 Methods

The basis for the present analysis was the 90Mb of *C.elegans* DNA sequence available as of October 1998, together with the published annotation in ACeDB [ED95] and the CeRep database of common worm repeats [CSC98].

7.3.1 Construction of the transposon family data set

Three principal techniques were used in constructing the data set of novel transposable elements:

- the identification of inverted repeats (following [OGB95]),
- sequence homology and clustering at the protein coding level (following [Smi96]) and
- sequence homology and clustering at the DNA level.

Inverted repeats

A list of all invreps was constructed by screening each cosmid against itself using `blastn` version 1.4.7 [AGM$^+$90]. The cosmid-by-cosmid approach introduces a coarse-graining over the ideal approach of comparing chromosomes whole; however, the separation of most *C.elegans* transposon invreps is considerably less (around 3kb) than the typical length of a cosmid (around 40kb) and the artificial length cutoff introduced should be negligible; this differs from the approach in [OGB95]. Before performing the BLAST search, low-complexity regions (using the `cfilter.pl` program described in Appendix A) and tandem repeats (using the `tandem` program from the GCG package) were identifed and masked out.

The invrep data were reduced by a factor of approximately two by taking the closest invrep-pair wherever a conflict arose, using the `gffintersect.pl` and `intersectlookup.pl` programs described in Appendix A. This left approximately 72000 invreps.

Protein sequence homology

To reduce the size of the invrep list, homology information was used to restrict the search to elements encoding a transposase protein of the D35E superfamily. This family is widely diverged [DDJH94] and it is anticipated that there will also be pseudogene-containing variants with internal deletions or insertions [OGB96]. For these reasons the **blastx** program, which searches the six-frame conceptual translation of the genomic sequence, may not be sensitive enough. A more sensitive program is GeneWise [BD97], which finds the optimal alignment of conceptually translated genomic sequence to a hidden Markov model (HMM) and is robust to gaps and frameshifts. HMMs were trained individually on three separate seed alignments: the first produced using CLUSTALW [THG94a] from all the *mariner* transposases in SP-TREMBL and the latter two derived from previous analyses of D35E subfamilies [DDJH94, SR96] and homologous sequences in SP-TREMBL. The shortest seed alignment is shown in Figure 7.2.

The available worm DNA was searched with these HMMs using GeneWise, and the matches combined with the list of invreps by dynamic programming using the `gffdp.pl` program described in Appendix A, to yield a set of predicted transposons. This set was partitioned into single-linkage clusters by flanking invrep sequence similarity using the `seqcluster.pl` program described in Appendix A.

DNA sequence homology

From the transposon family data set, a set of canonical invrep sequences for each family was extracted. Each set was used to train an HMM, which was then searched against the available worm DNA in order to obtain comprehensive data

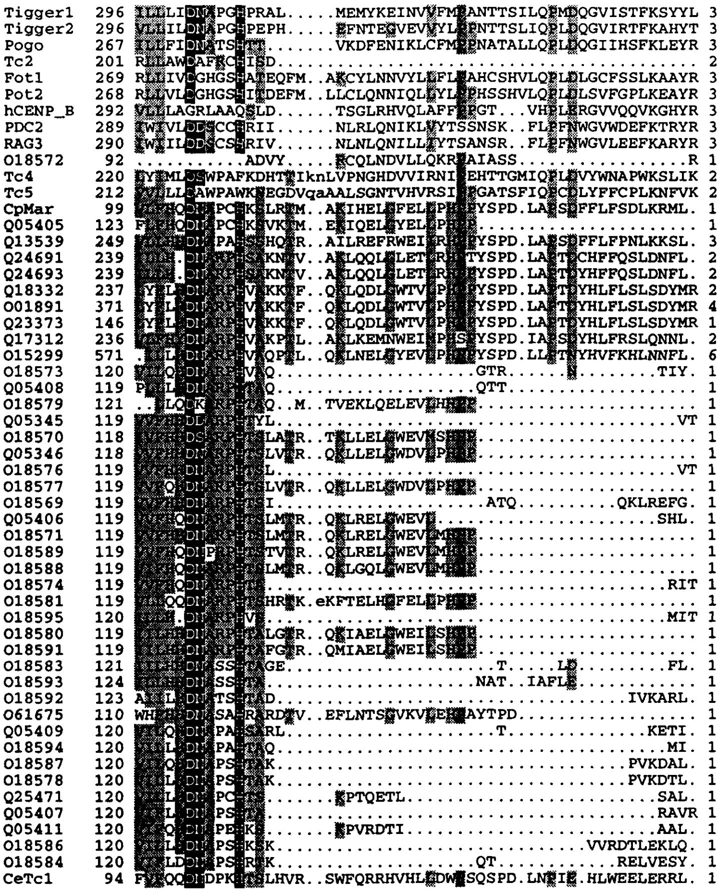

Figure 7.2: Alignment of D35E motifs from [DDJH94] and similar proteins from SP-TREMBL.

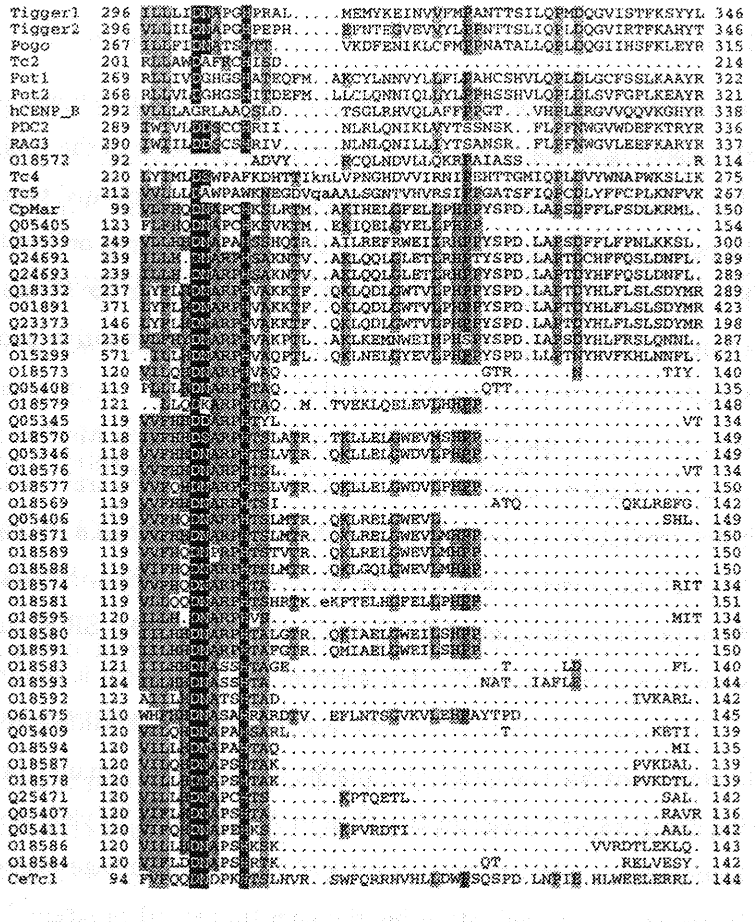

Figure 7.2: Alignment of D35E motifs from [DDJH94] and similar proteins from SP-TREMBL.

on the representation of each family in the worm genome, including instances where one half of the invrep had been deleted. HMMs trained on sequences from previously described transposon families (including Tcl-Tc7 [PvL97, RvLDP97] and Celel-Cele7 [OGB95]) were also searched against the worm genome. Invreps were paired together and associated with GeneWise-predicted transposase genes using the GFF dynamic programming software `gffdp.pl` described in Appendix A.

7.3.2 Analysis of the transposon family data set

For each autonomous transposon family, a multiple alignment of the predicted transposase genes was made using CLUSTALW. These alignments were phylogenetically analysed by the UPGMA method using BELVU [SD94].

7.4 Results

7.4.1 Previously characterised transposon families

Table 7.1 lists the results of searching the *C.elegans* DNA with HMMs constructed from inverted repeat sequences typical to known transposon families Tcl-Tc7 [PvL97, RvLDP97], Celel-Cele7 [OGB95] and Cele11-Cele14 [OGB96].

It is interesting to compare the results of this computational analysis with the element counts predicted from experimental data [PvL97]. The only element whose count is lower than experimentally predicted is Tcl. If a direct `blastn` search is performed using the Tcl sequence as a query, the higher, predicted count is obtained. A possible explanation for this is that the "missing" Tcl sequences do not fit the pattern of transposase homology flanked by invreps. Closer inspection reveals this to be so: in most cases, one of the invrep sequences is missing and in one case (*C.elegans* cosmid T10B5, invreps start at 37632) both invreps are present but do not flank the (usually) internal sequence (adjacent at position 39465).

Name	Example invrep sequence	Copies		Typical length	
		Pairs	Single invreps	Invrep only	Whole element
Tc1 (‡)	TACAGTGCTGGCCAAAAAGA...	25	10	81	1620
Tc2 (†♠)	CCGTATATTCTCTATTAGTG...	49	108	24	120
Tc3 (†♣)	TACAGTGTGGGAAAGTTCTA...	28	21	469	2350
Tc4 (†§)	CTAGGGAATGACCAGAATAA...	20	6	139	1610
Tc5 (¶)	CAAGGGAAGTCAAAAAACTG...	50	29	137	640
Tc6	CAGTGCTCCACATAATGATA...	22	886	656	1610
Tc7	TACAGTGCTGGCCAAAAAGA...	54	67	346	930
Cele1	CAAAATATCTCGTAGCGAAA...	73	280	36	230
Cele2	TACCHGGTCTCGACACGACA...	141	464	85	260
Cele4	TGGGTCTCGTTAGGTATTHG...	43	163	37	150
Cele5	GGTCTCGAAACGAYYGAAAY...	5	37	37	200
Cele6	TATTAMGRRAHCAHNARWTC...	19	42	32	150
Cele7	TAGTGHNAAANTATAGAAAA...	33	83	66	150
Cele14 (†)	CACGTGGAGTCAAAAAGTCC...	669	1095	36	180

Table 7.1: Previously characterised transposon families in the worm genome. Notes: (†) more copies than predicted [PvL97, OGB96]; (‡) 22 of the pairs enclose a transposase with 2 exons, lengths 155/875bp; (♠) includes 3 Cele11 and 32 Cele12 elements described in [OGB96], blastn searches reveal an additional 9 Cele11 and 7 Cele12 elements (approx.); (♣) 14 pairs enclose a transposase with 2 exons, lengths 416/572bp, 3 pairs form a putative nonautonomous subfamily, the rest appear internally heterogenous; (§) includes 4 copies of the putatively autonomous element Tc4v (3kb long); (¶) includes 20 copies of 1400bp and 25 copies of 600bp variants described in [OGB96], only 4 copies are "genuine" Tc5.

In all other cases, the database searches find about as many copies as experimentally predicted, with the exceptions of Tc2, Tc3, Tc5 and Cele14, whose copy numbers are elevated. In the former three cases this is due to the presence of putatively nonautonomous families sharing homology with the named families in the terminal regions of the flanking inverted repeats. The families associated with Tc2 and Tc5 have been previously described [OGB96], but the Tc3-associated family is new. Tc3 has been predicted to occur approximately 15 times in the Bristol N2 strain of *C.elegans* [CFA89] and 14 transposase-carrying copies are indeed found in this search; however, this only accounts for half the paired hits to the invrep HMM. Three of the remaining 14 pairs were found to share strong (over 90%) internal sequence identity, forming a new family of 1400bp proposed Tc3-hitchhikers with 574bp invreps, the terminal 247bp of which are similar to the Tc3 invrep. No strong internal similarity between the other Tc3-like elements was found.

The number of copies of the Cele14 invrep is an order of magnitude greater than predicted in [OGB96], probably because of the increased sensitivity of an HMM-based search over a BLAST search.

7.4.2 Previously uncharacterised transposon families

The search procedure described in 7.3.1 revealed six new Tc1/*mariner*-like families of transposon, named Tc11-Tc16 (this continues the Tc naming convention but leaves Tc8-Tc10 unused, allowing for independent transposon discoveries). Tc11-Tc16 contain coding sequences homologous to the *mariner* transposase flanked by characteristic inverted repeats. The definition of a family that was used - a group of transposons with near-identical invrep sequences - was supported by the phylogeny of the genes bracketed by these invreps, which clustered in the same way as the invrep sequences. Representation data for these transposons are listed in Table 7.2.

The exon structure of the predicted Tc11-Tc16 transposase genes in *C.elegans*

Name	Example invrep sequence	Copies		Typical length	
		Pairs (coding)	Single invreps	Invrep only	Whole element
Tc11 (†)	`TATTAGGTTGAACCGGAAGT...`	24 (11)	14	34	1230
Tc12 (‡)	`TATTAGGTTGGTCGAAAAGT...`	36 (19)	17	34	1250
Tc13 (♠)	`TATCAGGTCGTCCCATAAGT...`	59 (33)	16	34	1240
Tc14	`TACAGGGTGAGTCAAAATTA...`	12 (6)	47	30	1290
Tc15	`CTCGGCAATTCGTATCGTAC...`	4 (1)	7	40	1110
Tc16	`TATTAGGTTGTGAAAAAAGT...`	4 (2)	2	33	1260

Table 7.2: Previously uncharacterised Tc1/*mariner*-like transposon families in the worm genome. The numbers of coding-sequence containing pairs shown in brackets in the third column are based on the conservative *C.elegans* annotation rather than the Genewise predictions. Notes: (†) invrep similar to Tc13 and very similar to Tc12; (‡) invrep similar to Tc13 and very similar to Tc11; (♠) invrep similar to Tc11 and Tc12, tree suggests recent dispersion.

varies. The tranposase is most often predicted as a single exon over 1000bp long. None of these families has been characterised in the literature, although Oosumi *et al* found several copies of Tc13 after a `blastn` search with Cele14 as a probe [OGB96].

The chromosomal loci of the members of the transposon families listed in Tables 7.1 and 7.2 are published on the Wormdup website at the following URL:

`http://www.sanger.ac.uk/Users/ihh/Wormdup/`

7.4.3 Variation between transposon families

Transposon families Tc11-Tc13 display considerable similarity in their invrep sequences. This contrasts with previously described transposon families in *C.elegans*, which form distinct groups whether clustered by invrep or by internal sequence similarity. In particular the terminal 6 bases of the Tc11, Tc12, Tc13 and Tc16 invreps are almost completely conserved. These bases are thought to be important for catalysis in Tc1 [VP94].

Figure 7.3 shows an alignment of representative transposase proteins from the Tc1, Tc3 and Tc11-Tc15 families (the Tc16 coding sequence did not align

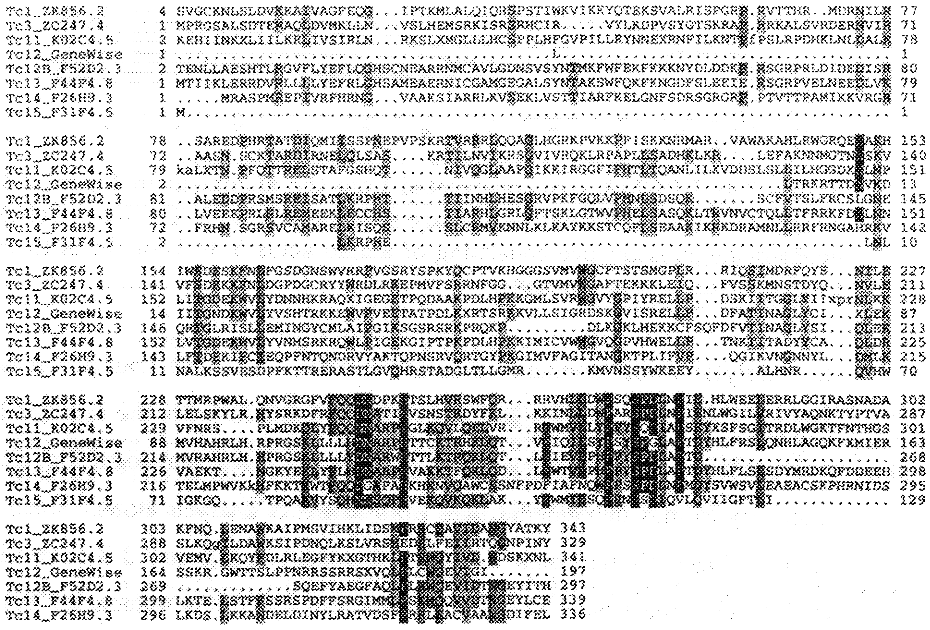

Figure 7.3: Alignment of transposase proteins from the Tc1, Tc3 and Tc11-Tc15 families. This alignment was constructed by removing redundant sequences from a multiple alignment of all predicted tranposase genes in *C.elegans* produced by CLUSTALW [THG94a]. Two distinct variants of the Tc12 transposase (Tc12 and Tc12B) are included. Residues 1 to 63 of Tc1 contain the *paired* DNA-binding domain, residues 71 to 207 contain the non-specific DNA-binding domain and residues 247-282 contain the D35E domain. Note that the Tc12 transposase gene was predicted using a version of GeneWise that does not extend the gene prediction beyond the optimal match to the HMM [BD97].

well to this set and was excluded). Two variants of the Tc12 protein appear (labelled Tc12 and Tc12B), as it was observed that this family forms two distinct subgroups when clustered by coding sequence. Where possible, the gene predictions from the ACeDB annotation were used in preference to the GeneWise predictions, as they tended to be more complete (GeneWise currently does not predict exons outside the region of homology).

Although the proteins are divergent, with the closest neighbours (the two Tc12 transposases) sharing under 50% sequence identity, they are also clearly homologous. Furthermore, the sequence homology is considerably greater over

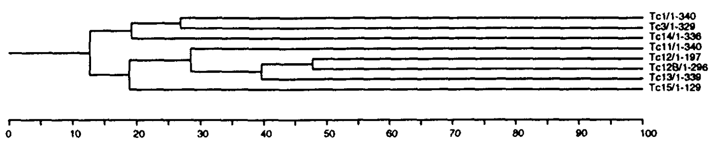

Figure 7.4: UPGMA tree constructed from the alignment in Figure 7.3. The horizontal scale marks out percentage sequence identity. The label TcN.X/Y-Z denotes the subsequence consisting of amino acids Y to Z of the X'th copy of the TcN transposase protein.

the catalytic D35E domain. This is consistent with the transposon families sharing a catalytic mechanism. It is also consistent with the families having evolved different invrep sequences for specific transposase recognition, although specificity cannot be demonstrated from sequence analysis alone.

A phylogenetic tree built from the alignment in Figure 7.3 is shown in Figure 7.4. The tree groups Tcl5 with Tcll-Tcl3 and (more tenuously) Tcl4 with Tcl/Tc3.

7.4.4 Variation within transposon families

The Tcl, Tc3, Tcll, Tcl2 and Tcl3 families are sufficiently numerous that the intra-family variation - that is, the variation between coding sequences for members of the same family - can also be analysed (Figures 7.5 and 7.6). Several interesting points emerge from a study of these trees. All of the trees tend to be skewed, favouring the view that most new duplicates of an element are inactive, doomed to accumulate mutations while transposition is dominated by a few active copies [YWB97, HLL97, LC97, HLNL97]. The short branch lengths of the Tcl, Tc3 and Tcl3 trees are evidence that these elements have been active in recent history (indeed, Tcl and Tc3 are known to be currently active in the Bristol N2 strain [PvL97]), whereas the longer branch lengths of Tcll and Tcl2 suggest that they ceased activity earlier. It can also be seen that Tcl2B subgroup of Tcl2 transposases form a distinct group, as mentioned above.

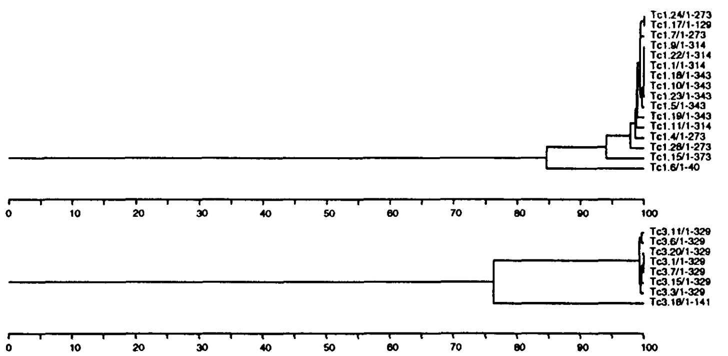

Figure 7.5: UPGMA trees constructed from alignments of the Tcl (top) and Tc3 (bottom) transposases. The horizontal scale marks out percentage sequence identity. The label TcN.X/Y-Z denotes the subsequence consisting of amino acids Y to Z of the X'th copy of the TcN transposase protein.

7.4.5 Location of transposons within the *C.elegans* genome

The distribution of transposons within the genome is of interest, not only because the insertion of a transposon into a gene or regulatory sequence can disrupt its function [Wes89, OGB95, BHP89], but also because the presence of transposons has been suggested to precipitate meiotic recombination [NCC$^+$92, YWB97]. The present study finds no evidence that the chromosomal location of a transposon is correlated with that of its nearest intra-familial relative. However, significant numbers of transposons were found within coding sequences and 5' upstream regions (Table 7.3). The high number of Tcl and Tcl3 elements overlapping with exons may be due to mispredicted genes in the *C.elegans* database.

The total fraction of transposons in or near coding sequences (68%) is higher than the proportion expected by chance (55%). DNA transposons thus display a clear preference for coding sequence in their choice of integration site.

Different types of repeats are often found to be associated together [Jur, PvL97]. As part of the preliminary screen for repetitive elements, the `gfffilter.pl` and `gffintersect.pl` programs (Appendix A) were used to find

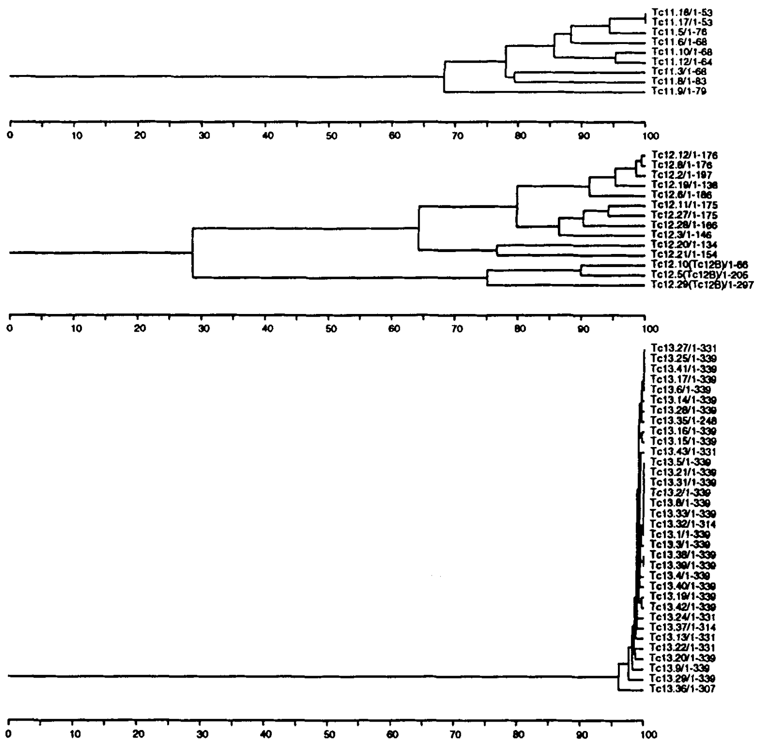

Figure 7.6: UPGMA trees constructed from alignments of the Tc11 (top), Tc12 (middle) and Tc13 (bottom) transposases. The horizontal scale marks out percentage sequence identity. The label TcN.X/Y-Z denotes the subsequence consisting of amino acids Y to Z of the X'th copy of the TcN transposase protein.

Transposon family	Copies in exons	Copies in introns	Copies in 1kb 5' regions
Tc1	5 (15%)	2 (6%)	19 (59%)
Tc2	0 (0%)	15 (37%)	8 (20%)
Tc3	1 (5%)	2 (10%)	12 (60%)
Tc4	0 (0%)	4 (28%)	3 (21%)
Tc5	0 (0%)	12 (30%)	14 (35%)
Tc6	1 (6%)	2 (12%)	1 (6%)
Tc7	1 (2%)	11 (26%)	7 (16%)
Cele1	1 (2%)	16 (38%)	10 (23%)
Cele2	1 (0%)	48 (42%)	37 (32%)
Cele4	0 (0%)	16 (53%)	6 (20%)
Cele5	0 (0%)	3 (60%)	2 (40%)
Cele6	1 (6%)	5 (31%)	3 (18%)
Cele7	2 (8%)	13 (54%)	7 (29%)
Cele11	0 (0%)	4 (8%)	16 (32%)
Cele12	1 (2%)	3 (7%)	10 (25%)
Cele14	16 (3%)	190 (37%)	133 (26%)
Tc11	3 (14%)	2 (9%)	14 (66%)
Tc12	4 (13%)	1 (3%)	19 (63%)
Tc13	13 (30%)	0 (0%)	24 (55%)
Tc14	2 (22%)	0 (0%)	6 (66%)
Tc15	1 (25%)	1 (25%)	2 (50%)
Tc16	1 (33%)	0 (0%)	2 (66%)

Table 7.3: Proximity of transposon families to coding sequence. The percentages in brackets indicate the fraction of the total copy number in each category.

the propensities of different repeats to associate with one another. An *association score* $\log\left[\frac{f_{xy}f}{f_x f_y}\right]$ (where f_{xy} is the frequency with which repeat x is associated with repeat y, f_x is the frequency with which x is associated with any other repeat and f is the total number of associations) was calculated for every pair of repeats x and y; some pairs of repeats with scores over 10 bits are listed in Table 7.4. There are clear clusters of repeats that are often found together, for example CeRep43, CeRep34 and CeRep23. These association propensities may indicate co-dependencies or similiarities in the mechanisms or preferred sites of integration.

7.5 Discussion

An exhaustive list of the chromosomal loci of all known DNA transposons in the Bristol N2 strain of *Caenhorabditis elegans* has been published on the Internet. In general DNA transposons display a clear preference for gene-proximal sequence in their choice of integration site. Statistical patterns of association between different classes of repetitive element have also been demonstrated. For example, 30% of Cele11 repeats are found to be near a copy of Tc5; and CeRep34, CeRep23 and CeRep43 are often found together. These association patterns may be indicative of similarities in the mechanisms of transposition.

A search using hidden Markov models has revealed putative new families of autonomous DNA transposon and one new subgroup of Tc3 elements in the *C.elegans* genome. Phylogenetic evidence suggests recent activity on behalf of one of the new families. The existence of several distinct species of transposon in the same genome with such striking homology between their flanking sequences has implications for the study of transposon ecology and evolution. There are several known mechanisms by which transposons could competitively interact. Transposase proteins of other members of the Tc1/*mariner* family bind specifically to the invrep sequences of transposons of that family *in vitro* [PvL97]. Furthermore, excessive expression of Tc1 transposase protein induces the phe-

Repeat type	Associated repeats (association score/bits)
CeRep10	Cele2 (11.6), CeRep14 (10.6), CeRep11 (10.4), CeRep37 (10.2)
CeRep11	Cele4 (11.1), CeRep10 (10.4)
CeRep12	CeRep13 (11.4)
CeRep13	CeRep18 (12.2), CeRep12 (11.4), CeRep30 (10.6), CeRep33 (10.4)
CeRep14	CeRep10 (10.6), Cele1 (10.4)
CeRep15	Cele7 (11.1)
CeRep17	CeRep19 (12.1), CeRep32 (11.9)
CeRep18	CeRep13 (12.2), CeRep33 (11.1), CeRep30 (11)
CeRep19	CeRep32 (12.2), CeRep17 (12.1)
CeRep22	CeRep37 (11)
CeRep23	CeRep34 (11.8), CeRep43 (11.8)
CeRep24	CeRep38 (12.5), Cele14 (12)
CeRep29	CeRep36 (12.7), CeRep35 (11)
CeRep30	CeRep18 (11), CeRep13 (10.6)
CeRep32	CeRep19 (12.2), CeRep17 (11.9)
CeRep33	CeRep18 (11.1), CeRep13 (10.4)
CeRep34	CeRep43 (12.4), CeRep23 (11.8)
CeRep35	CeRep36 (11.1), CeRep29 (11), CeRep40 (11)
CeRep36	CeRep29 (12.7), CeRep35 (11.1)
CeRep37	CeRep22 (11), CeRep10 (10.2)
CeRep38	CeRep24 (12.5), Cele14 (11.4)
CeRep40	CeRep35 (11)
CeRep41	Tc3 (11.8)
CeRep43	CeRep34 (12.4), CeRep23 (11.8)
Cele1	CeRep14 (10.4)
Cele2	CeRep10 (11.6)
Cele4	CeRep11 (11.1)
Cele7	CeRep15 (11.1)
Cele11	Tc5 (10.5)
Cele14	CeRep24 (12), CeRep38 (11.4)
Tc3	CeRep41 (11.8)
Tc5	Cele11 (10.5)

Table 7.4: Propensities for *C. elegans* repeat types to be found within 1kb of each other. The association scores in brackets are logs of the odds-ratios $\frac{f_{xy}f}{f_x f_y}$ where f_{xy} is the frequency of association of x and y, f_x is the number of associations for x and f is the total number of associations for everything. Only association scores over 10 bits are reported.

nomenon of "overproduction inhibition", reducing transpositional activity in what arguably functions as a regulatory negative-feedback mechanism [HLL97]. It has also been observed that missense mutations in the *mariner*-like MOS1 transposase gene have a dominant-negative effect; the "poisoning" of transposase oligomeric complexes by inactive subunits has been proposed as a mechanism to explain this [HLNL97]. All these mechanisms may work together with host-specific mechanisms to regulate transpositional activity [LC97, HLNL97]. The discovery of dormant *mariner* subfamilies with slight variations in their putative DNA-binding domains and transposase-binding nucleotide sequences may offer new opportunities to study the evolution of DNA-protein specificity in transposon ecology.