Evolutionarily conserved regions in *Caenorhabditis* transposable elements deduced by sequence comparison

SHIV S. PRASAD
Department of Medical Genetics, 226-6174 University Boulevard, University of British Columbia, Vancouver, B.C., Canada V6T 1W5

LINDA J. HARRIS
Plant Research Centre, Central Experimental Farm, Ottawa, Ont., Canada K1A 0C6

DAVID L. BAILLIE
Institute of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, B.C., Canada V5A 1S6

AND

ANN M. ROSE[1]
Department of Medical Genetics, 226-6174 University Boulevard, University of British Columbia, Vancouver, B.C., Canada V6T 1W5

Corresponding Editor: P. B. Moens

Received April 10, 1990

Accepted August 9, 1990

PRASAD, S. S., HARRIS, L. J., BAILLIE, D. L., and ROSE, A. M. 1991. Evolutionarily conserved regions in *Caenorhabditis* transposable elements deduced by sequence comparison. Genome, **34**: 6-12.

In this paper we present the sequence of an intact *Caenorhabditis briggsae* transposable element, Tcb2. Tcb2 is 1606 base pairs in length and contains 80 base pair imperfect terminal repeats and a single open reading frame. We have identified blocks of T-rich repeats in the regions 150–200 and 1421–1476 of this element which are conserved in the *Caenorhabditis elegans* element Tc1. The sequence conservation of these regions in elements from different *Caenorhabditis* species suggests that they are of functional importance. A single open reading frame corresponding to the major open reading frame of Tc1 is conserved among Tc1, Tcb1, and Tcb2. Comparison of the first 550 nucleotides of the sequence among the three elements has allowed the evaluation of a model proposing an extension of the major open reading frame. Our data support the suggestion that Tc1 is capable of producing a 335 amino acid protein. A comparison of the sequence coding for the amino and carboxy termini of the 273 amino acid transposase from *Caenorhabditis* Tc1-like elements and *Drosophila* HB1 showed different amounts of divergence for each of these regions, indicating that the two functional domains have undergone different amounts of selection. Our data are not compatible with the proposal that Tc1-related sequences have been acquired via horizontal transmission. The divergence of Tc1 from the two *C. briggsae* elements, Tcb1 and Tcb2, indicated that all three elements have been diverging from each other for approximately the same amount of time as the genomes of the two species.

Key words: *Caenorhabditis*, transposable element, sequence comparison.

PRASAD, S. S., HARRIS, L. J., BAILLIE, D. L., et ROSE, A. M. 1991. Evolutionarily conserved regions in *Caenorhabditis* transposable elements deduced by sequence comparison. Genome, **34** : 6-12.

Cet article comporte la présentation de la séquence d'un élément transposable intact de *Caenorhabditis briggsae*, le Tcb2. Cet élément a une longueur de 1606 paires de bases et contient des répétitions terminales imparfaites de 80 paires de bases, ainsi qu'un seul cadre ouvert de lecture. Des blocs de répétitions riches en T dans les régions 150 à 200 et 1421 à 1476 de cet élément ont été identifiés; ces blocs ont été conservés dans l'élément Tc1 de *Caenorhabditis elegans*. La conservation des séquences de ces régions dans des éléments de diverses espèces de *Caenorhabditis* suggère qu'elles ont une importance fonctionnelle. Un unique cadre ouvert de lecture, correspondant au principal cadre ouvert de lecture de Tc1, est commun au Tc1, au Tcb1 et au Tcb2. Une comparaison des premiers 550 nucléotides de cette séquence chez ces trois éléments a permis d'établir un modèle approximatif qui propose une extension du principal cadre ouvert de lecture. Les résultats obtenus appuient la suggestion que le Tc1 est capable de produire une protéine de 335 acides aminés. Une comparaison des séquences qui codent les terminaux carboxyles et amines d'une transposase de 273 acides aminés, provenant d'éléments semblables au Tc1 de *Caenorhabditis* et de *Drosophila* HB1, a montré des quantités différentes de divergence pour chacune de ces régions, indiquant donc que deux domaines fonctionnels ont subi des quantités différentes de sélection. Les résultats soumis ne sont pas compatibles avec la proposition avancée que les séquences reliées au Tc1 ont été acquises par transmission horizontale. Le divergence de Tc1 avec les deux éléments de *C. briggsae*, Tcb1 et Tcb2, indique que les trois éléments ont chacun divergé les uns des autres pour à peu près la même quantité de temps que les génomes des deux espèces.

Mots clés : *Caenorhabditis*, élément transposable, comparaison de séquences.

[Traduit par la Rédaction]

Introduction

In the genome of *Caenorhabditis briggsae*, two families of transposable elements with homology to the Tc1 element of *C. elegans*, namely Tcb1 (Barney) and Tcb2, have been identified (Harris et al. 1990). Tc1-like elements are conserved between *Caenorhabditis* and *Drosophila* and define a family of inverted repeat transposable elements (Harris et al. 1988; Henikoff and Plasterk 1988). Both Tc1 and Tcb1 were found to have sequence similarity to the HB1 element of *Drosophila melanogaster* (Harris et al. 1988). Recently

[1]Author to whom all correspondence should be addressed.

a new element, Uhu, has been identified in *Drosophila heteroneura* which shows extensive sequence similarity to Tc1, Tcb1, and HB1 (Brezinsky et al. 1990). Tcb1 and Tcb2 are present in 15 and 33 copies, respectively, in the genome of the *C. briggsae* strain G16 (Harris et al. 1990). Both elements show many genomic hybridization pattern differences between two *C. briggsae* strains, G16 (Fodor et al. 1983) and Z, suggesting that they, like Tc1, are capable of transposition.

Tc1 was first described by Emmons et al. (1983) and Liao et al. (1983). Although it is inactive in most strains of *C. elegans*, Tc1 has mutator activity in a few strains (reviewed in Moerman and Waterston 1989). The same population of elements exists in low and high copy number strains (Rose et al. 1985), supporting the proposal that the high copy number mutator strains grew in number from low copy number quiescent strains (Emmons et al. 1983). Mutator activity has arisen from three sources, spontaneous (Moerman and Waterston 1984) and ethyl methyl sulphonate treated (Eide and Anderson 1985) derivatives of the var. Bergerac strain BO and spontaneous occurrences in two var. Bristol strains, CB30 (C. Trent in Moerman and Waterston 1989) and KR579 (Babity et al. 1990). Extrachromosomal copies of Tc1, which may be transposition intermediates, have been detected (Rose and Snutch 1984; Ruan and Emmons 1984) and mutator loci that are responsible for order of magnitude increases in Tc1 activity have been mapped (Mori et al. 1988). Although the mechanism of mutator activity is unknown, a favored hypothesis is that changes in the element itself have led to increased expression of the transposase. If so, it would be of value to look for conserved noncoding sequences that could be responsible for the control of Tc1 expression. *Caenorhabditis briggsae* has been used for this kind of comparison to elucidate the functionally important portions of the *C. elegans* genome (Snutch 1984; Heine and Blumenthal 1986; Prasad and Baillie 1989; Starr 1989; Zucker-Aprison and Blumenthal 1989; Heschl and Baillie 1990). The sequence of Tcb1 analyzed in previous studies (Harris et al. 1990) did not come from a complete element but was constructed from two defective elements, one which did not contain the 5′ region and another with a large deletion in the open reading frame. This composite element would not be the best choice for searching for the regulatory sequences. To examine the conserved regions of the Tc1-related elements, we describe the complete nucleotide sequence of an intact *C. briggsae* transposable element from the Tcb2 family of elements and compare its sequence with that of Tcb1 in *C. briggsae* and Tc1 in *C. elegans*.

Materials and methods

Construction of plasmids

The isolation and characterization of *C. briggsae* transposable elements (Tcb1 and Tcb2) have been described elsewhere (Harris et al. 1990). Plasmids pCbh1 to pCbh5 contain the Tcb2 class of elements cloned in the pUC19 vector. The plasmid pCbh1 contains a 2.4-kb *Eco*RI insert. Two fragments, 1.3 and 1.1 kb, were generated on digestion of this plasmid with *Bam*HI and *Eco*RI. Both of these fragments were electroeluted as described in Maniatis et al. (1982) and correspondingly ligated into the Bluescript− vector (Stratagene). All plasmids were transformed into *E. coli* JM83 [ara(lacpro)rpsLthi80dlacZM15] cells (Messing 1979). The plasmids pCbh13 and pCbh15 contain the 1.3- and 1.1-kb *Bam*HI/*Eco*RI inserts, respectively. Overlapping deletions for DNA sequencing were generated by the Exonuclease III/S1 nuclease method of Henikoff (1987).

Plasmid DNA preparation

Plasmid DNA for restriction mapping, deletion, and dideoxy sequencing was prepared by the mini alkaline lysis method (Maniatis et al. 1982). The supercoiled plasmid DNA for sequencing was further purified from the low melting point agarose gels by extraction with phenol at 70°C (Maniatis et al. 1982).

DNA sequencing

The supercoiled plasmids were denatured in 0.2 M NaOH, 1 mM EDTA for 5 min at room temperature, neutralized with 2 M ammonium acetate pH 7.4, and immediately precipitated with ethanol at −70°C for 5 min. The pellets were washed with 70% ethanol, dried under vacuum, and stored at −20°C. The collapsed plasmids were sequenced by the chain termination method (Sanger et al. 1977) outlined in Hattori and Sakaki (1986) using DNA polymerase (Klenow fragment) and [^{35}S]dATP. Each portion of the analyzed DNA was sequenced twice from overlapping deletion clones on the same strand.

Sequence analysis

The FASTA sequence program (Pearson and Lipman 1988) and Delaney SEQNCE2, DB system (Staden 1980) were used for sequence data compilation and to search for sequence identity, both in nucleic acid and amino acid coding potential. Sequences were aligned and formatted using the eyeball sequence editor program, ESEE (Cabot and Beckenbach 1989).

Results and discussion

Characterization of the Tcb2 elements from C. briggsae

Screening of a partial *Eco*RI Charon 4 *C. briggsae* genomic library (constructed by T. Snutch, Simon Fraser University) with *C. elegans* Tc1 resulted in the purification of 10 different phage (Harris et al. 1990). The individual *Eco*RI fragments with homology to Tc1 were subcloned into pUC19 vector. Hybridization of individual subclones to each other revealed two classes of *C. briggsae* transposable elements based upon the intensities of hybridization signals. The predicted sequence of one class Tcb1 (constructed from two defective members of this class) has been described earlier (Harris et al. 1990). The other 5 of the 10 elements (pCbh1 to pCbh5) correspond to a second class, Tcb2. To examine this element in detail, we have sequenced one member of Tcb2 class derived from the subclone pCbh1. The sequence of the Tcb2 element is presented in Fig. 1 and is aligned with the sequence of *C. briggsae* Tcb1 and *C. elegans* Tc1 (Rosenzweig et al. 1983). Tc1 is 1610 base pairs (bp) long, including 54 bp perfect terminal inverted repeats and a long open reading frame (ORF1) with a coding potential for a 273 amino acid protein. A second, shorter ORF in another reading frame of the same strand with a coding potential for a 112 amino acid protein and putative TATA and CAAT boxes adjacent to the longer ORF were described. Tc1 elements are flanked by TA dinucleotide, which represents the duplicated target sites (Rosenzweig et al. 1983).

The sequence of Tcb2 as shown in Fig. 1 is 1606 bp in length. As observed in Tcb1 (Harris et al. 1990), Tcb2 has 80 nucleotide imperfect terminal repeats (Fig. 2). The right and left termini of the inverted repeats are 82% identical. The open reading frame (ORF) of Tcb2 starts at nucleotide 521 and stops at nucleotide 1340, the same position as that of the major ORFs of Tc1 and Tcb1.

```
TC1      CAGTGCTGGCCAAAAAGATATCCACTTTTGGTTTTTTGTGTGTAACTTTTTTCTCAAGCATCCATTTGACTTGAATTTTTCCGTGTGCATAAAGCGAAATGTTACGCAAA 110
TCb1     CAGTACTGGCCATAAAGAATGCGACAACTTGTTTTTTGACGATAACTTTTTGAAAACTCAACTTTTCAACTCGAATTTTTTACTGTAGACACATTGGAAAGTTACGAAAA 110
TCb2     CAGTACTGGCCAAAAAGGTATTCGCAAACGGTTTTTTGAGAGTAACTTTGCTGAAACTTAAGGGATCGATCTGAAAATTTTACTGTGCATGTCGTAGACGGTGACGACAA 110
TC1 -pro
TCb1-pro
TCb2-pro

TC1      TTTGCGGACCAA-ACATTACATGATTATCGATTTTTTCTGAATTTTATTTCAATTTTTTGATTTTTTCGTTTTTCCAATTTTCATTATTTTTTTTGAATTATCAATAAAA 219
TCb1     CTTGCGTACTTGGAAATTGCATCATCTTGCAAACAAAATTTTTTTTCAATTGTTTTCATTTTTTTTTTCATTTTGCATATTTTTCATGATTTTTGGTTTTTTTAAATAAA 220
TCb2     ATTGCGTACAAGGAAATTGCATCATTTT--AAGAAATTTTTTGTTTTTTTCAATTTTTTCAATTTTTTCGTTTTTTCACTTTTTTCATATTTTTATATTTTTGGATATTA 218
TC1 -pro
TCb1-pro
TCb2-pro

TC1      CGCACTCTGTTTGTTGCA----CTGGATTTGTTTGGTTGATAAATTATTTTTAAGGTATGGTAAAATCTGTTGGGTGTAAAAATCTTTCCTTGGACGTCAAGAAAGCCAT 325
TCb1     GTTGAAAGACAACTAATTGGAATTAGTTTTAGTTGATTATTTATTGTTTTTCAGATATGGTTAGACCATTCGGAAAAAATCCTTCTCTGTGGACGAGCGAAAAAGCATT 330
TCb2     TGGTTTCTAAGAATCTTGGT--TTTGGCTCCTATGTATTTAAGACCTGATTACAGACATGGCCAGACCCCTCGGAAAAAAAAACATCTCTCAGGACATCCGTAAAGCCAT 326
TC1 -pro                                                          M  V  K  S  V  G  C  K  N  L  S  L  D  V  K  K  A  I 18
TCb1-pro                                                          M  V  R  P  F  G  K  N  P  S  L  W  T  S  E  K  A  L 18
TCb2-pro                                                          M  A  R  P  L  G  K  K  N  I  S  Q  D  I  R  K  A  I 18

TC1      TGTAGCTGGCTTCGAACAAGGAATACCCACGAAAAGTCTCGCGCTGCAAATTCAACGTTCTCCGTCGACTATTTGGAAAgtaatcaagaa-gtaccaaactgaggtgagt 434
TCb1     GTGCGTGGTCATGAACTCGGCGCGCATCCGAAGATACTTGCAACTCAGTTTAGTTGTTCTTCCAGCCAAATTTACCgtattctgaagaaaaatcgagatgatgtgaggtt 440
TCb2     CGTCATCGCTTCAGAAGTGATGTTCCAAGACCTC---TTGCGCAGCAGTTCAATGTTACTAAACGTGCCATCGAg+aagatcttgaagca-gtttcgggatgaggtaata 431
TC1 -pro  V  A  G  F  E  Q  G  I  P  T  K  S  L  A  L  Q  I  Q  R  S  P  S  T  I  W  K                                 44
TCb1-pro  C  V  V  M  N  S  A  R  I  R  R  Y  L  Q  L  S  L  V  V  L  P  A  K  F  T  -                                 43
TCb2-pro  V  I  A  S  E  V  M  F  Q  D  L  -  L  R  S  S  S  M  L  L  N  V  P  S  -  -                                 41

TC1      tcgaaaaatattatttttttaataataaatgtttagAAATCCGTCGCTTTGAGAATCTCGCCCGGCAGGCCTCGAGTGACAACCCATAGGATGGATCGCAACATCCTCCGA 544
TCb1     tcaatcagctttctcaactaaatttaacaatttagAATGGCGTTGTTCATCGCCAATCTCCAGGAAGACCCCGAACCACTTCCAGAAATATGGATCGGAACATTTTGCGC 550
TCb2     ttcaaaaatagtatagttatggaaaaaattgtttagGGAGCCTTGGTGGTCGAAAATCGCCTGGAAGACCGCTTATCACGTCGAAGAACATGGATCGGAACATCGTCCGA 541
TC1 -pro                                     K  S  V  A  L  R  I  S  P  G  R  P  R  V  T  T  H  R  M  D  R  N  I  L  R 69
TCb1-pro                                     N  G  V  V  H  R  Q  S  P  G  R  P  R  T  T  S  R  N  M  D  R  N  I  L  R 68
TCb2-pro                                     R  S  L  G  G  R  K  S  P  G  R  P  L  I  T  S  K  N  M  D  R  N  I  V  R 66

TC1      TCAGCAAGAGAAGATCCGCATAGGACCGCCACGGATATTCAAATGATTATAAGTTCTCCAAATGAACCTGTACCAAGTAAACGAACTGTTCGTCGACGTTTACAGCAAGC 654
TCb1     GCATGCCGAGAAGATCCTAGACGCACATCTACGGGTATTCAACTTTCTGTGACGTCTCCAAATGAACCGGTACCATCGAGAAGAACTATTAGAAGACGTTTGCAAGTTGC 660
TCb2     GTTTCAAGAGAGGATCCCCAGACAAGTTCTACGGACATTCAGATGGTCGTGAAGACTCCAAATGAAGTCACACCATCCCTGAGAACCGTCAGAAGACGTCTTCAAGATGC 651
TC1 -pro  S  A  R  E  D  P  H  R  T  A  T  D  I  Q  M  I  I  S  S  P  N  E  P  V  P  S  K  R  T  V  R  R  R  L  Q  Q  A 106
TCb1-pro  A  C  R  E  D  P  R  R  T  S  T  G  I  Q  L  S  V  T  S  P  N  E  P  V  P  S  R  R  T  I  R  R  R  L  Q  V  A 105
TCb2-pro  V  S  R  E  D  P  Q  T  S  S  T  D  I  Q  M  V  V  K  T  P  N  E  V  T  P  S  L  R  T  V  R  R  R  L  Q  D  A 103

TC1      AGGACTACACGGACGAAAGCCAGTCAAGAAACCGTTCATCAGTAAGAAAAATCGCATGGCTCGAGTTGCGTGGGCAAAAGCGCATCTTCGTTGGGGACGTCAGGAATGGG 764
TCb1     TGGACTGCACGGACGAAGACCAGTCAAAAAACCACTCGTCAGTTTGAAAAACCGAAAAGCTCGCGTTGAATGGGCTAAACAGCACTTGTCCTGGGGCCCCCGTGAGTGGG 770
TCb2     TGGACTCCATGGACGACGGCCAGCTAAGAAACCATCGATCAGCAAGAAGAACAGAATCGCCCGCGTAGCATGGGCCAGAGCTCATCTCCACTGGGGACGTCAGGATTGGG 761
TC1 -pro  G  L  H  G  R  K  P  V  K  K  P  F  I  S  K  K  N  R  M  A  R  V  A  W  A  K  A  H  L  R  W  G  R  Q  E  W  142
TCb1-pro  G  L  H  G  R  R  P  V  K  K  P  L  V  S  L  K  N  R  K  A  R  V  E  W  A  K  Q  H  L  S  W  G  P  R  E  W  141
TCb2-pro  G  L  H  G  R  R  P  A  K  K  P  S  I  S  K  K  N  R  I  A  R  V  A  W  A  R  A  H  L  H  W  G  R  Q  D  W  139
```

FIG. 1. The complete sequence of Tcb2 aligned with the sequence of *C. elegans* Tc1 (from Rosenzweig et al. (1983) with the insertion of a T after position 361) and Tcb1 (from Harris et al. 1990). Amino acid alignment is presented below the nucleotide sequences. Gaps made to achieve the maximum alignment are shown by a dash (-), the position of an added T required to produce a functional splice acceptor site in Tcb2 by a plus (+), and nucleotide matches by a vertical rule (|). Lower case letters indicate the position of the nucleotides within the intron observed by Schukkink and Plasterk (1990).

```
TC1       CTAAACACATCTGGTCTGACGAAAGCAAGTTCAATTTGTTCGGGAGTGATGGAAATTCCTGGGTACGTCGTCCTGTTGGCTCTAGGTACTCTCCAAAGTATCAATGCCCA 874
TCb1      CAAATCACATCTGGAGCGATGAATCGAAGTTCAATATGTTCGGAACTGATGGTATTCAGTGGATTCGACGTCCCATTGGCTCCAGGTATGCTCCACAGTACCAATGTCCA 880
TCb2      CTAATCACGTGTTCAGCGACGAATCCAAGTTCAATCTGTTCGGTACTGACGGTATCAAGTGGATTCGCCGTCCGGTTGGATGTCGGTTCGATCCCAGCTACCAGCTCCAA 871
TC1 -pro  A K H I W S D E S K F N L F G S D G N S W V R R P V G S R Y S P K Y Q C P 179
TCb1-pro  A N H I W S D E S K F N M F G T D G I Q W I R R P I G S R Y A P Q Y Q C P 178
TCb2-pro  A N H V F S D E S K F N L F G T D G I K W I R R P V G C R F D P S Y Q L Q 176

TC1       ACCGTTAAGCATGGAGGTGGAGCGTCATGGTGTGGGGGTGCTTCACCAGCACTTCCATGGGCCCACTAAGGAGAATCCAAAGCATTATGGATCGTTTTCAATACGAAAA 984
TCb1      ACTGTGAAACATGGAGGTGGATCTGTGATGGTTTGGGGATGCTTCTCCGACACTTCTATGGGACCATTGAAGAGAATCGTTGGAACCATGGATCGATATGTGTACGAAGA 990
TCb2      ACCGTCAAGCATGGAGGTGGATCGGTGATGGTCTGGGGCTGCTTCTCTGGGACTTCAATGGACCCGCTTCGAAGGATCGACAGCATCATGGATCGCTTTGTCTATGAAGA 981
TC1 -pro  T V K H G G G S V M V W G C F T S T S M G P L R R I Q S I M D R F Q Y E N 216
TCb1-pro  T V K H G G G S V M V W G C F S D T S M G P L K R I V G T M D R Y V Y E D 215
TCb2-pro  T V K H G G G S V M V W G C F S G T S M D P L R R I D S I M D R F V Y E D 213

TC1       CATCTTTGAAACTACAATGCGACCCTGGGCACTTCAAAATGTGGGCCGTGGCTTCGTGTTTCAGCAGGATAACGATCCTAAGCATACTTCTCTTCATGTGCGTTCATGGT 1094
TCb1      CATCCTGGAGAACACAATGAGACCATGGGCAAGAGCAAATTTGGGCCGATCGTGGGTGTTCCAACAGGACAATGACCCGAAGCATACTTCGGGTCATGTCGCCAATTGGT 1100
TCb2      CATCTTGGAGAACACAATGAGACCCTGGGCAAGGTCTACTGTGGGCCGGGCGTTCGTTTTCCAACAGGATAACGACCCGAAGCACACCTCGAAGCACATCAAGGAGTGGT 1091
TC1 -pro  I F E T T M R P W A L Q N V G R G F V F Q Q D N D P K H T S L H V R S W 252
TCb1-pro  I L E N T M R P W A R A N L G R S W V F Q Q D N D P K H T S G H V A N W 251
TCb2-pro  I L E N T M R P W A R S T V G R A F V F Q Q D N D P K H T S K H I K E W 249

TC1       TTCAACGTCGTCATGTGCATTTGCTCGATTGGCCAAGTCAGTCTCCGGACTTGAATCCAATAGAGCATTTGTGGGAAGAGTTGGAAAGACGTCTTGGAGGTATTCGGGCT 1204
TCb1      TCAGACGTCGCCGTGTGAACCTCCTAGAATGGCCAAGTCAATCTCCAGACTTGAATCCCATCGAGCATATGTGGGAGGAGCTGGAACGCCGCCTCAAAGGAGTCAGAGCA 1210
TCb2      TCCGACGCCGCCACGTGGATCTTCTGGACTGGCCAAGTCAATCTCCGGACTTGAATCCGATCGAACACCTCTGGGAGCATGTCGAGAGACACGTTAGAGGGGTTCGTGCC 1201
TC1 -pro  F Q R R H V H L L D W P S Q S P D L N P I E H L W E E L E R R L G G I R A 289
TCb1-pro  F R R R R V N L L E W P S Q S P D L N P I E H M W E E L E R R L K G V R A 288
TCb2-pro  F R R R H V D L L D W P S Q S P D L N P I E H L W E H V E R H V R G V R A 286

TC1       TCAAATGCAGATGCCAAATTCAACCAGTTGGAAAACGCTTGGAAAGCTATCCCCATGTCAGTTATTCACAAGCTGATCGACTCGATGCCACGTCGTTGTCAAGCTGTTAT 1314
TCb1      TCCAATGCCAATCAAAAGTTTGCTCAACTCGAAGCTGCTTGGAAGAGTATCCCGATGACGGTGGTTCAGACGCTCCTGGAGTCGATGCCACGTAGATGCAAGGCTGTTAT 1320
TCb2      TCGAATGCCAACCAGAAGTTCGACCAACTCCAAGATGTCTGGCAGGCCATCCCGATGTCTGTCATCGACACCATTCTCGACTCGATGCCAAGAAGATGTCAAACGGTAAT 1311
TC1 -pro  S N A D A K F N Q L E N A W K A I P M S V I H K L I D S M P R R C Q A V I 326
TCb1-pro  S N A N Q K F A Q L E A A W K S I P M T V V Q T L L E S M P R R C K A V I 325
TCb2-pro  S N A N Q K F D Q L Q D V W Q A I P M S V I D T I L D S M P R R C Q T V I 323

TC1       TGATGCAAACGGATACGCGACAAAGTATTAAGCATAATTATGTTGTTTTTAAATCCAATTGCTCATATTCCGGTACTTTAATTGTCATTTCCTTGCAACCTCGGTTTTTT 1424
TCb1      CGATGCGAAAGGATACCCAACGAAATATTAAATCACTTTTTATTGTTACCTTTATTTTGTTTTTGACCTCCGGTGTTTCTTGAGCTTCTTTGTATAGATCTCCAAAAATG 1430
TCb2      CGACGCCAAAGGCTTCCCAACGAAGTACTAAAGCTA-TTTTATTCATTTTCTATTATTATTTTTCGTTTACCGGTCTATCCTAGATCTCAGAAGTCAAAGCTTCCATTTA 1420
TC1 -pro  D A N G Y A T K Y *                                                                                              335
TCb1-pro  D A K G Y P T K Y *                                                                                              334
TCb2-pro  D A K G F P T K Y *                                                                                              332

TC1       CAATATTTCTAGTTTTTCGATTTTTTTGAATTTTTCTGAAGTTTTTTCAAAATCTGTTGAACATTTTTGATGAATATTGTGTTTTTAGATTTTGTGAACACTGTGGTGAA 1534
TCb1      TCAT-TTTCATGAATTTTCGAAAAAAATTAATTTTCAAAAAAATTGTAGCAGAATATTAAAACCAACTTAGCCTAAAAGTTTTTGATAGTGCTGAAAAAATCTTACCAAA 1539
TCb2      CCGGTTTTTTGCAATTTTCATGAATTTTTCGATTTTTTTTCAATTTTTTGAAATTTTGACTATTCTTGTTGTTGAATCATTTTCTTTTACCGTTTTAAACTTTGTCTTAA 1530
TC1 -pro
TCb1-pro
TCb2-pro

TC1       GTTTCAAAACAAAATAACCACTTAGAAAAAAGTTACACACAAAAAACCAAAAGTGGATATCTTTTTGGCCAGCACTG                                  1611
TCb1      AATTCGGATTGAAAAGTTGAGTTTTCAAAAAGTTATCTTCAAAAAACAAGTTGTCGCATTCTTTATGGCCAGTACTG                                  1616
TCb2      AATATCAGATTGATCGGGAAGTTTCAGAAAAGTTATGCTCGAAAAACTATTTGCGAATAGCTTTTTGGCCAGTACTG                                  1607
TC1 -pro
TCb1-pro
TCb2-pro
```

Fig. 1 (concluded).

```
CAGTACTGGCCAAAAAGGTATTCGCAAACGGTTTTTTGAGAGTAACTTTGCTGAAACTTAAGGGATCGATCTGAAAATTT
||||||||||||||||| ||||||||||  |||||| |||  ||||||| |||||||||     |||| |||||| | |||
CAGTACTGGCCAAAAAGCTATTCGCAAATAGTTTTTCGAGCATAACTTTTCTGAAACTTC-CCGATCAATCTGATATTTT
```

FIG. 2. Alignment of the 5′ (top) and the 3′ (bottom) inverted repeats in the sequence of Tcb2. Nucleotide matches are shown by the vertical rule (|).

```
(A)

Tcb2 5'  GTTTTTTTCAATTTTTTTCAATTTTTTTCGTTTTTT
         ||||||| |||||||     | |||||||| |||||
Tcb2 3'  GTTTTTTGCAATTTTCATGAATTTTTTCGATTTTT

(B)

Tcl 5'  TATTTCAATTTTTT-GATTTTTTCG--TTTTTCC-AA-TTTT
        |||||| | ||||| |||||||| |  ||||||  || ||||
Tcl 3'  TATTTCTAGTTTTTCGATTTTTTTGAATTTTTCTGAAGTTTT
```

FIG. 3. (A) Alignment of the region 151–185 (Tcb2 5′) with the region 1424–1458 (Tcb2 3′) of the Tcb2 element is presented here. Matches are shown by the vertical rule (|). (B) Alignment of the region 155–200 (Tcl 5′) with the region 1427–1469 (Tcl 3′) of the Tcl element is shown here. Gaps (–) in the sequences are introduced to maintain maximum identity. Matches are shown by a vertical rule (|).

The major ORF of Tcl is conserved in Tcbl and Tcb2

A single ORF is conserved among Tcl, Tcbl and Tcb2, suggesting that it encodes the transposase. The sequence of Tcl has two ORFs on the same strand, a major and a minor (Rosenzweig et al. 1983), while Tcbl has two ORFs on different strands, the larger of which corresponds to the major ORF of Tcl (Harris et al. 1990). Neither of the minor ORFs is conserved in Tcb2. Recently Schukkink and Plasterk (1990) and A. Radice, N. Sherif, and S.W. Emmons (personal communication) have suggested that a longer transcript may be generated by splicing. Additional elements from the Bergerac strain, which were sequenced by Plasterk (1987) and I. Mori and R.H. Waterston (personal communication) had an extra T nucleotide after position 361. Schukkink and Plasterk (1990) observed that the insertion of this nucleotide creates a longer open reading frame, a potential intron sequence with a consensus splice donor at nucleotide 404, and a splice acceptor at 468. If splicing were to occur, a 335 amino acid protein would be produced. Examination of both Tcbl and Tcb2 sequences showed that an ATG at position 273 and the proposed splice acceptor site are conserved. The potential donor site is present in Tcbl and can be generated in Tcb2 with the addition of a T after position 398. When the potential intron sequence is removed, an open reading frame can be generated for all three elements. In pairwise comparisons between elements this extended ORF shows 30–45% amino acid identity for the first third, 5–20% for the middle third, and 35–45% for the last third. These observations lend support to the proposal that a transcript longer than that coding for 273 amino acids could be produced by Tcl.

We have examined our sequence data to determine the amount of conservation in each of the amino and carboxy termini of the transposase. Schukkink and Plasterk (1990) have proposed that the Tcl element encodes a protein that has two independent functional domains, the N-terminal peptides specifying the DNA binding structural domain and the C-terminal peptides forming the enzymatic functional domain, i.e., transposase function. They demonstrated that deletion of the first 39 N-terminal amino acid abolishes the DNA binding properties of the protein synthesized by Tcl of *C. elegans*, whereas deletion of 108 C-terminal amino acids does not affect the DNA binding properties. We compared the first and the last 40 amino acids encoded by Tcl, Tcbl, and Tcb2 of *Caenorhabditis* and HB1 of *Drosophila*. No significant variation was observed within the *Caenorhabditis* genus. However, a comparison with the HB1 element showed that the Tcl, Tcbl, and Tcb2 elements diverged at a significantly higher rate (17.5–20.0% conservation) in the first 40 N-terminal amino acids than in the last 40 C-terminal amino acids (37.5–42.5% conservation). Although HB1 is not now a functional element, it presumably has come from a functional ancestral element, which may have diverged from Tcl prior to becoming defective. It is possible that the reaction site function has been more highly conserved than the DNA binding domain. Our observation that the amino acid divergence differs for the amino and carboxy termini supports the proposal of Schukkink and Plasterk (1990) that the protein has two independent functional domains and that these are under different amounts of selection.

Tcb2 sequence has a direct repeat in the 5′ and 3′ regions

Unexpectedly, the Tcb2 element has a direct repeat flanking the ORF (Fig. 3A). Twenty-seven of the 34 nucleotides within the region 151–185 are repeated in the region 1424–1458 when deletions or insertions are allowed. These repeats are also present in Tcl. Thirty-three of the 42 nucleotides within the region 155–200 are repeated in the region 1427–1469 when deletions or insertions are allowed (Fig. 3B). We have compared the repeats between Tcl and Tcb2. Forty-three of the 51 nucleotides matched perfectly in the region 150–200, while 46 of the 55 nucleotides were identical within the region 1421–1476 (Fig. 1).

These repeats were not identifiable in the Tcbl element. The comparison of the 3′-flanking region of Tcbl with either Tcb2 or Tcl showed no significant identity. However, the 5′ sequences were conserved. The repeat within nucleotide region 151–212 of Tcb2 was compared with the similar region of Tcbl; 48 of the 61 nucleotides were found to be identical (Fig. 4A). When a similar comparison was done between Tcl and Tcbl, 46 of the 60 compared nucleotides were found to be identical (Fig. 4B). The absence of the repeat in Tcbl is puzzling, although it should be noted that the Tcbl sequence was assembled from two defective members of this family (Harris et al. 1990). The sequences in these conserved blocks are T-rich, although other nucleotides are also included at some positions. The possibility that this repeat is not functionally important and fortuitously conserved in Tcl and Tcb2 cannot be ruled out. However, we favor the interpretation that since the sequences in these regions have diverged less than those of surrounding regions, they are of functional importance. Conservation may be of runs of T's rather than of a particular nucleotide sequence and these runs may play a role in DNA structure or thermal stability. A diagram of the conserved regions in the Tcl-like elements is shown in the Fig. 5.

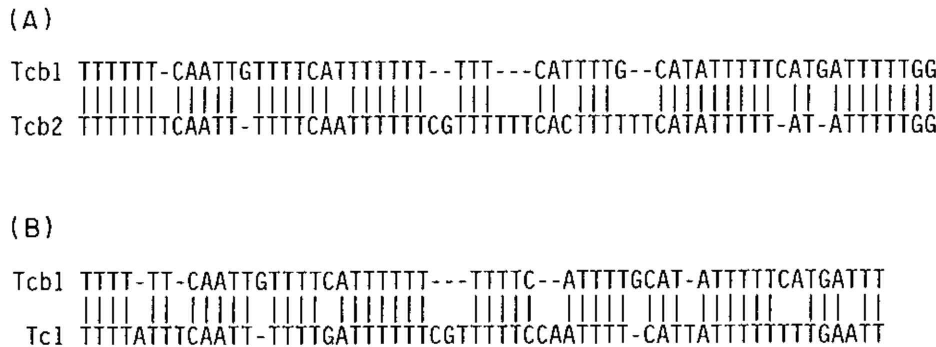

FIG. 4. (A) Alignment of the region 151–212 between Tcb1 (Tcb1 5′) and Tcb2 (Tcb2 5′). Gaps (-) introduced to maintain maximum identity are indicated. Matches are shown by a vertical rule (|). (B) Alignment of the region 151–212 between Tcb1 (Tcb1 5′) and Tc1 (Tc1 5′). Gaps introduced to maintain maximum identity are denoted by -. Matches are shown by a vertical rule (|).

FIG. 5. A drawing of the Tc1-related elements showing the three conserved features: (*i*) inverted repeats shown by hatched arrowheads, (*ii*) conserved blocks of T-rich repeats shown by closed arrowheads, and (*iii*) an open reading frame shown by closed box.

Rosenzweig et al. (1983) identified the TATA and CAAT box sequences at the 5′ end of the major ORF of Tc1. The sequences required for transcription initiation have not, as yet, been confirmed by functional assay. These TATA and CAAT boxes have not been conserved between Tc1 and Tcb1, casting doubt on their functional importance. It is possible that Tc1 and Tcb2 elements utilize different regulatory sequences. However, in other gene systems sequence conservation has been a useful tool for identifying promoter and enhancer elements; for example, the hsp-70 genes (Heschl and Baillie 1990) and the vitelogenin genes (Zucker-Aprison and Blumenthal 1989).

Estimates of divergene of Tc1, Tcb1, and Tcb2

Overall sequence identity in the region between the first inverted repeat and start of the ORF was found to be 54.3% between Tc1 and Tcb1, 55.2% between Tc1 and Tcb2, and 57.8% between Tcb1 and Tcb2. It is possible to estimate the time of divergence for the three different elements derived from *C. elegans* and *C. briggsae*. An estimate was made by determining the nucleotide divergence of the silent third positions of fourfold degenerate codons. The nucleotide divergence calculated for these silent sites between Tc1 and Tcb1 is 88.6% when corrected for multiple mutational events (Jukes and Cantor 1969). A similar comparison shows that the nucleotide divergence between Tcb1 and Tcb2 is 92%, while that between Tc1 and Tcb2 is 120.7%. These results provide the evidence that the Tc1 related sequences were not acquired as a result of horizontal transmission. In contrast, Daniels et al. (1990) showed that P elements of *Drosophila melanogaster* are not only present in *D. willistoni* and *D. saltans*, they are very similar in sequence as well. In addition previous studies have shown that the P element is absent in much more closely related species to *D. melanogaster* (Brookfield et al. 1984) and old laboratory strains of *D. melanogaster* (Bingham et al. 1982). Daniels et al. (1990) have thus provided evidence for horizontal transmission of the P transposable element between *Drosophila* species.

Nucleotide substitution rates at the silent third position of codons have been determined for bacteria (0.8% per 1 000 000 years (Ma) (Ochman and Wilson 1987)), insects (1.1%/Ma (Bodmer and Ashburner 1984)), sea urchins (1.1%/Ma (Busslinger et al. 1982)), nuclear genes of mammals (0.9%/Ma (Li et al. 1985*a*, 1985*b*)), and flowering plants (1%/Ma (Chang and Meyerowitz 1986)). Considering that all these organisms have a very similar nucleotide divergence (0.8–1.1%/Ma), it is reasonable to assume that the *Caenorhabditis* genes would evolve at similar rates. Therefore, we suggest that the Tcb2 element has been separated from Tc1 for 96–135 Ma, while Tcb1 has been separated from Tc1 for 72–99 Ma. An explanation for these results is that two diverged elements, Tcb1 and Tcb2, were acquired in *C. briggsae* from a common ancestor, while a Tcb1-like element, i.e., Tc1 was acquired in *C. elegans*. Another explanation would be that the two species in question began with a single related element, duplicated at some time in *C. briggsae*, which have since diverged significantly. Therefore, the changes in the silent third codon positions have been saturated. A test of this proposal would be to examine DNA divergence for other coding sequences.

Mechanisms that maintain sequence identity between members of a multigene family could be expected to be acting on these elements (Maeda and Smithies 1986). In the *C. briggsae* genome, our data indicate that these mechanisms are occurring within Tcb1 and Tcb2, but not between these two groups. For example, Harris et al. (1990) showed that two defective Tcb1 elements were identical in their 3′ regions. However, no sequence conservation with Tcb2 was observed in the 3′ region of the element. These results support the suggestion that concerted evolution has occurred within the Tcb1 and Tcb2 groups, but not between Tcb1 and Tcb2, and that sequence conservation between Tcb2 and Tc1 is due to selection acting on functionally important sequences.

Acknowledgements

The authors thank R. Plasterk and R. Schukkink for sharing their results before publication. The authors also thank the reviewers for helpful comments on the manuscript. This research was supported by a Medical Research Council of Canada operating grant to A.M. Rose.

BABITY, J., STARR, T., and ROSE, A.M. 1990. Transposition of Tc1 in a Bristol strain of *Caenorhabditis elegans*. Mol. Gen. Genet. **222**: 65–70.

BINGHAM, P.M., KIDWELL, M.G., and RUBIN, G.M. 1982. The molecular basis of P-M hybrid dysgenesis: the role of the *P* element, a *P*-strain specific transposon family. Cell, **29**: 995–1004.

BODMER, M., and ASHBURNER, M. 1984. Conservation and change in the DNA sequence coding for alcohol dehydrogenase in sibling species of *Drosophila*. Nature (London), **309**: 425–430.

BREZINSKY, L., WANG, G.V.L., HUMPHREYS, T., and HUNT, J. 1990. The transposable element Uhu from Hawaiian *Drosophila*—a member of the widely dispersed class of Tc1-like transposon. Nucleic Acids Res. **18**: 2053–2059.

BROOKFIELD, J.F.Y., MONTGOMERY, E., and LANGLEY, C.H. 1984. Apparent absence of transposable element related to the *P* elements of *D. melanogaster* in other species of *Drosophila*. Nature (London), **310**: 330–332.

BUSSLINGER, M., RUSCONI, S., and BIRNSTEIL, M.L. 1982. An unusual evolutionary behaviour of a sea urchin histone gene cluster. EMBO J. **1**: 27–33.

CABOT, E.L., and BECKENBACH, A.T. 1989. Simultaneous editing of multiple nucleic acids and protein sequences with ESEE. Comput. Appl. Biosci. **5**: 233–234.

CHANG, C., and MEYEROWITZ, E.M. 1986. Molecular cloning and DNA sequence of the *Arabidopsis thaliana* alcohol dehydrogenase gene. Proc. Natl. Acad. Sci. U.S.A. **83**: 1408–1412.

DANIELS, S.B., PETERSON, K.R., STRAUSBAUGH, L.D., KIDWELL, M.G., and CHOVNICK, A. 1990. Evidence for horizontal transmission of the *P* transposable element between *Drosophila* species. Genetics, **124**: 339–355.

EIDE, D., and ANDERSON, P. 1985. Transposition of Tc1 in the nematode *C. elegans*. Proc. Natl. Acad. Sci. U.S.A. **82**: 1756–1760.

EMMONS, S.W., YESNER, L., RUAN, K.S., and KATZENBERG, D. 1983. Evidence for a transposon in *Caenorhabditis elegans*. Cell, **32**: 55–65.

FODOR, A., RIDDLE, D.L., NELSON, F.K., and GOLDEN, J.W. 1983. Comparison of a new wild-type *Caenorhabditis briggsae* with laboratory strains of *C. briggsae* and *C. elegans*. Nematologica **29**: 203–217.

HARRIS, L.J., BAILLIE, D.L., and ROSE, A.M. 1988. Sequence identity between an inverted repeat family of transposable elements in *Drosophila* and *Caenorhabditis*. Nucleic Acids Res. **16**: 5991–5997.

HARRIS, L.J., PRASAD, S.S., and ROSE, A.M. 1990. Isolation and sequence analysis of *C. briggsae* repetitive elements related to the *C. elegans* transposon Tc1. J. Mol. Evol. **30**: 359–369.

HATTORI, M., and SAKAKI, Y. 1986. Dideoxy sequencing method using denatured plasmid templates. Anal. Biochem. **152**: 232–238.

HEINE, U., and BLUMENTHAL, T. 1986. Characterization of regions of the *Caenorhabditis elegans* X chromosome containing vitellogenin genes. J. Mol. Biol. **188**: 301–312.

HENIKOFF, S. 1987. Unidirectional digestion with exonuclease III in DNA sequence analysis. Methods Enzymol. **155**: 156–165.

HENIKOFF, S., and PLASTERK, R.H.A. 1988. Related transposons in *C. elegans* and *D. melanogaster*. Nucleic Acids Res. **16**: 6234.

HESCHL, M.F.P., and BAILLIE, D.L. 1990. Functional elements and domains inferred from sequence comparisons of a heat shock gene in two nematodes. J. Mol. Evol. **31**: 3–8.

JUKES, T.H., and CANTOR, C.R. 1969. Evolution of protein molecules. *In* Mammalian protein metabolism. Vol. 3. *Edited by* H.N. Munro. Academic Press, New York.

LI, W.H., LUO, C.C., and WU, C.I. 1985*a*. Evolution of DNA sequences. *In* Molecular evolutionary genetics. *Edited by* R.J. MacIntyre. Plenum, New York. pp. 1–94.

LI, W.H., WU, C.I., and LUO, C.C. 1985*b*. A new method for estimating synonymous and non-synonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**: 150–174.

LIAO, L.W., ROSENZWEIG, B., and HIRSH, D. 1983. Analysis of a transposable element in *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. U.S.A. **80**: 3583–3589.

MAEDA, N., and SMITHIES, O. 1986. The evolution of multigene families: human haptoglobin genes. Annu. Rev. Genet. **20**: 81–108.

MANIATIS, T., FRITSCH, E.F., and SAMBROOK, J. 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

MESSING, J. 1979. A multipurpose cloning system based on the single-stranded DNA bacteriophage M13. National Institute of Health Publication No. 79-99.2. Recomb. DNA Tech. Bull. **2**: 43–48.

MOERMAN, D.G., and WATERSTON, R.H. 1984. Spontaneous unstable *unc*-22 IV mutations in *C. elegans* var. Bergerac. Genetics, **108**: 859–877.

______ 1989. Mobile elements in *Caenorhabditis elegans* and other nematodes. *In* Mobile DNA. *Edited by* D.E. Berg and M.M. Howe. American Society for Microbiology, Washington, DC. pp. 538–556.

MORI, I., MOERMAN, D.G., and WATERSTON, R.H. 1988. Analysis of a mutator activity necessary for germline transposition and excision of Tc1 transposable element in *Caenorhabditis elegans*. Genetics, **120**: 397–407.

OCHMAN, H., and WILSON, A.C. 1987. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J. Mol. Evol. **26**: 74–86.

PEARSON, W.R., and LIPMAN, D.J. 1988. Improved tools for biological sequence analysis. Proc. Natl. Acad. Sci. U.S.A. **85**: 2444–2448.

PLASTERK, R.H.A. 1987. Differences between Tc1 elements from the *C. elegans* strain Bergerac. Nucleic Acids Res. **15**: 10 050 – 10 051.

PRASAD, S.S., and BAILLIE, D.L. 1989. Evolutionarily conserved coding sequences in the *dpy*-20 - *unc*-22 region of *Caenorhabditis elegans*. Genomics, **5**: 185–198.

ROSE, A.M., and SNUTCH, T.P. 1984. Isolation of the closed circular form of the transposable element Tc1 of *Caenorhabditis elegans*. Nature (London), **311**: 485–486.

ROSE, A.M., HARRIS, L.J., MAWJI, N.R., and MORRIS, W.R. 1985. Tc1(Hin): a form of the transposable element Tc1 in *Caenorhabditis elegans*. Can. J. Biochem. Cell Biol. **63**: 752–756.

ROSENZWEIG, B., LIAO, L., and HIRSH, D. 1983. Sequence of the *C. elegans* transposable element Tc1. Nucleic Acids Res. **11**: 4201–4209.

RUAN, K.S., and EMMONS, S.W. 1984. Extrachromosomal copies of transposon Tc1 in the nematode *Caenorhabditis elegans*. Proc. Natl. Acad. Sci. U.S.A. **81**: 4018–4022.

SANGER, F., COULSON, A.R., BARRELL, B.G., SMITH, A.J.H., and ROE, B.A. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. **74**: 5463–5467.

SCHUKKINK, R.F., and PLASTERK, R.H.A. 1990. TcA, the putative transposase of the *C. elegans* Tc1 transposon, has an N-terminal DNA binding domain. Nucleic Acids Res. **18**: 895–900.

SNUTCH, T.P. 1984. Molecular and genetic analysis of the heat shock response of *Caenorhabditis elegans*. Ph.D. thesis, Simon Fraser University, Burnaby, B.C.

STADEN, R. 1980. A new computer method for the storage and manipulation of DNA gel reading data. Nucleic Acids Res. **8**: 3673–3694.

STARR, T. 1989. Molecular analysis of the *dpy*-14 region of chromosome I in *Caenorhabditis elegans*. Ph.D. thesis, University of British Columbia, Vancouver, B.C.

ZUCKER-APRISON, E., and BLUMENTHAL, T. 1989. Potential regulatory elements of nematode vitellogenin genes revealed by interspecies sequence comparison. J. Mol. Evol. **28**: 487–496.