# 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<sup>+</sup>92, YWB97, FBT<sup>+</sup>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 transposons, 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 Tc1/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 Tc1 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]. Tc1 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 Tc1 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 Tc1 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 Tc1 transposase protein. The 63 N-terminal residues bind specifically between bases 5 and 26 of the Tc1 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 Tc1 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<sup>+</sup>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 Tc1/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

Tigger1	296	INCLINE POPPAL MEMYKEINVYFMEANTTSILQPMDQGVISTFKSYYL	346
Tigger2	296	VILLILDNAPG PEPHEFNTEGVEVVYLEPNTTSLIQPLDQGVIRTFKAHYT	346
Pogo	267	ILLFIONATS TT VKDFENIKLCFMPPNATALLQPLDQGIIHSFKLEYR	315
Tc2	201	RIIIAWAFKCIISD.	214
Fot1	269	RIMINGHGS ATEOFM. AKCYLNNVYLEFLEAHCSHVLOPLELGCFSSLKAAYR	322
Pot2	268	RUNVLOGHGSHITTDEFM LLCLONNIOLUYLEPHSSHVLOPLULSVFGPLKEAYR	321
hCENP B	292	THUAGRIAAOSLD TSGLEHVOLAFFERGT VHRUREGVVOOVKGHYR	338
PDC2	289	WEVI DECORTI NURLONIKI WYTSSNSK FLEENWGUWDEFKTRYR	336
RAGE	290	WETLARSCSERTV NUMERICATION NUME	337
018572	92	ADVV #COLNDVILOKR#ATASS	114
TC4	220	EVENT DEVENTOR FROM THE FOR WORK OF THE FUT	275
705	212	WILL PAWPAWKWEGDUGSA AL SCHTVHVEST BEGATSELOPCOLVEEC PLKNEVK	267
CoMar	200	THE REPORT OF TH	150
005405	123	FURIONIC POLYNYKYM FETOFLEYEL PHER	154
013530	249	THE PARTY AND AT PERFORMET PROPERTY OF LASSIFFI FONLYKY	300
024691	220	THE DESCRIPTION AT AN AUTOMOUS TOTOL AND OUT DEPARTMENT.	200
024693	233	AND AND AND AN	209
010333	233	VELODING WINNE OF OD CHEMICAN STOLDED IN THE STOLDARD.	209
010332	23/	SILLING AND SUCCESSION OF ADD THE STORE STOLEN AND A STREET STOLEN	407
001891	3/1	DIRLEMERT VERKET. OR OUL WITH PETISPD. LA TURLISSSDIRK	425
023373	140	SILLIGHERPSVERREF. ORLODLIWTVIPE SISPD. LASIDIRLFLSLSDIRR	198
01/312	230	AND AND AND A AND AND AND AND AND AND AN	28/
015299	5/1		021
018573	120	OTTO STRATE OF CONTRACT OF CONTRACT.	140
005408	119	Priliping and a second se	135
018579	121	ILOCK REPERQ. M. TVERLQELEVIA	148
Q05345	119	VT	134
018570	118	MERIOSING TALAR. TELLELEWEVNSHIP	149
Q05346	118	A CHE DIG REPARTING LIVER QELLELGWDVI PHER	149
018576	119	VCC DIATENTSLVT	134
018577	119	V QUDA REPEALVER. QLLELGWDV PHEP	150
018569	119	AVALUDIT RPTTSIQKLREFG.	142
Q05406	119	COMPRESSION SLMTR QCLRELOWEVI	149
018571	119	WITH DURRENTSLMER OKLRELIWEVIMENE	150
018589	119	QDI PREDATV R QKLRELIWEVIMILE	150
018588	119	CHOOL REPORT R. Q LGQL WEVIMEN	150
018574	119	QDHAREFIL	134
018581	119	QQDHARPETSHRK.eKFTELHAFELAPH	151
018595	120	MIT	134
018580	119	TANDUARP TALGERQEIAELWEISHER	150
018591	119	TLENDINRE TAFGTR. OMIAELIWEILSHEP	150
018583	121	TINDUSSETAGEFL.	140
018593	124	TANEDN SSTA	144
018592	123	AT IL DIA TSETAD	142
061675	110	WHETH DHASA RARDEV EFLNTSOVKVER AYTPD	145
Q05409	120	TO DISPASARL	139
Õ18594	120	LIDH PATTOMI.	135
018587	120	ODHAPSTAKPVKDAL.	139
018578	120	PVKDTL.	139
025471	120	SAL.	142
005407	120	RAVR	136
005411	120	AAL.	142
018586	120	VVRDTLEKLO.	143
018584	120	OT	142
CeTc1	94	FYROODIDPKETSLHVR. SWFORRHVHLEDWESOSPD. LNEIE. HLWEELERRI.	144

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

Tiggerl	296	ILULI REPORTERAL MEMYKEINVÜFMÄANTTSILOFMÖOGVISTEKSYYL	346
Ti agar?	200	WEITTER STORES STORES STORES TO STOR	316
6199001 N	200		2.40
8090	207	ZEEPI MAINTEL VKDFENIKECPPRENATALLOPELEGGIIHSPKLETH	313
TC2	201	Kanawaafkoulud	214
Potl	269	REDIVESHGSEATROPM. ARCYLNNVYLEFLEAHCSHVLORLECFSSLKAAYR	322
Fot2	268	PLEVI GHOS ITDEFM. LUCLONNICL YL PHSSHVLOR ALSVFGPLKEAYR	323
BOSNE R	202		320
00000	200		320
82.0.2	409	AWAVIAN COMPLEX NLRLQNIKLAYISSNSK FLEFRMAVWDEPKIRIR	338
RAGS	\$30	WEILANSCSERIVNLNLQNILLEYTSANSRPLEFRWGVLEEFRARYR	337
018572	- 92		214
TC 4	220	SYMMEWPAFKDHTWIKNLVPNGHDVVIRNI#EHTTGMIOPLSVYWNAPWKSLIK	275
Tes	212	WILL ANDAWENECOUTAAALSINETVELVEST BIATSFLOR DLYPECH, KNPVR	257
mMar	99	THE REPORT OF A PROPERTY AND A PROPE	150
006405	122		250
003403	***	Frank Provide the State of Sta	434
613222	649	SULAR SAFALSHOR, ALLREFRWEISSBURTSPD. LARSDFPLFPNLKKSL.	300
Q24691	239	LIM . A AR A SAKNEY A LOOLDLET PHETYSPD. LAPTICHPPOSLDNPL.	289
Q24693	239	ALL AND SAKNEY, ALLOUL LET RHERYSPD. LARTDYHFPOSLDNFL.	283
018332	237	INFLATER VAKETP. OKLODIGWTVEPHEPYSPD. LATTYNLFLSLSDYMR	289
001891	371	AVAILABLE AVAILABLE OF OT OT OT ATTACK AND A AVAILABLE OF AVAILABLE OF A AVAILABLE OF AVAILABLE OF A AVAILABLE OF AVAI	422
******	140	A REAL OF A REAL	100
V43313	7.40	ALL CONTRACT CONDUCTION OF A CARD LOCAL DESIGNATION	290
Q1/314	230	ALTH MART VARPEL. AKLKEMNWEIPPESEISPD. LAPSDIHLPRSLONNL.	287
015299	571	LAL DOARP VAOPTL QELNELGYEVEPHERYSPD. LLETNYHVPKHLNNPL.	621
018573	120	VICTOR VIC	140
Q05408	119	TALL TANK TRO	135
018579	121	NIOFFERENTIAO. M. TVENLORLEVERHER	148
005345	318		3 2 4
010670	320		3.39
0103/0	140		197
Q05346	113	WYFHINING TELVER. OKLLELEWDVEPAPP.	149
018576	119	V	134
018577	119	VIONE ARE ILVIR. OKLLELENDVOPHER.	150
018569	113	ATO OKLREFG.	142
005406	119	CAN DE CALLER OF RELEVENTS	149
018571	110	A DOLLAR DE AND	160
n10000	110		200
040303	113	The second s	100
019299	113	ATAR AND THE REAL OF THE AND A TAR A TA	120
018574	119		134
018581	119	WILCOM AND ISHPER. eKPTELHIPELEPHER.	151
018595	120	MIN MARYNA,	134
018580	119	TITER THE PART OF CRIMETESHEP	150
018681	310	THE REPORT OF THE PARTY SHE	360
010603	121		340
020303	101		3.4.0
018533	124	A A A A A A A A A A A A A A A A A A A	144
018532	123	ATTLE TO AD IVKARL.	142
061675	110	WHILL SA RARDY EFLNTSCVKVEH AYTPD	145
005409	120	KETI.	139
018594	120	MILLING PARTAO. MY	135
018587	120	OTVETAL DOMESTIC	139
010670	120	restriction of the second se	3.20
026378	120	and an	2.27
243471	130	Senter Purster	282
Q05407	120	WIRLEY POLICE	136
Q05411	120	WIROWNIAPEEKS APVRDTI	1.42
018586	120	WIELEWERSEREK	143
018584	120	DIRLIGUE POR RELVESY	342
Catters	64	PROCESSION AND CORRECTION AND CORRECTION AND THE PROPERTY OF	3 8 4
*****	3.40	* 20 * * ******************************	1.4.4

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 Tc1-Tc7 [PvL97, RvLDP97] and Cele1-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 Tc1-Tc7 [PvL97, RvLDP97], Cele1-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 Tc1. If a direct blastn search is performed using the Tc1 sequence as a query, the higher, predicted count is obtained. A possible explanation for this is that the "missing" Tc1 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	Co	opies	Typical length		
		Pairs	Single	Invrep	Whole	
			invreps	only	element	
Tc1 (‡)	TACAGTGCTGGCCAAAAAGA	25	10	81	1620	
Tc2 (†♠)	CCGTATATTCTCTATTAGTG	49	108	24	120	
Tc3 (†♣)	TACAGTGTGGGGAAAGTTCTA	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; ( $\bigstar$ ) includes 3 Cele11 and 32 Cele12 elements described in [OGB96], blastn searches reveal an additional 9 Cele11 and 7 Cele12 elements (approx.); ( $\clubsuit$ ) 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); ( $\P$ ) 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	Single	Invrep	Whole
		(coding)	invreps	only	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; ( $\blacklozenge$ ) 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

Tcl_28856.2 Tc3_20247.4	4 1	evocknislovakaavagpeoaiptkmialdigrapstiwkvikkyoteksvalrispora.avvitkamennika mprosalsitteaaqdvmnilnvslhemsakisrankirvvilkopvsigtskaa, akklsvrderavia	77 71
Tc11_K02C4.5		KEHIINKZLIILKEĞIVSIRINRKSLXMULLINCÖPPLHFGVPILLEYNNEXRNFILKNTĞIPELRPDHKLNIĞALĞ	78
Tcl2_GeneWise	3		2
10165 05202.3 W-15 94894 6	2	IDBULALSHILD Y LI DY LL MARS, NDARRAM, AV LOUND Y SINGHAF ME DATA NAN CULARA A STARAKULI HE DI SA L	30
3013_24424.0	2	BILLADARD/VENIBULETRILLERADALCALLALUNGSSUMDICHEMADWOUTCHEDESSENDIGENEMAMMANALUNG	33
1028 F2003.3			\$
101.0	ંગ		•
Tel_2X856.2	78	SAREDEHREATERONIESSIEDVPSKREVREREOOAELHGPNFVKKEPISKNIMAR VAWAKAHLEWGROEEARH	153
Te3_20247.4	72	AASM. SCATARIERNE ZLSÄS RATTLÄVIKES VIVEQKLEPÄPLESADHELKE LEPAKINMOTNESEV	140
Tell_K02C4.5	79	kalxtv.pfottemstaF5sectnvvgCLAAPikKIRGGFIPHTMIOANLILXVDBSLSLBILHG6DX	151
Tc12_GeneWise	2		13 - 13
Tc12B_F5202.3	81		145
To13_F44F4.8	80	.LVEREPRISLERENCERS	151
Tc14_F26H9.3	72	FREM SGREVCAMARE EISON	142
TO15_F31F4.5	2		10
TH1 28856 3	34.8	THE THE PROPERTY STATEMENT OF THE STATEMENT	377
303 20243 4	361	VIET REAL DODOCORYNERIU REFEMUES ENFOR	331
Tc11 802C4.5	152	LINDERRY YONNEXBACKIGEDETPODAARPOLEERAWLSVARIVYERIVELLE. DSKITTCHISTING	228
Tell Conewise	34	TTERNING VYSHTRENEW WERTATION EXETSEEVIJSIGHDSKWYISHELEE. DEATINASI OOL 2008	87
Te128 F52D2.3	346	OPERING AND	213
Tc13 F44F4.8	3.52	LATISER BY VINNERS FOR ALC KOTPTER POLICEK IN ICV WAY ON PVHWELSE. TNX TTADY CA. OGD	225
Tel4 \$2689.3	143	LPROBRIEC SOPPHYONDRYVANTOPNERVERTCYPEGIMVFAGITANERTPLIPESOGINVNENNYL	215
Tc15_F31F4.5	11	NALKSSVESÖPPKTTRÉRASTLOVOHRSTADÖLTLLÖRR	70
	·		
Tc1_28856.2	228	TTMEPWAL, ONVORGPV XX BUDY TELEVISHED A NEVHILLARS OF VEN 18 . HLWEE ERRIGOIRASNADA	302
TC3_ZC247.4	312	LELSKYLR, NYSKADPARKA, STI VSNIS ADVISC KKINA GMAADMADA I NLWGIL KIVYA (NKTY)PIVA	287
TC11_RO2C4.5	233	VPNNSPLMDADITECTION AND CLARATED AND CONTRACT AND A CONTRACT AND CONTRACT	301
TC12_Genewise	88	KVHAHRLH, KPRGS BLULL PARTE TTLE REPORT VICOUSTING AND PREPARE SECOND AGAY ANTER	103
10128_10204.3	234	AVRAINTAL PRESS CLLL PRAY (11A PROVIDE 1	200
1013_24424.0	220	VALKI POKINGI LI POMO VANGE DI LI PINE EN PINE PINE PINE PINE PINE PINE	220
1014_FA0N7.3	410	TELEVENTATION IN THE CONTRACT AND	473
3615 1311413	1° 2	TRUCT	229
TC1_2K856.2	303	KPNQ SENAPAIPMSVIHKLIDEN RECOATION YATKY 343	
Te3_2C247.4	288	SLKQ2 LDANKSIPUNQLKSLVASHECHLPE KATONNPINY 329	
Tell_K02C4.5	302	VENV. EROTERILEGFIKKETHELET	
Tc12_GeneWise	164	SSAR.GWTTŠLPPNARSSRRSXVC CLC CLC CLC CLC CLC CLC CLC CLC CLC C	
Tc128_F52D2.3	269	Sgefyaegfacter and the second	
Tc13_F44F4.8	299	LKTE. SSTPESSRSPDFFSRGIMME STERSYLCE 139	
Tel4_F26H9.3	296	lkds. Kakameldinylratvüsemelkackaakkoiifel 336	

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 transposase 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 Tc15 with Tc11-Tc13 and (more tenuously) Tc14 with Tc1/Tc3.

#### 7.4.4 Variation within transposon families

The Tc1, Tc3, Tc11, Tc12 and Tc13 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 Tc1, Tc3 and Tc13 trees are evidence that these elements have been active in recent history (indeed, Tc1 and Tc3 are known to be currently active in the Bristol N2 strain [PvL97]), whereas the longer branch lengths of Tc11 and Tc12 suggest that they ceased activity earlier. It can also be seen that Tc12B subgroup of Tc12 transposases form a distinct group, as mentioned above.



Figure 7.5: UPGMA trees constructed from alignments of the Tc1 (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<sup>+</sup>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 Tc1 and Tc13 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	Copies in		Copies in		Copies in	
family	exor	ıs	intro	ns	$1 \mathrm{kb} 5$	' 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	Associated repeats
type	(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.