Evolutionary Relationships Among the Members of an Ancient Class of Non-LTR Retrotransposons Found in the Nematode *Caenorhabditis elegans*

Ignacio Marín, Piedad Plata-Rengifo,† Mariano Labrador,‡ and Antonio Fontdevila†*

*Departamento de Genética and Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Universidad de Valencia, Spain; †Departamento de Genética y Microbiología, Universidad Autónoma de Barcelona, Spain; and ‡Department of Biology, Johns Hopkins University

We took advantage of the massive amount of sequence information generated by the *Caenorhabditis elegans* genome project to perform a comprehensive analysis of a group of over 100 related sequences that has allowed us to describe two new *C. elegans* non-LTR retrotransposons. We named them *Sam* and *Frodo.* We also determined that several highly divergent subfamilies of both elements exist in *C. elegans.* It is likely that several master copies have been active at the same time in *C. elegans,* although only a few copies of both *Sam* and *Frodo* have characteristics that are compatible with them being active today. We discuss whether it is more appropriate under these circumstances to define only 2 elements corresponding to the most divergent groups of sequences or up to 16, considering each subfamily a different element. The *C. elegans* elements are related to other previously described non-LTR retrotransposons (*CR1,* found in different vertebrates; *SR1,* from the trematode *Schistosoma*; *Q* and *T1,* from the mosquito *Anopheles*). All of these elements, according to the analysis of their reverse transcriptases, form a monophyletic cluster that we call the ''*T1/CR1* subgroup.'' Elements of this subgroup are thus ancient components of the genome of animal species. However, we discuss the possibility that these elements may occasionally be horizontally transmitted.

Introduction

Retrotransposons can be divided in two main classes, according to their presence or absence at both ends of the elements of long terminal repeats (LTRs). The structure of many non-LTR retrotransposons, also called LINEs, is reminiscent of that of a retrovirus (reviewed in Eickbush 1992, 1994). They often have two open reading frames (ORFs) and end in a 3′ untranslated region with repetitious adenine-rich sequences. However, there is substantial structural variability among non-LTR retrotransposons. The reverse transcriptase (RT), generally found in the second ORF, is the only protein encoded by all them. It has been shown that most elements also encode an endonuclease (EN) (Martín, Olivares, and López 1996; Feng et al. 1996). In some other elements, however, no EN is apparent, and whether these elements actually encode a highly divergent EN is unknown (Feng et al. 1996). ORF2 of some elements also contains an RNAse H. Moreover, several elements totally lack ORF1 (Feng et al. 1996) (throughout the text, we will use the convention of referring to the ORF containing the RT as ''ORF2,'' even in the absence of ORF1). Finally, unrelated sequences, produced by recombinational events, at the 5′ ends of otherwise almost identical elements have also been observed (Adey et al. 1991, 1994; Hayward, Zavanelli, and Furano 1997 and references therein). How all of these structural types have arisen is still poorly understood. The current evidence, based on the comparative analysis of RTs, strongly suggest that all non-LTR retrotransposons have evolved from a single ancestral lineage (Xiong and Eickbush 1990; Eickbush 1994). Thus, all of these structural differences must have been acquired secondarily, in some cases very recently.

An additional source of sequence variation arises because most of the copies produced by non-LTR retrotransposons become inactivated, suffering truncations at their 5′ ends. These copies then diverge randomly and rapidly (reviewed in Smit 1996; Finnegan 1997). In fact, the number of active retrotransposons per genome is very low (reviewed in Deininger et al. 1992). In the case of human LINEs, it has been recently estimated that only 30–60 copies of the 100,000 found in a genome are active (Sassaman et al. 1997).

Because non-LTR retrotransposons may have different structures, only a few copies are active, and most copies are highly divergent and inactive, it turns out that conventional methods—those based on detection by DNA hybridization and subsequent cloning and sequence comparisons—are often insufficient to obtain a clear picture of the evolutionary dynamics of these elements. Considering this situation, the global information rendered by the sequencing of complete genomes may be very useful, because it allows exhaustive comparative studies of the elements inhabiting the analyzed organism. In this work, we present the first results of such an approach for non-LTR retrotransposons of the nematode *Caenorhabditis elegans.* This is a particularly interesting species to study, because the sequencing project of the *C. elegans* genome is about 70% complete (for a review, see Blumenthal and Spieth 1996).

We previously described (Marín and Fontdevila 1996) certain sequences of the fly *Drosophila koepferae* that were closely related to those described for two *Anopheles* non-LTR retrotransposons, called *T1* and *Q* (Besansky 1990; Besansky, Bedell, and Mukabayire 1994). Several studies have established that other ele-

Abbreviations: LTR, long terminal repeat; EN, endonuclease; RNAse H, ribonuclease H; RT, reverse transcriptase.

Key words: *Caenorhabditis elegans,* non-LTR retrotransposon, reverse transcriptase, endonuclease, genome project, horizontal transmission.

Address for correspondence and reprints: Ignacio Marín, Departamento de Genética, Universidad de Valencia, Dr. Moliner, 50, Burjassot 46100, Valencia, Spain. E-mail: ignacio.marin@uv.es.

Mol. Biol. Evol. 15(11):1390–1402. 1998

ments related to *T1* and *Q* occur in vertebrates (element *CR1*: Stumpf et al. 1981; Chen et al. 1991; Burch, Davis, and Haas 1993; Vandergon and Reitman 1994; Ohshima et al. 1996; Haas et al. 1997; Kajikawa, Ohshima, and Okada 1997) and also in the trematode *Schistosoma mansoni* (element *SR1*, Drew and Brindley 1997). From now on, we will call this group of elements the ‘‘*T1/CR1* subgroup.’’ Recently, we observed that sequences generated in the genome project of the nematode *C. elegans* show substantial similarity to the elements of the *T1/CR1* subgroup. We describe here these new sequences, demonstrating that they belong to two different elements that are represented multiple times in the *C. elegans* genome. Different subfamilies of these elements are present in the databases. Sequence comparisons suggest that the number of potentially functional (‘‘master’’) copies of these two elements, which we will call *Sam* and *Frodo* (Tolkien 1954), is small. We will discuss the implications of these results with regard to the evolutionary history of these elements, the phylogenetic range of the *T1/CR1* subgroup of non-LTR retrotransposons, and the possibility of horizontal transmission of some of its members.

Materials and Methods

The *C. elegans* sequences considered in this work were first found to be related to known transposable elements when the putative products of the *T1/Q*-related sequences of *Drosophila koepferae* (Marín and Fontdevila 1996) were compared using TBLASTN to the sequences of the nonredundant database at the National Center for Biotechnology Information (NCBI; for all of the BLAST programs, version 1.4.9MP [Altschul et al. 1990], implemented online at the NCBI page (http://www.ncbi.nlm.nhm.gov/) was used). We found later, using the same type of analysis, that the *C. elegans* sequences were related to those of the *Anopheles gambiae* elements *T1* and *Q*, as well as to the elements *CR1* (chicken and turtle) and *SR1* (RT sequence only; from *Schistosoma*). Only a low degree of similarity to other non-LTR retrotransposons, mainly in the RT region, was detected.

A combination of nucleotide searches with BLASTN and searches with TBLASTN and BLASTP (Altschul et al. 1990) using the sequences of the putative products encoded by the *C. elegans* sequences was used to establish that they can be divided into two main types, corresponding to two different elements. The same procedures were used to compare the sequences belonging to each of these elements in order to define subfamilies. Nucleotide BLASTN searches were also used to establish the lengths and structures of the different sequences found in *C. elegans*. ORF identification was also performed online using the ORF Finder tool at NCBI.

Once the best candidates for active elements were established (see *Results*), multiple-sequence alignments were performed for both the EN- and RT-related sequences of a number of retrotransposons using CLUSTAL W (version 1.6; Thompson, Higgins, and Gibson 1994). The output of the program was refined by hand after comparison with the previous TBLASTN results as well as with a database of aligned sequences of other non-LTR retrotransposons. To formally establish the phylogenetic relationship among the elements considered in this work, the program implementing the neighbor-joining method (Saitou and Nei 1987) included in CLUSTAL W was used. Those positions containing gaps in one or more sequences were excluded from such analysis. Programs RETREE and DRAWGRAM from the PHYLIP package (Felsenstein 1989, 1993) were used to draw the phylogenetic trees presented in figures 2 and 4. The bootstrapping routine also included in CLUSTAL W was used to determine the values presented in those figures. The program GeneDoc (Nicholas and Nicholas 1997) was used to highlight the similarities among the multiple-aligned sequences in figures 1 and 3.

Results

New Members of the *T1/CR1* Subgroup Are Detected in *C. elegans*

In Marín, Labrador, and Fontdevila (1992), we described middle repetitive DNA-containing clones obtained from the genomes of *Drosophila buzzatii* and *Drosophila koepferae*, two sibling species of the *repleta* group. Among the transposable-element-related sequences detected when some of those clones were sequenced (Labrador and Fontdevila 1994; Marín and Fontdevila 1995, 1996), we were able to determine that the conceptual translation of two clones defined products related to those found in non-LTR retrotransposons. In particular, a clear similarity to the *Anopheles* elements *T1* and *Q* was observed (Marín and Fontdevila 1996). Two years later, in a routine search using the *D. koepferae* sequences, we found that *C. elegans* sequences recently included in the publicly available databases also show a substantial similarity to those putative products. Further analyses (see *Materials and Methods*) using the whole sequences of the EN and RT of the fully characterized *Anopheles* elements confirmed this finding.

Preliminary searches using some of the *C. elegans* sequences showed that they were also related to the other two characterized members of the *T1/CR1* subgroup, *CR1* and *SR1*, while their relationships with other non-LTR retrotransposons were much more distant. In particular, they exhibit only a very weak relationship to the only other non-LTR retrotransposon previously studied in *C. elegans*, the element *Rte-1* (Youngman, van Luenen, and Plasterk 1996). Deeper analyses determined that the *C. elegans* sequences can be classified into two groups. When their putative protein products are compared, sequences in each of these two groups show a high level of similarity, while the similarity between groups is much lower. For example, for the RT domain, around 60% of the amino acids are identical (and 75%–80% biochemically similar) when random members of the same group are compared. Among sequences of different groups, however, the similarity is about the same as that found when related, but different, non-LTR retrotransposons are compared (around 30% identity and 50% biochemical similarity). We interpreted the substantial differences between these two groups of sequences, as well as their similarities within a group, to correspond to the fact that two different non-LTR retrotransposons of the *T1/CR1* subgroup are present in the genome of *C. elegans*. We named these two elements *Sam* and *Frodo* (Tolkien 1954).

Structural Analysis of the *C. elegans* Sequences

To further characterize these sequences, and especially to confirm that two different elements were present, a priority was to establish whether any copy of these sequences in the databases could correspond to active transposable elements. To solve this problem, we developed a strategy based on three premises: (1) Active elements (or those sequences most similar to active elements) should behave as ‘‘master copies.’’ Therefore, a certain number of sequences almost identical to a structurally complete copy, but with particular truncations in their 5′ ends, should be found in the genome. (2) By comparison of the nucleotide sequences, it should be possible to extend the elements from zones known to correspond to their 3′ ends (such as those encoding the RT, which are generally found toward the end of non-LTR retrotransposons) to upstream zones. In this way, we should be able to approximately define the 5′ ends of the elements. (3) Finally, the ORFs of active elements should be free of truncations, gaps, frameshifts, and stop codons.

In order to further constrain our search for active elements, we started our analysis considering only those sequences that contain complete EN sequences. EN sequences are found upstream of RT sequences (this has been confirmed for all of the *C. elegans* sequences considered in this study), so they are expected to be intact only in those elements without, or with relatively small, 5′ truncations. To establish that EN domains were complete, we compared the conceptual translations of the available sequences with the ENs of the other elements of the *T1/CR1* subgroup, and we looked for the known conserved domains of the protein (Martín, Olivares, and López 1996; Feng et al. 1996). A total of 14 full-length EN sequences were found, 11 for the group of sequences that we defined as corresponding to the *Sam* element and 3 corresponding to *Frodo* elements (tables 1 and 2).

At Least Five Subfamilies of *Frodo* Coexist in the *C. elegans* Genome

When we studied the nucleotide sequences of the different EN-containing copies in detail, we discovered that sometimes even those that we had classified as belonging to the same element according to their protein sequences were highly heterogeneous at the nucleotide level. Thus, two of the *Frodo* sequences with full-length ENs, which we call *Frodo2.1* and *Frodo2.2*, are almost identical along 3,300 bp (see table 1; henceforth, we will call the sequences by a name [*Sam* or *Frodo*] that indicates the main group to which they belong; a number to indicate the subfamily; and, if more than one EN-containing copy of a subfamily is present, a second number to indicate the particular copy). However, the other *Frodo* sequence that contains a complete EN (*Frodo1*) is only distantly related to *Frodo2.1* and *Frodo2.2*

Table 1
Summary of the Most Complete *Frodo* Copies and Their Longest Truncated relatives

Subfamily	Name of Master Copy	Name of Clone Containing Master Copy (position in the clone, 5′–3′)	Accession No.	Endonuclease	Reverse Transcriptase	3′-Terminal Sequence	Other Clones Containing Fragments of Elements of this Subfamily
I	FRODO1	CEK06A4 (>23,780–20,784)	Z70755	Complete	Complete? (see text)	5′-TAATAAATACAATACAATACAA-3′	CEF58E6, CEC06A12, CEB0513, CELF54D8, CET09F5, CELF47D2, CEK07F5, CEY7A9C, CELF54C1, CEC50B8, CER13H4, CEF43C1, CEM199, CELK12C11, CELM03D4, CELK03F8, CELT21F2
II	FRODO2.1	CEAH6 (>24,687–21,408)	Z48009	With frameshifts	With frameshifts	5′-TTAAATTAAATTAAAATTAAA-3′	CEM01E11, CELZC196, CELF08B1, CET04B2, CEZK892, CEC33D9, CEC49F5, CEC50F4, CELK07E12, CELZC487, CEC50C10, CELC23H3, CELY102E9, CELF49D11, CELF07G11
	FRODO2.2	CEZK1073 (<11,700–14,936)	Z68135	Complete	Truncated	Same as FRODO2.1	

Table 2
Summary of the Most Complete *Sam* Copies and Their Longest Truncated Relatives

Sub-family	Name of Master Copy	Name of Clone Containing Master Copy (position in the clone, 5′–3′)	Accession No.	Endonuclease	Reverse Transcriptase	3′-Terminal Sequence	Other Clones Containing Fragments of Elements of this Subfamily
I	SAM1	CELTO7E3 (<19,600–22,449)	U13643	Complete	Complete	5′-CAAATAAACAATTATTAAA-3′	CER17, CEK11H3, CELC52E12, CELF25B4, CEC08F11, CELF38E1, CEF56H9, CED1086, CELT01B11, CELC15C7
II	SAM2	CELB0478 (>20,000–17,169)	U57054	Complete	With frameshifts	5′-TTTGAATAAATATATATAT-3′	CELC12D12, CET05F1, CET05F1A, CELF47D2, CET13F3, CELT27A10, CEF09C8, CEC03D6, CER04D3, CELK05G3
III	SAM3	CELF38E9 (<18,500–21,336)	U46668	With frameshifts	Complete	5′-TGAATGAATGATATATATATATATAT-3′	CEY7A9B, CELF30B5, CEF07C6, CELT02G5, CET19C4, CELW02D7, CELC31B8, CELR08E3
IV	SAM4	CET19C9 (>17,262–13,825)	Z92972	Complete	Complete	5′-TAAATTATTATTA-3′	CELW03F9, CELF36A4
V	SAM5.1	CEF58D12 (>4,800–1,125)	Z81092	With frameshifts	Complete	5′-TAAATAAATTAAA-3′	CEF23A7, CEC06H5, CELEGAP4, CEC50H2, CELF3GH5, CET02C1, CELC25F6, CELC09G12, CEK06G5
	SAM5.2	CELZK250 (<26,200–truncated)	AF025472	With frameshifts	With frameshifts	None (truncated 3′ end)	
VI	SAM6	CER05A10/ CEK01G12 (<31,943–32,011 (END)/ 1–>3,364)	Z82280/ Z82275	Complete	Complete	Unknown	CELM01D7
VII	SAM7.1	CEZK337 (<7,625–10,613)	Z82090	Complete	Complete	5′-TCAAGGCAATTAA-3′	CELC23G10, CEF28B1, CELT23E7, CELK02F3, CELF10G2, CEC35D6, CEC06C3
	SAM7.2	CEF38B2/ CEF08G12 (<17,400–19,805 (END)/ 1–720)	Z50045/ Z66561	Complete	With frameshifts	Same as SAM7.1	
VIII	SAM8	CELF21E9 (<12,000–15,060)	AF016663	Complete	With frameshifts	5′-TCAAATTCAATTCAATTCAA-3′	CEM04C7, CET02E9, CEK08G2, CEF54D5, CELF39F10, CEK03F8, CEC50B6, CELF59E11, CEC54C8, CELF58A6, CET01C1, CET07H6
IX	SAM9	CEF16B12 (<7,400–10,100)	Z81064	Complete	Truncated	5′-TTAAATAAATCAATCAATCAA-3′	CEF58A4, CEF59A1, CEF28H7, CEF26E4, CEZK673, CELZK488, CEZK218, CELC06E4, CEF57G4

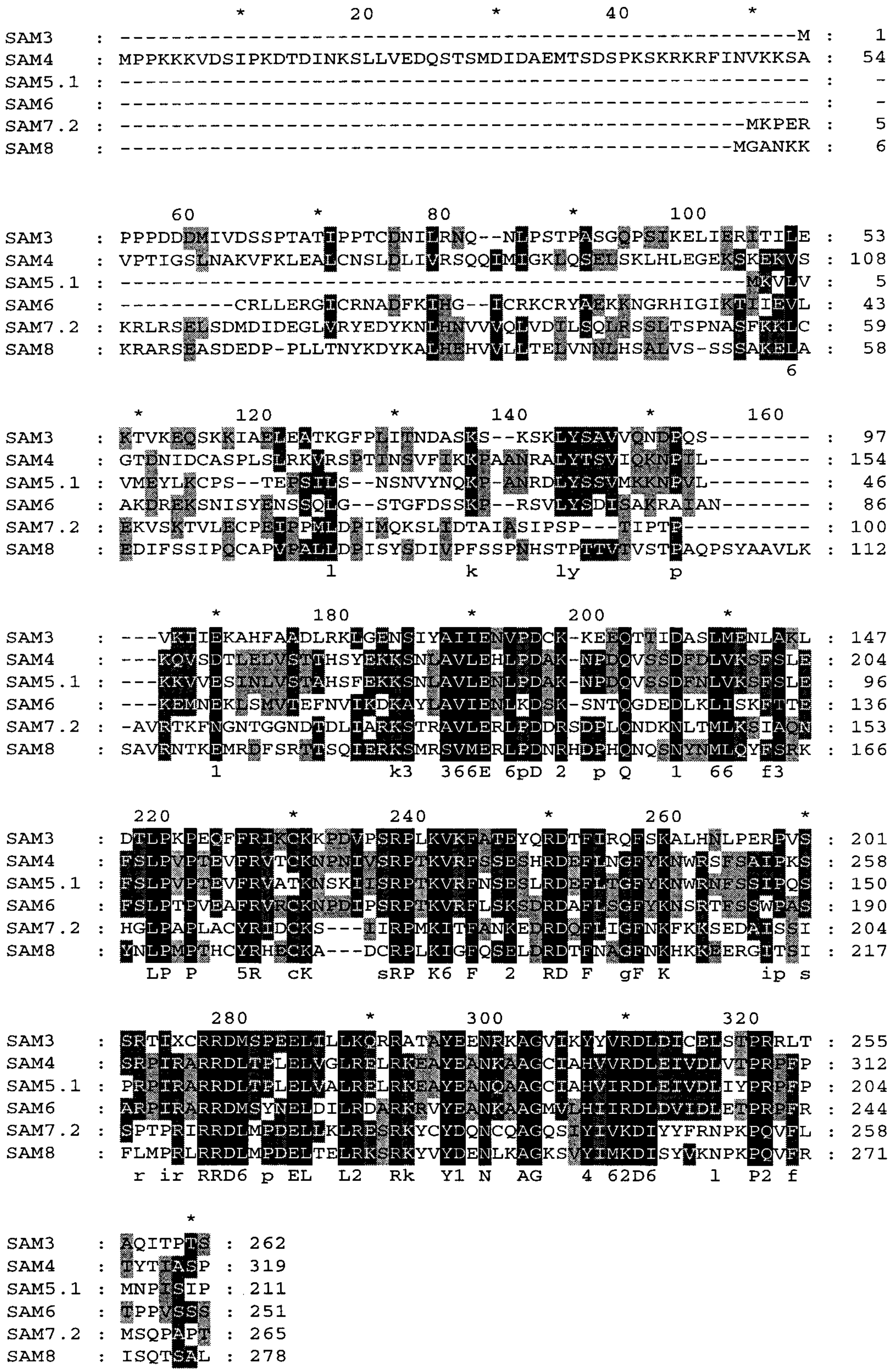

FIG. 1.—Putative ORF1 found in six *Sam* sequences. A seventh copy, *Sam9,* contains almost 60 amino acids corresponding to the C-terminal end of this putative protein (not shown), while no related sequences were found in the other *Sam* copies. The letters and numbers under the alignments summarize the most conserved positions. Numbers refer to biochemically similar amino acids according to the Blosum62 matrix (1: D, N, E; 2: E, Q, K, R; 3: S, T, A; 4: H, Y; 5: F, Y, W; 6: L, I, V, M).

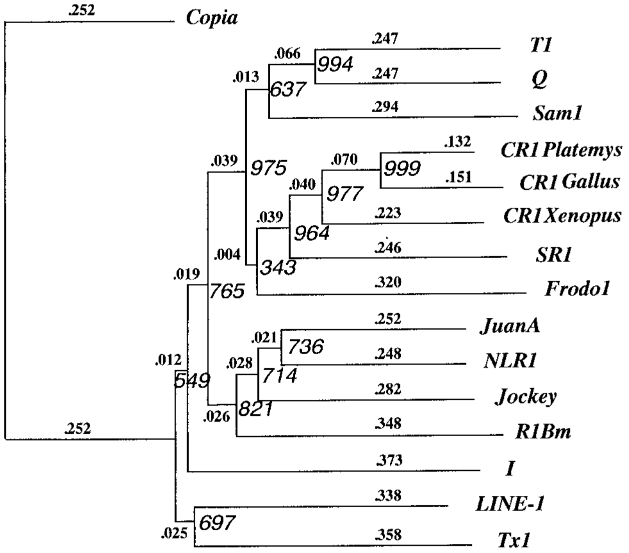

FIG. 2.—Phenogram obtained using the RT sequences of selected non-LTR retrotransposons. The branch lengths are shown above the lines in small numerals. The numbers of cases out of 1,000 in which the adjacent branch was supported in bootstrap analysis are shown in italics (see *Materials and Methods*).

(about 55%–60% nucleotide identity along their length). 5′-truncated versions of both *Frodo1* and *Frodo2.1/2.2* with very high nucleotide identities (90%–100%) to those "master" (EN-containing) copies are present in the databases (table 1). Therefore, at least two highly differentiated subfamilies of the *Frodo* element exist in *C. elegans.* Henceforth, we will follow the convention of considering those sequences with a nucleotide identity of at least 90% members of the same subfamily.

With regard to their coding sequences, of the three longest *Frodo* copies, *Frodo2.1* is clearly an inactive copy, because its putative coding regions have several frameshifts and stop codons, while *Frodo2.2* seems to be almost complete but lacks a small stretch of the N-terminal end of the RT. Only *Frodo1* seems to have a structure compatible with being an active element. However, *Frodo1* has a frameshift in its RT domain (at nucleotide 21901 in the clone). Interestingly, a frameshift in the same position is found in *Frodo2.2.* The fact that the two sequences that seem to be structurally closest to our expectation for an active element contain the same frameshift suggests that it might be present in active *Frodo* retrotransposons.

As we have already mentioned, clones *Frodo2.1* and *Frodo2.2* have 98% nucleotide identity along 3,300 bp. This includes 300 bp upstream of the N-terminal end of the EN. We have not been able to extend the similarity of any *Frodo* sequence farther upstream. This does not necessarily mean that an ORF1 is absent in *Frodo.* In fact, upstream of the ENs of *Frodo2.2,* we detected a putative ORF of 275 amino acids (96 amino acids of this putative ORF1 are also present in *Frodo2.1,* which thus could be a truncated copy). It is unclear, however, whether this sequence actually encodes a functional product, or even part of one. No similarity to any of the proteins found in the ORF1s of other known non-LTR retrotransposons (including those of the *T1/CR1* subgroup and the putative ORF1 of *Sam,* see below) has been detected. No long putative ORF1 has been detected in *Frodo1.*

Fortunately it has been much easier to define the 3′ ends of the *Frodo* elements. The similarity between *Frodo2.1* and *Frodo2.2* finishes in a repetitious sequence, which, in *Frodo2.2,* is 5′-TAAATAAAT-TAAATTAAATTAAAATTAAA-3′. The similarity between these two sequences and the other truncated copies of the subfamily (summarized in table 1) commonly ends at the same sequence, although nucleotide substitutions are occasionally observed. This type of 3′ repetition is also found in the other available long copy, *Frodo1,* as is easily determined by comparing that copy with its truncated variants (table 1). However, the repetitions are different for each subfamily. In the case of subfamily I, similarity among copies ends at the sequence 5′-TAATAAATACAATACAATACAA-3′. As we mentioned in the introduction, this type of repetitious, often A-rich, sequence is typical of the 3′ ends of non-LTR retrotransposons, including other elements of the *T1/CR1* subgroup.

More than 30 other *Frodo* truncated copies are present in the databases that are not 90% identical to the three EN-containing copies and, therefore, have not been included in table 1. About three quarters of them can be grouped, following the criterion of >90% nucleotide identity, in three other subfamilies whose longest representatives are the sequences found in the clones CELR05F9 (nucleotides 11499–13784), CELK07E8 (nucleotides 24552–27068), and CEC43D7 (nucleotides 15086–16719), respectively. The rest are very small fragments, at most a few hundred base pairs long.

Caenorhabditis elegans Contains No Less than 11 Highly Divergent Subfamilies of *Sam*

An even more complex situation is found when *Sam* sequences are analyzed (table 2). Eleven sequences contain putative full-length ENs, and 10 of these copies also seem to possess a complete RT. However, like *Frodo* sequences, *Sam* sequences are quite heterogenous. Although some copies are up to 95% identical along thousands of nucleotides, others are much less related, such that even the nucleotide sequences of their RTs, probably the slowest-evolving part of the elements, are, in some comparisons, as low as 55% identical. Sequence analysis has established that, again using the criterion of 90% nucleotide identity, the 11 longest copies can be classified as belonging to 9 different subfamilies. We have found long, almost identical copies in only two cases (sufamilies V and VII; see table 2). Ten of the 11 most complete sequences have 5′-truncated versions, the one exception being *Sam6,* which is truncated in its 3′ end. Regarding the 3′ ends of the elements, it turns out that each of the *Sam* subfamilies, like those of *Frodo,* have different 3′ repetitious ends (table 2). Also like *Frodo* sequences, there are over 20 other *Sam* sequences not included in any of these subfamilies; about half of them belong to two other subfamilies whose longest available copies are found in the clones CELF55A4 (nu-

```
                     *        20         *        40         *
T1        : LSLSLGVFPALWKSCWLFPVHKKGC-RSIVSNYRGITQTCATAKTFELCIFP :  51
Q         : ASLANGYFPKAWRKSWMVPIYKKGD-RTDAINYRGITSLCAIAKVEELVIYK :  51
CR1P      : XSWRLGEVPDDWKKANIVPIFKKGK-KEDPGNYRPVSLTSXPGKIMEQVLKE :  51
CR1G      : KSWLSGEVPDDWRKGYVTPIYKKGS-KEDPGNYRPVSLTSVPGKIMEQILLD :  51
CR1X      : DSLRSGIVPRDWRIANVVPLFKKGS-RSQPQNYRPVSLTSVLGKLFEGLIKD :  51
SR1       : HSLSRGKLPEIWKLAHITPIFKGGR-RSEPSSYRPVALLSIPSKIMESLVFD :  51
SAM1      : ESFSSSFFPTPWKTSIIVPLHKKGS-TTNPGNYRPITLTHPLSRLFERVVLD :  51
SAM2      : ESYDSGRFPEIRKTSVVIPVHKKGK-MCDAKNFRPISLTHPLSRLFEKLIMD :  51
SAM3      : ESYASGQFPEIWKSSIIVPVHKKGC-RTDANNYRPISLTHPLSRVFEKFIVE :  51
SAM4      : EFLTSKSFPSCWKTSIVVPVYKKGN-SSDPKNFRPIALTHPLSRLFERIVLN :  51
SAM5.1    : EFITTNSFPSCWKTSIVIPVHKKGN-SSDPKNYRPIALTHPLSRLFERIVLN :  51
SAM6      : EFLSTKTFPVCWKESTVVPIFKKGN-ASDPENYRPISLTHPLSRLFEKLVLQ :  51
SAM7.1    : HSLQTSSVPKAWKHAFIIPVPKKGN-LAEPCNYRPISITDPFCRVFERILCT :  51
SAM8      : QSFADSKIPDCWKKAIIIPIPKKGN-PSSPSNYRPISLTDPFVRVMERIICS :  51
FRODO1    : ISMMRGEVPECWRRSIVIPLNKVPK-PKLASQFRPISLTSHLCRIYERCLLK :  51
FRODO2.1  : LSMMHGVVPRSWTQSVVITLSKVPK-PKESVDFRPISITSNICRVYERCLLR :  51
Jockey    : SVLDVGYFPKAWKSASIIMIHKTGKTPTDVDSYRPTSLLPSLGKIMERLILN :  52
                    P  w    6 p6 K g        5Rp 3 t    2  E  6

                  60         *        80         *       100
T1        : TILHSCSS--AISPKQHGFMPGRSTSTNLMSFVTNIFRSFEAGTQ-LDAIYT : 100
Q         : NLLHACRS--YLSPYQHGFVPKKST-TNLVEFVTYCTSQIDAGAQ-VDAIYT :  99
CR1P      : SILRHLEERKVIRXSQHGFTKGKSCLTNLIAFYEEVTGSVDVGKA-VDVIYL : 102
CR1G      : DMLDHMRNERVIRDSQHGFTRGRSCLTNLVAFYDGVTALVDEGKA-TDVIYL : 102
CR1X      : KILDFIANHNTMSLCQHGFMRNRSCQTNLISFYENVSRDLDSGMA-VNVIYL : 102
SR1       : GILEHLSSSKFFSPQQHGFRKGHSCTTNLLTAVDKWTTILDRKGK-VDVIYL : 102
SAM1      : ALKQKLSPL--LSSKQFGFLPARSCTLAVLESTSKIQKVLLDKSKYVDTIYF : 101
SAM2      : KLRSQSISK--LSRFQFGFMNSRSCTLALLNSISKAYEILSNRSNYVDMVYF : 101
SAM3      : KLRKECSSK--ISKSQFGFMNSRSCTLALLNACSKILDSLTIRSKYVDAIYL : 101
SAM4      : PIMKTFSPK--ISKFQFGFLNNRSCTTALLNASSIYHKILADPNQYIDIVLF : 101
SAM5.1    : PTRKTFSHK--ISKFQFGFLNNRSCITAHLNASSNYHKILSVPNQFVDIILF : 101
SAM6      : SIKKTCANK--LSKIQFGFLNNRSCTLAHLNSSTFCHSILSHPRKFLDIVLF : 101
SAM7.1    : RIKIDFAHL--FSKFQHGFLARRSCTNSLVLSHANYYHIL-HAHKSLYVIFF : 100
SAM8      : RIRADFXPP--PPPHQHGFLHYRSCPSSLVRSVSLYHSLLKKEKS-LDIIFF : 100
FRODO1    : QILPYLEEKHFWSNAQHGFRPKKSTITCTLEAVNDWSDAIDTGNQ-VDVIYL : 102
FRODO2.1  : KLEEFLKNENFWSPTQHGFRRNKSTITNLLDSLNYWTQSIDEGAQ-VDIIYL : 102
Jockey    : RLLTCKDVTKAIPKFQFGFRLQHGTPEQLHRVVNFALEAMENKEY-AVGAFL : 103
                           Q GF    s

                 *       120         *       140         *
T1        : DFHAAFDSLPHSLLLAKLSKLGFGDGIISWLSSYLSNRSCRVKTGSYLS--E : 150
Q         : DLKAAFDSLPHAILLAKLDKLGIPSPLVQWLKSYLIHRTYIVKIDKHMS--K : 149
CR1P      : DFSKAFDTVSHSILISKLKKYGLDEWTIKWIENWLXHRAQRVVINGSMS--S : 152
CR1G      : DLTKAFDMVPHHILISKLEGCGFDGWTTHWIRNWLKGRRQRVVTNGSMS--R : 152
CR1X      : DFAKAFDTVPHKRLLVKLRNVGLEHSICTWIENWLKDRLQRVVVNGTFS--N : 152
SR1       : DFSKAFDRVNHACLIKKLGRLGIKPPLIDWITSYLENRHFKVRINFTLS--Q : 152
SAM1      : DFKKAFDSVPHNLLLLKLLNVGLDSGCCEWISSFLSNRSSKIKIDHFIS-DN : 152
SAM2      : DFRKAFDCVPHNLLLCKMQLFGMDQKICNWFQSFLCGKXSIIKVGDYVSDQN : 153
SAM3      : DFKKAFDSVPHNLLLCKLELFGLDVKMCNWFRSFLSNRTSSIKVCDHVS-KN : 152
SAM4      : DFAKAFDSVPHELLLSKIRNFGVDDNLCAWFRSFLSNRVSKVKIGDCLS-DS : 152
SAM5.1    : DFAKAFDSVPHELLLCKLRNFGFDDNICLWFKAFISDRVSKVKIGDSLS-NS : 152
SAM6      : DFRKAFDTVPHNLLILKLKNFGIDDNLCDWFQSFISNRTSTVKVSNSMS-TC : 152
SAM7.1    : DFQKAFDRVPHSLLLQKLELFGIPLPIINWFGDFLPERSFSVKVNDSIV-KS : 151
SAM8      : DFAKAFDKVSHSILLGKLAKFGLDSLTCSWYKEFLENRTFSVKINKFVS-KN : 151
FRODO1    : DFAKAFDRIPHCQLLDRLINLKINKNLINWIIAFLTNRVFPVKVGDSLS--S : 152
FRODO2.1  : DYSKAFDRIAHDLLLEKLVKIGLNKNLLIWFKQFLANRTFRVKIGNTLS--S : 152
Jockey    : DIQQAFDRVWHPGLLYKAKRL-FPPQLYLVVKSFLEERTFHVSVDGYKS--S : 152
            D   AFD 6 H  L6 2    g       w   56  2   6      s
```

FIG. 3.—Alignments of the RTs of the *T1/CR1* subgroup elements, including members of the different *Sam* and *Frodo* subfamilies, plus the more distantly related non-LTR retrotransposon *Jockey*. These alignments were used to build the tree in figure 4. Conventions are as in figure 1.

```
             160         *       180         *       200
T1        : EFFCTSGVPQGCVLSPLLFSLFINDVCNVLPPDG-------HLLYADDIKIF : 195
Q         : EIVSSSGVPQGSNIGPLLFILFINDVTLALPPDS-------ISLFADDAKIF : 194
CR1P      : WQPVTSGVPQGSVLGPVLFNIFINDLEDGVDCT--------LSKFADDTKLG : 196
CR1G      : WRPVMSGVPQGSVLGPVLFNIFINDIDDGIECT--------LSKFADDTKLS : 196
CR1X      : WTSVVSGVPQGSVLGPLLFNLFINDLEVGIEST--------VSIFANDTKLC : 196
SR1       : AMECPSGVPQGSILGPLLFLIYINDLPQQVSSD--------LLLFADDVKLW : 196
SAM1      : SFNVMSGVPQGSVTGPFLFLLYINDLLDLFPPDVH------VTAFADDLKLL : 198
SAM2      : YIDVISGVPQGSVSGPFLFLIYINDLLELIPSDVH------VYAFADDIKLL : 199
SAM3      : KLEVLSGVPQGSVCGPFLFLIYINDLLGMLPPDVQ------ISAFADDIKIY : 198
SAM4      : SYKNSSGVLQGTVTGPFLFLIYINDILEQFPPDVH------AMAFADDLKLF : 198
SAM5.1    : SFKNSSGVLQGTVTGPFLFLVYINDLLDLFPPDVH------VIAFADDLKLF : 198
SAM6      : KYSISSGVLQGTVTGPFLFLIYINDLLEQFPADVH------VTAFADDVKIS : 198
SAM7.1    : SQPIISGVPQGSVSGPLLFLIFINDLLLSLPTELC------FCAFADDIKLY : 197
SAM8      : SYPISSGVPQGSVSGPLLFILFINNLLIDLAPTIN------ISCFADDVKIF : 197
FRODO1    : FKRADCGVPQGXVLSPLLFGIFVNEIPNILPPAIK------CKQFADDLKLY : 198
FRODO2.1  : VKQAICGVPQGAVLSPVLFGMFVNEISSILPENVQ------CQQFADDTKLY : 198
Jockey    : IKPIAAGVPQGSVLGPTLYSVFASDMPTHTPVTEVDEEDVLIATYADDTAVL : 204
                  GVpQG   gP L5 65 n16                  5A1D k6

             *       220         *       240         *       260
T1        : LPVSS-SSDCMSLQHYLNAFVHWCSSNLLRLCPDKCSVISFSHSLSPISFNY : 246
Q         : APINN-TGDCTFLQDCIEIFCSWCKRNGLTICIEKCYCVSFSRCRSPVTGTY : 245
CR1P      : GVVDT-LEGRDRIQKDLNKLEDWAKRNLMRFNKDKCRVLHLGW--KNPMHSY : 245
CR1G      : GAVDT-EEGRDAIQRDLDRLERWARVNLMRFNTAKCRVLHLGW--RNPRHLY : 245
CR1X      : RTIGS-MQDAATLQSDLSKLENWAANWKMRFNVDKCKVMHFGK--NNINASY : 245
SR1       : REIRT-HNDILVLQEDLTRLQSWVDDNGLTFNTSKCKVVHLR---HVADHSY : 244
SAM1      : ------GSDPTSIQTSINIVADWCKKWRLNLAEHKTAVLHFGK--QNPRHKY : 242
SAM2      : ------GDNVSAIQKSINVVTDWCAKWKLNLAENKTNVIHFGK--RNPKNEY : 243
SAM3      : ------GDNSNSIQKSIDIVTDWCRKWSLNLAENKSVVVHYGK--NNPKFVY : 242
SAM4      : ------SNNCKSLKSSISLIEAWCTKWQLNLAENKTNVLHIGK--KNPKCEY : 242
SAM5.1    : ------SNNCQSLRSSIKIIENWCKIWQLKLAENKTKVLHVGK--KNPKYKY : 242
SAM6      : ------SENIESIKKSISIIEQWCDIWKLKLAENKTQVLHIGK--LNPKTDY : 242
SAM7.1    : ------SNDHVVLQKGIDTVVEWSISNSLPLAHAKTALLRLGS--KNPVHPY : 241
SAM8      : ------HTDPTIIQNSIDIIVSWSKLNELPLAPTKSALLALGT--RNKNQSY : 241
FRODO1    : TAIPSNSNNNVHLQKAIETIVEWSKSTKLALNNDKTVCISLGR--NTTEFQY : 248
FRODO2.1  : VKTPKKENENK-LQLAINLVSEWSKLSKLDLNNSKTVHMTLGT--NKKDFSY : 247
Jockey    : TKSKSILAATSGLQEYLDAFQQWAENWNVRINAEKCANVTFAN-RTGSCPGV : 255
                        62  6     W     6     K   6            y

                     *       280
T1        : TLSNSSLSRVLSIRDLGIIL : 266
Q         : FMDGTAVNRQNHAKDLGVLL : 265
CR1P      : RLGTDELGSSSAEKDLGV-- : 263
CR1G      : RLEGAVLESSSAEKDLGVLM : 265
CR1X      : TLNGNVLGVSLNEKDLGVFV : 265
SR1       : NLGNSPLEVSQVEKDLGVLV : 264
SAM1      : FVNSVEIKPRDSIRDLGIIV : 262
SAM2      : YANGMKVTKKDSVKDLGIFV : 263
SAM3      : TANGIIIAKKKSVKDLGIFV : 262
SAM4      : YVNGQKIHSCSKARDLGIWI : 262
SAM5.1    : YVNGQNIECCSKACDLGIWV : 262
SAM6      : LVNGHKISVCSKARDLGIWV : 262
SAM7.1    : LLQNKLIHETATVRDLGLIT : 261
SAM8      : SVDGVPITPSSTVRDLGLIT : 261
FRODO1    : TIENSLITRSTIVRDLGXXX : 268
FRODO2.1  : SINGQTIRKEDVARDLDFLI : 267
Jockey    : SLNGRLIRHHQAYKYLGITL : 275
                  6       dLg
```

FIG. 3 *(Continued)*

cleotides 27061–29672) and CEF40F12 (nucleotides 26767–28355) respectively, while the rest are very short copies of uncertain classification.

To determine the lengths of the elements, we performed detailed nucleotide comparisons of the regions upstream of the EN and downstream of the RT. These comparisons established that a low level of similarity among some *Sam* sequences extends at least 900 bp upstream of the sequence encoding the N-terminal end of the EN. The similarity also extends 150 bp downstream of the end of the ORF2, demonstrating that the minimum size of full-length *Sam* elements would be about 3,800 bp. When we searched for an ORF1, we were able to determine that in several *Sam* sequences, the region immediately upstream of the EN may encode a protein. An ORF can be postulated in those sequences that would start about 800–900 bp upstream of the EN, i.e., very close to the point where the similarity among *Sam* copies ceases to be detected, and it would end immediately adjacent to where the ORF2 starts. Figure 1 shows that the putative ORF1 products of several *Sam* elements are related, most clearly in their C-terminal ends, while the N-termini are much more variable. None of these putative ORF1 proteins showed significant similarity to any other sequence in BLASTP or TBLASTN database searches.

Because non-LTR retrotransposons often cause direct duplications upon insertion, we analyzed whether the sequences immediately adjacent to the 3′ ends of each putative master copy of *Sam* and *Frodo* are also present in regions upstream of the EN. We were able to detect such direct repeats for only three *Sam* copies. Around *Sam5.1,* six 150-bp repeats are detected, with one of them situated immediately adjacent to the 3′ end of the element (nucleotides 890–1125 in CEF58D12) and the other five situated 3.7 kb away, upstream of the EN and the putative ORF1 of the element (nucleotides 4847–5670). Assuming that these repeats mark the ends of *Sam5.1,* it would extend from nucleotide 1126 to nucleotide 4846 in CEF58D12, thus being 3721 bp long. These types of multiple repetitions at both sides of an element are not found in any other copy of *Sam* or *Frodo.* The only two other copies with a detectable single direct repeat at both sides are *Sam7.1* and *Sam7.2. Sam7.1* has 10 bp-repeats in positions 7170–7179 and 10620–10629 of CEZK337, which defines a 3,439-bp element. For *Sam7.2,* 105-bp repeats define a 3,992-bp element that would start in the clone CEF38B2 (nucleotide 16477) and end in the adjacent clone CEF08G12 (nucleotide 720). These results suggest that *Sam7.1* may be a 5′-truncated copy, and they explain the lack of a detectable ORF1 in *Sam7.1,* while the intimately related sequence *Sam7.2* probably has a full-length ORF1 (fig. 1).

Phylogenetic Position of *Sam* and *Frodo*

Regarding the structures of these elements, a final result is that no RNAse H-related sequences were found. Therefore, only two protein domains (EN and RT) can be used for comparison with other retroelements. We aligned the putative ENs and RTs of *Frodo* and *Sam* and those of the other elements of the *T1/CR1* subgroup (for *SR1,* however, only the RT has been described), plus the sequences of seven other non-LTR retrotransposons. These elements were chosen according to three criteria: (1) they contain EN; (2) five of them (*Jockey, R1Bm, I, LINE1,* and *Tx1*) span the whole range of non-LTR retrotransposons, according to Xiong and Eickbush (1990); (3) the last two, *Juan A* and *NLR1,* were the elements outside of the *T1/CR1* subgroup with the highest scores in comparisons with the *C. elegans* sequences. A *Xenopus laevis* sequence (accession number AF027962) that seems to contain an intact RT very similar to that of *CR1* was also included.

Figure 2 shows the phylogenetic tree for the RT sequences of these elements obtained using the neighbor-joining method (see *Materials and Methods*), and including the RT of the LTR-containing retrotransposon *Copia* as an outgroup. This tree was obtained by considering only conserved residues and eliminating those positions at which any sequence showed gaps, a strategy similar to that used by Xiong and Eickbush (1990) for their general study on RT relationships. The zones included in our analyses are 95% congruent with the conserved domains for the RT proteins defined by Xiong and Eickbush (1990), the minor differences being caused by the different sets of sequences considered. As we can see in figure 2, the elements that we have defined so far as belonging to the *T1/CR1* subgroup are together, and bootstrapping results strongly support this branch. Figure 3 shows the alignment of the RTs of the *T1/CR1* elements known so far, including a representative of each *Sam* and *Frodo* subfamily. We obtained similar trees using the EN sequences (not shown), but the information provided by this protein is not sufficient, according to the bootstrapping results, to obtain a well-supported topology. In particular, there is some ambiguity regarding the deepest branches, which separate the less-related groups of elements. This difference is probably due to the fact that ENs are evolving much faster than RTs. In any case, it is worth mentioning that the elements of the *T1/CR1* subgroup are also clustered together in the EN-based analyses, forming one of the best-supported branches (bootstrapping results: supported in 721 of 1,000 cases).

To establish the relationship among the *C. elegans* elements more precisely, we also used comparisons of their RT sequences. We used the 12 long EN-containing *Sam* and *Frodo* copies in which the RTs are not truncated to obtain the tree presented in figure 4. The relationships among the different *Sam* and *Frodo* sequences shown are well supported according to the bootstrapping results. Apart from the expected split among *Frodo* and *Sam* elements, we can also see that the *Sam* elements are divided into two groups that include, respectively, subfamilies I–VI and subfamilies VII and VIII. Within these groups, the RTs are, on average, 58% identical and 76% similar; between groups, however, the values drop to 41% identical and 62% similar. According to a comparison of this tree with the previous one in figure 2 and according to the bootstrapping data obtained, it can easily be seen that the precise phylogenetic positions of

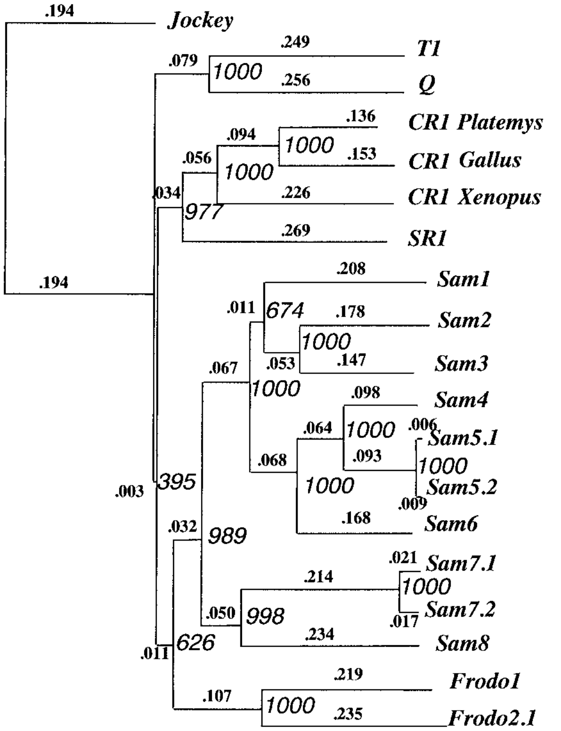

FIG. 4.—Phenogram obtained for the *T1/CR1* subgroup elements using the RT alignments shown in figure 3. Branch lengths and bootstrapping results are detailed as in figure 2.

Sam and *Frodo* inside the *T1/CR1* subgroup are ambiguous.

Discussion

In this work, we described two new non-LTR retrotransposons in *C. elegans.* Because no significant similarity has ever been observed in BLAST analyses, we can conclude that they are very different from *Rte-1,* the only non-LTR retrotransposon previously characterized in that nematode species (Youngman, van Luenen, and Plasterk 1996). These two elements, *Sam* and *Frodo,* are most closely related to a few other retroelements, forming what we call the *T1/CR1* subgroup. This subgroup so far includes members of the genomes of vertebrates, insects, nematodes, and trematodes. We have shown that this subgroup can be unambiguously defined by analysis of the RT sequences of those elements: when the RTs of non-LTR retrotransposons are compared, all the elements of this subgroup are grouped together in a branch that is highly supported by bootstrap analysis (fig. 2). This result confirms the BLAST analyses, in which comparisons among all these elements systematically produce scores that are much higher than those found for other non-LTR retrotransposons, both when their ENs and when their RTs are used for query purposes.

Interestingly, the topology of the tree presented in figure 2 is identical to that found by Xiong and Eickbush (1990), although those authors considered only one element of the *T1/CR1* subgroup, namely *T1* itself. This topology is also identical to that obtained by other authors who studied individual *T1* subgroup elements (see, e.g., Drew and Brindley 1997; Kajikawa, Ohshima, and Okada 1997). Kajikawa, Ohshima, and Okada (1997), during a study describing the *CR1* element of the turtle *Platemys spixii,* detected a copy of *Sam* (*Sam3,* which they called “CeCRT”). Their analysis situated *Sam3* in exactly the same topological position in which *Sam1* appears in figure 2, i.e., together with *T1* and *Q.* However, the bootstrapping results show that the particular topology for the *T1/CR1* subgroup presented in figure 2 is not well supported. In fact, when all of the *Sam* and *Frodo* subfamilies are included in the comparisons (fig. 4), it becomes clear that the precise phylogenetic position of the *C.elegans* elements with respect to the other members of the *T1/CR1* subgroup is uncertain.

We determined that both *Sam* and *Frodo* encode EN and RT proteins. We also showed that *Sam* and *Frodo* sequences are highly heterogeneous, and, when the two elements are taken together, up to 16 subfamilies can be defined according to their nucleotide sequences. We found not only that the subfamilies that we studied more precisely (those with at least one copy containing EN) are substantially different in their sequences, but also that one significant structural feature, their 3′-terminal repetitious sequence, is variable. The large amount of information provided by the available *Sam* sequences has allowed us to conclude that it is very likely that active *Sam* elements have an ORF1. For *Frodo,* however, the available sequences do not provide enough information to determine the precise structure of the 5′ end. For example, it is unclear whether *Frodo* has an ORF1. Although it is possible that *Frodo,* like other non-LTR retrotransposons, lacks an ORF1, this seems unlikely considering that all the best characterized elements of the *T1/CR1* subgroup seem to have one (Besansky 1990; Haas et al. 1997; Kajikawa, Ohshima, and Okada 1997) and we have found that *Sam* also probably has one. There are two ways to explain our difficulty in finding a *Frodo* ORF1. First, it is possible that the sequences upstream of the EN are evolving so rapidly that it is very difficult to detect any similarities among elements in that region, especially if there are only a few quite divergent copies available. As we pointed out in the introduction, for the best analyzed non-LTR retrotransposon, *LINE-1,* it has been shown that different subfamilies often differ in their 5′ ends. In particular, differences in ORF1 have been found in different *LINE-1* subfamilies (Schichman et al. 1992; Adey et al. 1994). Alternatively, it is conceivable that all the *Frodo* copies analyzed are just truncated elements and that complete ones are not yet in the databases, or even that they do not exist in the *C. elegans* strain used for genome sequencing (or in the species). In any case, our rigorous structural analysis has provided a clear picture of which ones are the best candidates for active copies.

An interesting point to consider is that our definition of only two new *C. elegans* elements may be very conservative. The fact is that the members of some subfamilies of *Sam* and *Frodo* are very divergent at the nucleotide level. We have also shown that the 3′ ends of elements of each subfamily are characterized by totally different terminal sequences. Therefore, we could have proposed the existence of several *Sam*- and *Frodo*-like elements by defining each subfamily as a different element. We preferred to define only two elements, because the structures of many of the copies found suggest that they are inactive. By structural analysis, it is impossible to precisely determine how many sequences correspond to active elements, and, thus, to define several elements would be premature.

However, the fact that there are truncated versions for each of the sequences that most closely resemble active elements demonstrates that relatively recent insertional events have occurred for each subfamily. That several *Sam* subfamilies were active recently, and may be still active, in the *C. elegans* genome is also supported by comparisons between the number of substitutions per synonymous site (K_S) and the number of substitutions per nonsynonymous site (K_A). For example, we have found, according to Li (1993), that $K_S = 1.84$ and $K_A = 0.10$ when the RT domains of *Sam4* and *Sam5.1* are compared. When we compared the RT domains of *Sam4* and *Sam6*, the values obtained were similarly biased ($K_S = 2.62$, $K_A = 0.22$). We found that these comparisons can be established only between elements of closely related subfamilies (when most sequences are compared, the number of substitutions in fourfold degenerate sites is so high that it precludes the determination of K_S and K_A). However, it is possible to determine for the RT domain of all the *Sam* sequences described in figure 3 that the proportion of synonymous changes is larger (ranging from 0.72 to 0.91) than the proportion of nonsynonymous changes (range 0.12–0.44; calculations according to Nei and Gojobori 1986). This large excess of synonymous changes suggests that all these sequences have been under purifying selection (and therefore were part of active elements) recently. It is thus very likely that several highly divergent master copies of these elements are still active in *C. elegans*, as has been suggested for *LINE-1* in several species (Kass, Berger, and Dawson 1992; Stanhope et al. 1993). It is therefore possible that the low number of potentially active elements found in our analysis is a peculiarity of the stock used for genome sequencing.

If functional studies demonstrate that elements of several subfamilies are indeed active, it would obviously be necessary to reconsider the classification presented in this work. For future studies, to avoid confusion, we suggest describing the *Frodo*- and *Sam*-related sequences with reference to the different subfamilies described in this work, and using the nomenclature that we have developed, i.e., with a name plus one or several numbers (*Frodo1*, *Sam5.1*, etc.). This will help to make clear which main lineage and which precise sequence we are under discussion. This nomenclature can be used even if we finally define several elements. The only problem regarding our convention is that *Sam3*, and, in general, the members of subfamily III of *Sam*, may be more properly called *CeCRT*, due to the priority of its description by Kajikawa, Ohshima, and Okada (1997). We suggest the use of ‘‘*CeCRT/Sam3*’’ for the members of this particular subfamily.

The situation discovered for *Sam* and *Frodo* is probably not exceptional. It is perfectly possible that if exhaustive searches are performed in other species, some described elements will simply correspond to highly diverged, and often inactive, copies of other elements. A possible example is the *G* element of *Drosophila melanogaster*, so far described only by inactive, highly degenerated copies that are quite similar to another element, *F* (Di Nocera 1988). In any case, all these complications emphasize the fact that the conventional molecular biology methods used to define and characterize retrotransposons may in some cases be providing a distorted picture of the evolutionary dynamics of these elements. For example, it is unlikely that the different subfamilies of the *C. elegans* elements that we characterized could be detected in standard DNA–DNA hybridization experiments unless multiple probes and very low stringency conditions are used.

In our previous study (Marín and Fontdevila 1996), we discussed whether the finding of *D. koepferae* sequences solely related to the *Anopheles* elements *T1* and *Q* but not to other *Drosophila* non-LTR retrotransposons could be a sign of horizontal transmission between dipteran species. In the last 2 years, the situation has changed substantially, with the characterization of new elements of this subgroup. This work, together with the description of complete protein sequences for *CR1* elements of vertebrates (Haas et al. 1997; Kajikawa, Ohshima, and Okada 1997), establishes that elements of the *T1/CR1* subgroup are found in many organisms. Because the *D. koepferae* sequences do not seem closer to *T1* or *Q* than to other elements of this subgroup (data not shown), we do not consider that our previous results have to be explained by horizontal transmission; we now think that it is more likely that *T1/CR1* subgroup elements are, in general, ancient components of the genome of all of these species. However, our comparative study of these elements actually gives some new support to the hypothesis that the element *SR1* of the trematode *Schistosoma mansoni* has been horizontally transmitted, as suggested by Drew and Brindley (1997). As we can see in figures 2 and 4, elements of the *T1/CR1* subgroup have been found in both nematodes and insects, organisms phylogenetically closer to vertebrates than are trematodes (reviewed in Freeman and Herron 1998). However, they are more distantly related to the vertebrate *CR1* elements than to the *Schistosoma SR1* element, which is a very close relative. Two simple explanations, among others, for these relationships follow: (1) The elements of the *T1/CR1* subgroup described so far in dipterans and *C. elegans* diverged long ago from the relatives of *CR1*, while an element very similar to *CR1* was present in vertebrates, in trematodes, and in the nematode-insect clade. This *CR1* element was subsequently lost in both nematodes and insects, or in the nematode-insect ancestor. (2) The profound differences between the vertebrate *CR1* elements and the *T1/CR1* subgroup elements found in nematodes and insects simply reflect the time of divergence between vertebrates and the nematode-insect clade. If this is the case, any element of the *T1* subgroup in *Schistosoma* should be even more divergent, and thus the *SR1* element must have been acquired recently. Which one of these two hypotheses (or another, more complex, hypothesis that includes different rates of sequence evolution; see Capy, Anxolabéhère, and Langin 1994) is correct will be known only after a careful study of the phylogenetic range of these and other elements of the *T1/CR1* subgroup.

In this work, we have demonstrated how the information generated by the genome projects can be advantageously used for studies on the evolution of retrotransposable elements. However, it is evident that the characterization presented in this study can be extended by more detailed comparisons of the multiple sequences of *Sam* and *Frodo* that are available in the databases. Precise studies such as those already performed for *LINE-1* (in different mammalian species) and for *CR1* (as in Vandergon and Reitman 1994) are possible for *Sam* and *Frodo.* In addition, when new sequences of this species are available, they could further contribute to an understanding of the evolutionary dynamics of these two elements. Finally, we look forward to the functional characterization of these elements, which can be greatly accelerated by using our structural data to devise strategies to screen for active elements.

Acknowledgments

This work was supported in part by grant PB96/1136 from DGES (Spain), awarded to A.F. We would like to thank Thomas Eickbush and an anonymous reviewer for their comments, which greatly improved the manuscript. We are also grateful to Eladio Barrio and Santiago Elena, who helped us with the literature and programs used to establish the proportions of synonymous and nonsynonymous substitutions.

LITERATURE CITED

Adey, N. B., S. A. Schichman, D. K. Graham, S. N. Peterson, M. H. Edgell, and C. A. Hutchison III. 1994. Rodent L1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. Mol. Biol. Evol. **11**:778–789.

Adey, N. B., S. A. Schichman, C. A. Hutchinson III, and M. H. Edgell. 1991. Composite of A and F-type 5′ terminal sequences defines a subfamily of mouse LINE-1 elements. J. Mol. Biol. **221**:367–373.

Agarwal, M., N. Bensaadi, J.-C. Salvado, K. Campbell, and C. Mouches. 1993. Characterization and genetic organization of full-length copies of a LINE retroposon family dispersed in the genome of *Culex pipiens* mosquitoes. Insect Biochem. Mol. Biol. **23**:621–629.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403–410.

Besansky, N. J. 1990. A retrotransposable element from the mosquito *Anopheles gambiae.* Mol. Cell. Biol. **10**:863–871.

Besansky, N. J., J. A. Bedell, and O. Mukabayire. 1994. Q: a new retrotransposon from the mosquito *Anopheles gambiae.* Insect Mol. Biol. **3**:49–56.

Blinov, A. G., Y. Y. Sobanov, S. S. Bogachev, A. P. Donchenko, and M. A. Filippova. 1993. The *Chironomus thummi* genome contains a non-LTR retrotransposon. Mol. Gen. Genet. **237**:412–420.

Blumenthal, T., and J. Spieth. 1996. Gene structure and organization in *Caenorhabditis elegans.* Curr. Opin. Genet. Dev. **6**:692–698.

Burch, J. B., E. D. L. Davis, and N. B. Haas. 1993. Chicken repeat 1 elements contain a *pol*-like open reading frame and belong to the non-long terminal repeat class of retrotransposons. Proc. Natl. Acad. Sci. USA **88**:5814–5818.

Capy, P., D. Anxolabéhère, and T. Langin. 1994. The strange phylogenies of transposable elements: are horizontal transfers the only explanation? Trends Genet. **10**:7–12.

Chen, Z.-Q., R. G. Ritzel, C. C. Lin, and R. B. Hodgetts. 1991. Sequence conservation in avian CR1: an interspersed repetitive DNA family evolving under functional constraints. Proc. Natl. Acad. Sci. USA **88**:5814–5818.

Deininger, P. L., M. A. Batzer, C. A. Hutchinson III, and M. H. Edgell. 1992. Master genes in mammalian repetitive DNA amplification. Trends Genet. **8**:307–311.

Di Nocera, P. P. 1988. Close relationship between non-viral retroposons in *Drosophila melanogaster.* Nucleic Acids Res. **16**:4041–4052

Drew, A. C., and P. J. Brindley. 1997. A retrotransposon of the non-long terminal repeat class from the human blood fluke *Schistosoma mansoni.* Similarities to the Chicken-Repeat-1 elements of vertebrates. Mol. Biol. Evol. **14**:602–610.

Eickbush, T. H. 1992. Transposing without ends: the non-LTR retrotransposable elements. New Biol. **4**:430–440.

———. 1994. Origin and evolutionary relationships of retroelements. Pp. 121–157 *in* S. S. Morse, ed. The evolutionary biology of viruses. Raven Press, New York.

Felsenstein, J. 1989. PHYLIP—Phylogeny Inference Package (version 3.2). Cladistics **5**:164–166

———. 1993. PHYLIP (Phylogeny Inference Package). Version 3.5c. Distributed by the author (http:/evolution.genetics.washington.edu/phylip.html). Department of Genetics. University of Washington, Seattle.

Feng, Q., J. V. Moran, H. H. Kazazian Jr., and J. D. Boeke. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell **87**:905–916.

Finnegan, D. J. 1997. Transposable elements: how non-LTR retrotransposons do it. Curr. Biol. **7**:R245–R248.

Freeman, S., and J. C. Herron. 1998. Evolutionary analysis. Prentice Hall, Upper Saddle River, N.J.

Haas, N. B., J. M. Grabowski, A. B. Sivitz, and J. B. E. Burch. 1997. Chicken repeat 1 (CR1) elements, which define an ancient family of vertebrate non-LTR retrotransposons, contain two closely spaced open reading frames. Gene **197**:305–309.

Hayward, B. E., M. Zavanelli, and A. V. Furano. 1997. Recombination creates novel L1 (LINE-1) elements in *Rattus norvegicus.* Genetics **146**:641–657.

Kajikawa, M., K. Ohshima, and N. Okada. 1997. Determination of the entire sequence of turtle CR1: the first open reading frame of the turtle CR1 element encodes a protein with a novel zinc finger motif. Mol. Biol. Evol. **14**:1206–1217.

Kass, D. H., F. G. Berger, and W. D. Dawson. 1992. The evolution of coexisting highly divergent LINE-1 subfamilies within the rodent genus Peromyscus. J. Mol. Evol. **35**: 472–485.

Labrador, M., and A. Fontdevila. 1994. High transposition rates of *Osvaldo,* a new *Drosophila buzzatii* retrotransposon. Mol. Gen. Genet. **245**:661–674.

Li, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**: 96–99.

Marín, I., and A. Fontdevila. 1995. Characterization of *Gandalf,* a new inverted-repeat transposable element of *Drosophila koepferae.* Mol. Gen. Genet. **248**:423–433.

———. 1996. Evolutionary conservation and molecular characteristics of repetitive sequences of *Drosophila koepferae.* Heredity **76**:355–366.

Marín, I., M. Labrador, and A. Fontdevila. 1992. The evolutionary history of *Drosophila buzzatii.* XXIII. High content of nonsatellite repetitive DNA in *D. buzzatii* and in its sibling *D. koepferae.* Genome **35**:957–964.

Martin, F., M. Olivares, and M. C. López. 1996. Do nonlong terminal repeat retrotransposons have nuclease activity? Trends Biochem. Sci. **21**:283–285.

Moran, J. V., S. E. Holmes, T. P. Naas, R. J. DeBerardinis, J. D. Boeke, and H. H. Kazazian Jr. 1996. High frequency retrotransposition in cultured mammalian cells. Cell **87**: 917–927.

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

Nicholas, K. B., and H. B. Nicholas Jr. 1997. GeneDoc: a tool for editing and annotating multiple sequence alignments. Distributed by the author (www.cris.com/∼ketchup/genedoc.shtml).

Ohshima, K., M. Hamada, Y. Terai, and N. Okada. 1996. The 3′ ends of tRNA-derived short interspersed repetitive elements are derived from the 3′ ends of long interspersed repetitive elements. Mol. Cell. Biol. **16**:3756–3764.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Sassaman, D. M., B. A. Dombroski, J. V. Moran, M. L. Kimberland, T. P. Naas, R. J. DeBerardinis, A. Gabriel, G. D. Swergold, and H. H. Kazazian Jr. 1997. Many human L1 elements are capable of retrotransposition. Nat. Genet. **16**:37–43.

Schichman, S. A., D. M. Severynse, M. H. Edgell, and C. A. Hutchinson III. 1992. Strand-specific LINE-1 transcription in mouse F9 cells originates from the youngest phylogenetic subgroup of LINE-1 elements. J. Mol. Biol. **224**: 559–574.

Smit, A. F. A. 1996 The origin of interpersed repeats in the human genome. Curr. Opin. Genet. Dev. **6**:743–748.

Stanhope, M. J., D. A. Tagle, M. S. Shivji, M. Hattori, Y. Sakaki, J. L. Slightom, and M. Goodman. 1993. Multiple L1 progenitors in prosimian primates: phylogenetic evidence from ORF1 sequences. J. Mol. Evol. **37**:179–189.

Stanhope, M. J., D. A. Tagle, M. S. Shivji, M. Hattori, Y. Sakaki, J. L. Slightom, and M. Goodman. 1993. Multiple L1 progenitors in prosimian primates: phylogenetic evidence from ORF1 sequences. J. Mol. Evol. **37**:179–189.

Stumph, W. E., P. Kristo, M.-J. Tsai, and B. W. O'Malley. 1981. A chicken middle-repetitive DNA sequence which shares homology with mammalian ubiquitous repeats. Nucleic Acids Res. **9**:5383–5397.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Tolkien, J. R. R. 1954. The lord of the rings. George Allen and Unwin, London.

Vandergon, T. L., and M. Reitman. 1994. Evolution of Chicken Repeat 1 (CR1) elements: evidence for ancient subfamilies and multiple progenitors. Mol. Biol. Evol. **11**: 886–898.

Xiong, Y., and T. H. Eickbush. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. **9**:3353–3362.

Youngman, S., H. G. A. M. van Luenen, and R. H. A. Plasterk. 1996. Rte-1, a retrotransposon-like element in *Caenorhabditis elegans.* FEBS Lett. **380**:1–7.

Thomas H. Eickbush, reviewing editor

Accepted July 7, 1998