Structural and Evolutionary Analysis of the copia-like Elements in the Arabidopsis thaliana Genome

Javier Terol, Mari Cruz Castillo, Mónica Bargues, Manuel Pérez-Alonso, and Rosa de Frutos

Departamento de Genética, Facultad de Ciencias Biológicas, Universitat de València, Valencia, Spain

The analysis of 460 kb of genomic sequence of Arabidopsis thaliana chromosome III allowed us to identify two new transposable elements named AtC1 and AtC2. AtC1 shows identical long terminal repeats (LTRs) and all the structural features characteristic of the copia-like active elements. AtC2 is also a full copia-like element, but a putative stop codon in the open reading frame (ORF) would produce a truncated protein. In order to identify the copia-like fraction of the A. thaliana genome, a careful computer-based analysis of the available sequences (which correspond to 92% of the genome) was performed. Approximately 300 nonredundant copia-like sequences homologous to AtC1 and AtC2 were detected, which showed an extreme heterogeneity in size and degree of conservation. This number of copies would correspond to approximately 1% of the A. thaliana genome. Seventy-one sequences were selected for further analysis, with 23 of them being full complete elements. Five corresponded to previously described ones, and the remaining ones, named AtC3 to AtC18 are new elements described in this work. Most of these elements presented a putative functional ORF, nearly identical LTRs, and the other elements necessary for retrotransposon activity. Phylogenetic trees, supported by high bootstrap values, indicated that these 23 elements could be considered separate families. In turn, these 23 families could be clustered into six major lineages, named copia I–VI. Most of the 71 analyzed sequences clustered into these six main clades. The widespread presence of these copia-like superfamilies throughout plant genomes is discussed.

Introduction

DNA repetitive sequences represent the major fraction of plant genomes. In some species, they constitute up to 50% of the genome, contributing directly to evolution and host genome organization (Pearce et al. 1996a; SanMiguel et al. 1996). Plant genomes are especially rich in copia-like retroelements, which have been found in all the species that have been analyzed (Flavell et al. 1992, 1997; Voytas et al. 1992). The copia-like elements are usually present in high copy numbers; for example, more than 10^5 copies of BARE-1 have been estimated to be present in barley (Manninen and Schulman 1993), 10^5 copies in the rye genome (which constitutes 3.5% of the genome) (Pearce et al. 1997), around 10^5 copies in Vicia faba (Pearce et al. 1996a), and 10^4 copies in Avena (Linares, Serna, and Fominaya 1999). However, in some cases, such as that of rice, the copy number of copia-like elements may be as low as 100 per genome (Wang et al. 1999), and the Tpv2 family of Phaseolus vulgaris seems to consist of 40 copies (Garber et al. 1999). Most of the copia-like plant elements tend to be dispersed throughout the eukaryotic genomes, although their specific location depends on the plant species and the particular retrotransposon (Brandes et al. 1997; Kumar et al. 1997; Miller et al. 1998; Kumar and Bennetzen 1999). In contrast, some of them tend to be concentrated at the paracentromeric regions. For example, most of the 71 analyzed sequences are concentrated at the paracentromeric regions, as is the case for Arabidopsis thaliana, Allium cepa, and Cicer arietinum (Pearce et al. 1996b; Brandes et al. 1997; Heslop-Harrison et al. 1997). A striking feature of the copia-like elements is that they are present as highly heterogeneous populations in plant genomes. An extreme example was obtained by sequencing 31 fragments of the reverse transcriptase region from the potato. It was found that all the sequences were different, with similarities varying from 5% to 75% (Flavell, Smith, and Kumar 1992). Extreme heterogeneity in sequences in a wide range of species from both dicotyledons and monocotyledons have also been found (Flavell et al. 1992, 1997; VanderWiel, Voytas, and Wendel 1993; Kumar 1996; Matsuoka and Tsunewaki 1996, 1999; Kumar et al. 1997; Pearce et al. 1997; Wang et al. 1997; Kuipers, Heslop-Harrison, and Jacobsen 1998; Yáñez et al. 1998; Kumar and Bennetzen 1999). Most of the works show that within the genome of a species, a particular family of copia-like elements is usually composed of multiple subfamilies that could be considered analogous to retroviral quasispecies.

The small genome of A. thaliana (130 Mb) has a low content of repetitive DNA. The interspersed DNA fraction is especially low, constituting 2% of the genome (Meyerowitz 1992). During the last decade, some families of transposable elements have been described in this species: the non-long-terminal-repeat (non-LTR) retrotransposons Tal1-1 (Wright et al. 1996) and TSL (Chye, Cheung, and Xu 1997), the LTR-retrotransposon gypsy-like elements Tat1 (Peelman et al. 1991; Wright and Voytas 1998) and Athila (Périsier et al. 1995, 1996), the transposon-like elements Limpet1 (Klimyk and Jones 1997), Tag1 (Tsay et al. 1993; Frank et al. 1997), and Tag2 (Henk, Warren, and Innes 1999), the superfamilies Arnold and Harbinger (Kapitonov and Jurka 1999) and Basho (Le et al. 2000), and the foldback transposon Hairpin (Adé and Belzile 1999). Some families of repetitive elements structurally related to the miniature inverted-repeat transposable elements (MITEs) have also been described (Casacuberta et al. 1998; Surzycki and Belknap 1999; Feschotte and Mouches 2000; Le et al. 2000). Referring to the copia-like elements, 10 related families, designated Tal1–Tal10, were

Key words: copia-like elements, Arabidopsis thaliana, evolution of transposable elements.

Address for correspondence and reprints: Rosa de Frutos, Departamento de Genética Facultad de Ciencias Biológicas, Dr. Moliner 50, 46100 Burjasot, Valencia, Spain. E-mail: rosa.frutos@uv.es.

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038
considered a superfamily within the genome of *A. thaliana* (Voytas and Ausubel 1988; Voytas et al. 1990; Konieczny et al. 1991). Some sequences related to *Ta1-Ta10* elements have been reported (Brandes et al. 1997), and the divergence between members of these families is high. Voytas (1992) considered that the elements that share >95% nucleotide identity can be regarded as members of the same family, and those that share <85% can be considered members of different families. The copy numbers are low, with only a few copies per family, compared with other *copia*-like plant families. It has been estimated that the *Ta1-Ta10* superfamily constitutes 0.1% of the *A. thaliana* genome (Konieczny et al. 1991), with most of the elements located in clusters at the paracentromeric heterochromatin (Brandes et al. 1997). These data are in accord with works that show that the *copia*-like elements concentrate in the centromeric regions (Heslop-Harrison et al. 1997) and, in a more generic reference, that regions flanking the centromeres are densely populated by transposable elements (Copenhaver et al. 1999; Cold Spring Harbor Laboratory et al. 2000). *copia*-like elements other than those of the *Ta1-Ta10* superfamily, such as *Evelknievel* (Henikoff and Comai 1998), *Art1* (Hervé et al. 1999), *Metal* (Kapitonov and Jurka 1999), and *ArtRE1-ArRE2* (Kuwahara, Kato, and Komeda 2000), have been described within the *A. thaliana* genome. It has also been described that the presence of *copia*-like elements extends to the mitochondrial genome of this species (Knoop et al. 1996).

The complete sequence of *A. thaliana* chromosomes II and IV have recently been reported (European Union Arabidopsis Genome Sequencing Consortium et al. 1999; Lin et al. 1999). The analysis of those sequences has shown the low content of repetitive DNA of the *A. thaliana* genome when compared with other plant species. Dispersal repeats, which consist of predominantly LTR and non-LTR retrotransposons, are found throughout the chromosome arms. However, the main fraction of transposable elements is concentrated at the pericentromeric heterochromatin, as had previously been described. This region is constituted mainly by a few genes and a high density of presumably inactive mobile elements (Lin et al. 1999). In addition to the complete sequences of chromosomes II and IV, studies of different stretches of the *A. thaliana* chromosomes have been performed (Quigley et al. 1996; Thompson, Schmidt, and Dean 1996; Comella et al. 1999; Terryn et al. 1999), as have computer-assisted analyses, looking for the presence of different families of transposable elements (Kapitonov and Jurka 1999; Le et al. 2000). In this paper, we report the results of a careful computer-based analysis performed in order to identify, characterize, and establish the evolutionary relationships among *copia*-like elements that exist in the *A. thaliana* genome.

**Materials and Methods**

**DNA Sequencing**

The P1 genomic clones T15C9, F11C1, F3A4, T20E23, and T12C14, from chromosome III, were completely sequenced in the frame of the European Arabidopsis Genome Sequencing Project. For each clone, a shotgun library was constructed and the DNA sequence was determined with an ABI automatic DNA sequencer (in collaboration with Sistemas Genómicos S.L.). The sequencing projects were managed with the Staden software package (Staden 1996), and a total of 460 kb were analyzed, with an average redundancy of 8.44 characters per position.

**Sequence Analyses**

The searches for open reading frames (ORFs) on the genomic clones were performed with the GENSCAN program (Burge and Karlin 1997) in collaboration with the Munich Information Center for Protein Sequences (MIPS). The GenBank (GB) and European Molecular Biology (EMB) databases were searched for sequence similarities using the BLAST programs at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/BLAST/) (Altschul et al. 1997). An estimate of the similarity between sequences was obtained with the GAP program from the GCG Software Package (Wisconsin University). Multiple alignments were performed with the CLUSTAL W program (Thompson, Higgins, and Gibson 1994). Genetic distances were calculated with the Poisson correction method (Nei and Chakraborty 1976) for amino acid sequences; the phylogenetic trees were constructed with the neighbor-joining (Saitou and Nei 1987) and UPGMA (Swofford and Selander 1981) methods, the bootstrap test was carried out with 1,000 iterations. These evolutionary analyses were performed with the MEGA platform (Kumar, Tamura, and Nei 1993).

**Results**

**Search and Characterization of *copia*-like Elements on 460 kb of Genomic Sequence**

Within the frame of an international consortium for sequencing of the *A. thaliana* genome, we determined the nucleotide sequence of the P1 genomic clones T15C9, F11C1, F3A4, T20E23, and T12C14, corresponding to the short arm of chromosome III (European Union Chromosome 3 Arabidopsis Sequencing Consortium et al. 2000). The accession numbers for these sequences are AL132970, AL132976, AL132978, AL133363, and AL162507. Information on performance of analysis and a detailed annotation of these database entries can be viewed at the Munich Information Center for Protein Sequences (MIPS; http://www.mips.biochem.mpg.de/proj/thal/). Our contribution to the sequencing of chromosome III has been the determination of the nucleotide sequences of the inserts of these five clones, which amount to a total of 540,430 bp. We considered it interesting to analyze the presence and distribution of known transposable elements on these sequences in an effort to understand their dynamic and evolution in the *A. thaliana* genome.

We used the GenScan program to predict the ORFs present on the sequences we had determined. We obtained a total of 113 putative ORFs, which is in agree-
significant similarity between coding for a hypothetical protein of 1,452 amino acids. Further analyses on them revealed copia-like retrotransposons. These comparisons revealed that AtC1 and AtC2 were named AtC1 (accession number AF287471) and AtC2 (accession number AF287472), and we performed further studies on them.

AtC1 was predicted as a single 4,359-bp-long ORF coding for a hypothetical protein of 1,452 amino acids. Computer searches of the GB and EMB databases found significant similarity between AtC1 and several known copia-like retrotransposons. These comparisons revealed that AtC1 contains all the amino acid domains found to be conserved among autonomously active retroelements. Both the amino acid conservation and the domain order served to identify AtC1 as a copia-like retrotransposon.

We analyzed the genomic sequence to determine the structure of this new element (fig. 1A). AtC1 has a long internal region of 4,629 bp, bounded by 335-bp LTRs, which present the canonical inverted repeats at their ends (TG-CA). The nucleotide sequences of the LTRs were found to be completely identical; a 5-bp direct repeat (CTGCT) flanking the LTRs could correspond to the target site duplication (TSD) (fig. 1B). The internal region contains one large ORF which could code for the genes gag and pol, encoding nucleic acid, protease, integrase, reverse transcriptase, and RNase H domains, respectively, in that order (fig. 1A). The primer-binding site is in accord with the one described by Gauss and Sprinzl (1983) for plant tRNA<sub>met</sub> (PBS) (fig. 1B). There is also a polypurine tract (PPT) at the end of the ORF. Thus, the structural analysis of AtC1 suggests that it could be an active element, as it displays all the main features described in the retrotransposons that have been shown to be able to transpose.

The prediction programs defined AtC2 as an ORF composed of two exons, with the protein deduced from this ORF being 1320 amino acids long. Comparisons of this protein sequence against the databases confirmed that AtC2 belonged to the copia-like family of retrotransposons. As the Ty-copia retroviral-like family is characterized by the encoding of a polyprotein with a single reading frame (Voytas and Ausubel 1988), we decided to check for the existence of the two exons predicted for AtC2. We found that a direct translation of the sequence beginning at the ORF initiation codon gave rise to a protein of 1,337 amino acids, 17 residues longer than the one predicted originally, with a stop codon at nucleotide position 1435. The alignments showed that the additional residues were conserved with respect to other copia-like proteins, confirming that the 51-bp intron predicted by the Gen Scan program was an artifact caused by the presence of the stop codon. The complete polyprotein could code for all of the amino acid domains described for copia-like elements; however, the presence of the stop codon would produce a truncated protein 478 amino acids long, which would lack the RT and RNaseH domains, indicating that AtC2 is not able to transpose autonomously. The LTRs are identical except for the presence of five indels in them, which causes a small difference in the sizes of the LTRs (3 nt). This difference in size and the fact that the LTRs have also lost the inverted repeats at their ends (fig. 1B) indicate that the element has been inactive for a long enough period to accumulate changes in its sequence.

The overall similarity at the amino acid level between AtC1 and AtC2 was 42.21%, although the conservation varied extensively along the sequence. The best conserved domains were the integrase (68.1% similarity) and the retrotranscriptase (60% similarity). The similarity data and the differences in the lengths and structures of the elements indicate that AtC1 and AtC2 diverged long ago, as the degree of conservation between them is similar to that found between these elements and the copia element from Drosophila melanogaster (43.2% and 44.15% similarity, respectively).

Search for Homologous Sequences

Database searches performed at the DNA and amino acid levels with the programs BLASTP and TBLASTN did not yield any sequences identical to AtC1 or to AtC2. However, the TBLASTN searches allowed us to identify two new copia-like sequences on clones T15G18 (accession number AC006567; located on chromosome IV) and K2N11 (accession number AB022213; located on chromosome V) that were very similar to AtC1 and AtC2, respectively. We named the
elements \textit{AtC3} (accession number AJ292423) and \textit{AtC4} (accession number AJ293575).

The BLAST P searches performed on the GB and EMB databases with \textit{AtC1} and \textit{AtC2} produced several hundreds of sequences related to the \textit{Ty-copia} family of retrotransposons, and we decided to study them in detail. Only a few of these sequences corresponded to previously described elements: \textit{Evelknievel} (Henikoff and Comai 1998), the \textit{Ta1} (Voytas et al. 1990; Konieczny et al. 1991) and \textit{AtRE1} (Kuwahara, Kato, and Komeda 2000) families from \textit{A. thaliana}, \textit{Hopscotch} from \textit{Zea mays} (White, Habera, and Wessler 1994), \textit{Bare-1} from barley (Manninen and Schulman 1993), \textit{Tnt} from tobacco (Grandbastien, Spielmann, and Caboche 1989), and \textit{SIRE-1} from soybean (Laten, Majumdar, and Gaucher 1998). Most of the other database entries corresponded to predicted proteins from \textit{A. thaliana} and had been produced by the automatic annotation of genomic clones. These sequences are the result of the work of the international consortium which is performing the sequencing of the genome of this model plant.

The \textit{copia}-like sequences we obtained showed extreme heterogeneity in length and degree of conservation. The size of the predicted proteins ranged from approximately 80 to 1,500 amino acids, which indicates the existence of many defective elements. Many proteins had been described as the products of ORFs, with a number of predicted exons that ranged from 2 to 8. We verified that, as it was the case with \textit{AtC2}, the predicted introns were artifacts caused by the presence of indels or stop codons. In this way, we established that the \textit{copia}-like polyproteins described as multixeonic ORFs indicated the existence of defective or mutated \textit{copia}-like elements that were unable to transpose autonomously. We decided then to select those \textit{copia}-like polyproteins coded by a single ORF in order to allow us to identify putative active elements.

We chose a total of 25 sequences which might correspond to active elements and studied them in detail. Eight of the elements belonged to \textit{copia}-like families previously described in \textit{A. thaliana}: \textit{Ta1} (Voytas et al. 1990; Konieczny et al. 1991), \textit{Evelknievel} (Henikoff and Comai 1998), and \textit{AtRE1} and \textit{AtRE2} (Kuwahara, Kato, and Komeda 2000). The remaining 17 sequences, including \textit{AtC1}, \textit{AtC3}, and \textit{AtC4}, were characterized in proof for the first time, and the 14 new elements were named \textit{AtC5–AtC18}. In spite of having described \textit{AtC2} as an element producing a truncated polyprotein, we included it in the study, as it presented most of the structural features of this kind of transposon and could be representative of a new \textit{copia}-like family.

Structural Characterization of the New \textit{copia}-like Elements

We identified the DNA sequences from which the analyzed polyproteins were derived and carried out a structural analysis in order to find all of the characteristic features of the \textit{copia}-like retrotransposons. The results are summarized in table 1, in which the known families \textit{Ta-1}, \textit{Evelknievel}, \textit{AtRE1}, and \textit{AtRE2} have been also included. The most striking observation is that of the high degree of heterogeneity between the 26 analyzed elements. They all have different sizes, ranging from 4,629 to 5,738 bp, more than 1 kb of difference in size. The 26 elements present many of the structural features that characterize the autonomously active elements. All them present LTRs, with 17 of them showing identical 5' and 3' repeats. In the remaining 9 elements (\textit{AtRE2}, \textit{AtC2}, \textit{AtC4}, \textit{AtC5}, \textit{AtC8}, \textit{AtC11}, \textit{AtC14}, \textit{AtC15}, and \textit{AtC17}), the 5' and 3' LTRs present small differences in size, mainly due to one-base indels. The inverted repeats are portrayed by 19 of our elements, while \textit{AtC2}, \textit{AtC8}, \textit{AtC11–AtC13}, \textit{AtC17}, and \textit{AtC18} present LTRs where the inverted repeats begin with the canonical TG sequence but have a 3' change. Only the elements from the families \textit{AtRE1} and \textit{Ta-1} present LTRs identical in size and sequence. The rest of the elements display LTRs very heterogeneous in size, ranging from 120 to 734 bp, and sequence, with no significative similarities between them. When we analyzed the genomic sequences adjacent to the LTRs we found that 17 presented direct repeats that could correspond to the TSD, suggesting recent transposition events. Finally, most of the elements presented both the primer-binding site (PBS) and the polypurine tract (PPT) (adjacent to the 5' LTR and before the 3' LTR, respectively), suggesting that they can transpose autonomously.

A multiple alignment with the 26 \textit{A. thaliana} polyproteins, the \textit{copia} element from \textit{D. melanogaster}, \textit{Hopscotch} from \textit{Zea mays}, \textit{Bare-1} from barley, \textit{Tnt} from tobacco, and \textit{SIRE1} from soybean was performed with the CLUSTAL W program (Thompson, Higgins, and Gibson 1994). The study of the alignment revealed that, in all but two cases, the predicted proteins displayed all of the highly conserved domains present in retrotransposons and that these domains were ordered as is usual in \textit{copia}-like elements: \textit{Gag} protein (Prats et al. 1988), \textit{protease} (Pearl and Taylor 1987), integrase (Johnson et al. 1986) (data not shown). Only the \textit{Gag} region of \textit{Ta1–Ta3} and \textit{AtC8} presented one change in the consensus zinc finger that characterizes this region (Prats et al. 1988), probably compromising the ability of these elements to engage in active transposition.

Copy Number and Chromosome Distribution

TBLASTN and BLASTN searches on the GB and EMB databases were performed for each of the elements studied in this work in order to find out how many identical copies existed in the \textit{A. thaliana} genome (table 1). Most of the elements were present as single copies in the genome: \textit{AtC3}, \textit{AtC5}, \textit{AtC7}, \textit{AtC13}, \textit{AtC15}, and \textit{AtC67} showed two identical copies in the same or different chromosomes, and only \textit{Evelknievel} and \textit{AtC10} could be considered multicopy elements, with three and six copies, respectively. Nevertheless, these data will have to be confirmed when the \textit{A. thaliana} genome is completely sequenced. The elements are widely dispersed on the five \textit{A. thaliana} chromosomes, with seven
copies on chromosome I, eight on chromosome II, seven on chromosome III, six on chromosome IV, and two on chromosome V. The smaller number of elements on chromosome V is probably due to the status of the sequencing project, with a smaller number of sequenced chromosome clones which carry the retrotransposons in- cludes in the databases. The chromosome positions of the genomic clones which carry the retrotransposons in-

Search for Expressed Sequence Tags

The DNA sequence of the newly described elements was used to perform searches on the expressed sequence tag (EST) databases in order to find the corresponding ESTs. Three of the analyzed copia-like sequences produced several identical ESTs: AtC7, with eight different ESTs; AtC10, with two ESTs; and AtC18, with 1 EST. The fact that AtC7 and AtC10 present both transcriptional activity and several copies (two and six, respectively) on the A. thaliana genome enforces the idea they might be active retrotransposons.

Phylogenetic Analysis

The comparison of the amino acid sequences of the polyproteins coded by the copia-like elements characterized in this work reflects the high heterogeneity found at the structural level. The degree of similarity varies from 99% to 41%. As has been mentioned above, the similarity was much higher when the conserved do-

Table 1
Summary of All the Features Analyzed in the Copia-like elements Studied in this Work

<table>
<thead>
<tr>
<th>Element</th>
<th>Accession</th>
<th>No.</th>
<th>Size (^a)</th>
<th>TSD (^b)</th>
<th>LTR (^c)</th>
<th>% ID (^d)</th>
<th>IR (^e)</th>
<th>PBS (^f)</th>
<th>ORF (^g)</th>
<th>PPT (^h)</th>
<th>CHR (^i)</th>
<th>CN (^j)</th>
<th>ESTs (^k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AtRE1-c</td>
<td>BAA78442</td>
<td>4,827</td>
<td>166 100</td>
<td>99</td>
<td>3 1</td>
<td>4,266</td>
<td>3 1</td>
<td>3,993</td>
<td>4 1</td>
<td>713</td>
<td>4,543</td>
<td>1 1</td>
<td>3</td>
</tr>
<tr>
<td>AtRE2-4</td>
<td>BAA78427</td>
<td>4,845</td>
<td>166 100</td>
<td>99</td>
<td>3 1</td>
<td>4,266</td>
<td>3 1</td>
<td>3,993</td>
<td>4 1</td>
<td>713</td>
<td>4,543</td>
<td>1 1</td>
<td>3</td>
</tr>
<tr>
<td>AtRE1-L</td>
<td>BAA78425</td>
<td>4,816</td>
<td>167 100</td>
<td>99</td>
<td>3 1</td>
<td>4,266</td>
<td>3 1</td>
<td>3,993</td>
<td>4 1</td>
<td>713</td>
<td>4,543</td>
<td>1 1</td>
<td>3</td>
</tr>
<tr>
<td>AtRE1-C</td>
<td>BAA78423</td>
<td>4,480</td>
<td>167 100</td>
<td>99</td>
<td>3 1</td>
<td>4,266</td>
<td>3 1</td>
<td>3,993</td>
<td>4 1</td>
<td>713</td>
<td>4,543</td>
<td>1 1</td>
<td>3</td>
</tr>
<tr>
<td>BtRE1-2</td>
<td>AAD26972</td>
<td>7,211</td>
<td>120 100</td>
<td>99</td>
<td>3 1</td>
<td>4,266</td>
<td>3 1</td>
<td>3,993</td>
<td>4 1</td>
<td>713</td>
<td>4,543</td>
<td>1 1</td>
<td>3</td>
</tr>
</tbody>
</table>

Note.

\(^a\) Size of the complete element, including the long terminal repeats (LTRs) (bp).
\(^b\) Target site duplication
\(^c\) Long terminal repeat length (bp).
\(^d\) Percentage of identity between the 5' and 3' LTRs.
\(^e\) Inverted repeats at the ends of the LTR.
\(^f\) Primer binding site for plant tRNA\(_{\text{Ama}}\) (Gauss and Sprinzl 1983).
\(^g\) Length of the open reading frame (bp).
\(^h\) Polyuridine tract.
\(^i\) Chromosomes where the element was found.
\(^j\) Number of identical copies.
\(^k\) Number of identical expressed sequence tags found.

Phylogenetic Analysis

The comparison of the amino acid sequences of the polyproteins coded by the copia-like elements characterized in this work reflects the high heterogeneity found at the structural level. The degree of similarity varies from 99% to 41%. As has been mentioned above, the similarity was much higher when the conserved do-

search into six major lineages, or families. Elements belonging to the same lineage showed similarities higher than 50%. The topology of the tree was supported by the high bootstrap values of the main branches. We decided to name the lineages copia I–VI.

The most homogeneous lineage is copia I, mainly composed of the AtRE1 and AtRE2 elements, which are
Fig. 2.—Phylogenetic tree constructed with the complete polyproteins from the complete *Arabidopsis thaliana* elements, as well as *Hopscotch* from maize (*ZMHOPS*), *Tnt* from tobacco (*TNT1*), and *SIRE*-1 from soybean (*GMSIRE1*). The proteins were aligned with the CLUSTAL W program (Thompson, Higgins, and Gibson 1994), the distances were calculated with the Poisson correction for amino acids, the tree was constructed with the neighbor-joining method, and the bootstrap values were calculated over 1,000 iterations. The six main lineages in which the sequences can be grouped are indicated with gray and white boxes.

almost identical. The other lineages are more heterogeneous both in structure and sequence, and they could be split into new lineages, as in the cases of *copia* II and V. For example, *copia* II could be divided into two new lineages, composed of *AtC6* and *AtC7*, and *AtC1*, *AtC5*, and *AtC4*, respectively. Following our previously established criteria, we maintained *copia* II as a single lineage, as the elements belonging to it present a similarity higher than 50%.

The tree also includes *copia* elements from different plant species: *Hopscotch* from maize (White, Habera, and Wessler 1994), *Tnt* from tobacco (Grandbastien, Spielmann, and Caboche 1989), and *SIRE*-1 from soybean (Laten, Majumdar, and Gaucher 1998). Each of these sequences group with one of the lineages we have described, indicating that the *copia* elements of those families are closer to the elements from different species than to the other *A. thaliana* families (fig. 2).

In an effort to clarify the structure and evolutionary relationships of the *copia*-like sequence population in *A. thaliana*, we identified the RT region used by Konieczny et al. (1991) in their study of the *Ta-I* family in all the *copia*-like sequences we analyzed. We performed a multiple alignment with the RT domains of 63 Arabidopsis polyproteins and included the 8 *TA* sequences (*TA2–* *TA10*) characterized in the above-mentioned work. The phylogenetic tree constructed with the alignment using the neighbor-joining method (fig. 3) displays a topology very similar to the ones obtained for the complete polyprotein, although the bootstrap values are much lower. We also obtained a similar cluster configuration when the UPGMA method was used to construct the tree. Most of the RT sequences could be assigned to the six major lineages described in figure 2, although there are a number them that could form other lineages different from the ones we have proposed.

**Discussion**

In this work, we performed an extensive analysis of the *copia*-like elements in the *A. thaliana* genome. To achieve this, we used information provided by the genome sequencing project, produced at our laboratory, and available at the databases. Early and recent works on *copia*-like elements in *A. thaliana* have been performed by cloning and sequencing genomic clones carrying *copia*-like sequences or by PCR amplification of conserved regions of the elements (Voytas and Ausubel 1988; Voytas et al. 1990; Konieczny et al. 1991; Kuwahara, Kato, and Komeda 2000), which has limited the
Fig. 3.—Phylogenetic tree constructed with the RT regions from 71 Arabidopsis thaliana copia-like retrotransposons. The RT sequences from the previously described elements and the ones described in this work are indicated by their names; the other sequences are indicated by their accession numbers. The sequences were aligned with the CLUSTAL W program (Thompson, Higgins, and Gibson 1994), the distances were calculated with the Poisson correction for amino acids, the tree was constructed with the neighbor-joining method, and the bootstrap values were calculated over 1,000 iterations. We have indicated the six major lineages we propose with gray and white boxes. The dashed lines on the right of the boxes indicate the sequences that might constitute additional major lineages.

The estimated size for the A. thaliana genome is 130 Mb and, to date, 120.7 Mb are available on the
Taiana genome. Previous data suggested that members of computer analysis represents at least 1% of the size of the used in the amplifications. If we assume that the mean differences can be explained because of the approach maintained the scope of the detection to sequences used in the previous works, for which the PCR technique limited the scope of the detection to sequences which had conserved the binding sites for the primers used in the amplifications. If we assume that the mean size of the copia-like sequences is between 4 and 5 kb and that the copy number is approximately 300, it can be deduced that the copia-like fraction detected in this computer analysis represents at least 1% of the Taiana genome. Previous data suggested that members of Ta1-Ta10 families constituted approximately 0.1% of the genome of this species (Konieczny et al. 1991). Our results agree with recently published works that suggest that the copia-like fraction would constitute 1% of the A. thaliana genome (Kapitonov and Jurka 1999). If this datum were confirmed, this percentage would constitute an extremely low proportion of the A. thaliana genome when compared with the copia-like components of other plants, such as barley (Manninen and Schulman 1993), maize (SanMiguel et al. 1996), rye (Pearce et al. 1997), and Avena (Linares, Serna, and Fominaya 1999). Several works have pointed out that there is a relationship between the overall total copy number of retrotransposons and the host genome size and that this could largely account for most genome size variation in plants (Katsiotis, Schmidt, and Heslop-Harrison 1996; Kumar et al. 1997; Linares, Serna, and Fominaya 1999). If this is the case, the low copy number of copia sequences in A. thaliana would be in correspondence with the small size of its genome.

We have analyzed 71 copia-like sequences, and most of them appeared to be highly heterogeneous in size and potentially inactive. Only 25 elements showed many of the structural characteristics necessary for potentially activity, with eight of these corresponding to families described in previous works. The presence of target site duplication in these 25 elements, along with the LTRs being identical or nearly identical for each element, indicates a predictable recent activity. In this respect, the high sequence identity between the 5’ and 3’ LTRs of retrotransposons from different families of maize (SanMiguel et al. 1998) led to the proposal of a recent burst of activity of the elements in the maize genome. Although most of the analyzed sequences indicate the presence of inactive elements that could be considered relicts of ancient active ones, it is interesting to note the presence of potentially active copia-like elements in the A. thaliana genome. Supporting the potential activity of these elements, three of them have associated ESTs, which implies transcriptional activity for them. None of the analyzed elements are located on the heterochromatin, indicating that they have inserted in coding regions, although it would be necessary to perform a more exhaustive analysis of the genomic sequences adjacent to the insertion points to find how many genes, if any, have been interrupted by the insertions.

The most noticeable characteristic of the 23 copia-like analyzed elements is their heterogeneity. All of these elements vary in sequence and size, although they maintain similar copia-like structures. Divergence values ranging from 5% to 40% are in accord with the extreme variability previously described for copia-like elements in plants (Flavell et al. 1992, 1997; Flavell, Smith, and Kumar 1992; VanderWiel, Voytas, and Wendel 1993; Kumar 1996; Matsuoka and Tsunewaki 1996, 1999; Kumar et al. 1997; Pearce et al. 1997; Wang et al. 1997; Kuipers, Heslop-Harrison, and Jacobsen 1998; Yañez et al. 1998; Kumar and Bennetzen 1999). Some of these works find that the nucleotide divergence among copia-like elements ranges from 0.4% to 57.8% in several grass species (Matsuoka and Tsunewaki 1999) and from 17% to 61% in rye (Pearce et al. 1997), and that the divergence at the amino acid level varies from 1% to 64% in rice (Wang et al. 1997) and from 33% to 58% in Lycopersicon chilense (Yañez et al. 1998). Nevertheless, a relatively high homogeneity has been found among members of the copia-like SIRE-1 family in the soybean genome (Laten, Majumdar, and Gaucher 1998). In A. thaliana, high divergence values between members of the Ta1-Ta10 families was reported early (Voytas and Ausubel 1988; Voytas et al. 1990; Konieczny et al. 1991). This high divergence is confirmed by the fact that all of the elements we analyzed in our work are very different ones. It could be proposed that all of the elements belong to different families, and therefore in the A. thaliana genome there would exist at least 23 families of different copia-like elements. The phylogenetic analysis indicated strong phylogenetic relationships among them, and six major clades were defined, supported by high bootstrap values. These results strongly agree with recent data from Le et al. (2000), which mention the presence of 27 groups of copia-like elements in the A. thaliana genome. In contrast, Kapitonov and Jurka
(1999) reported the presence of 100 copia-like diverse families in the genome of this species. These strong differences in the estimation of family number are probably caused by the criteria used to establish families of copia-like elements.

Ten families, Ta1–Ta10, were described early in the A. thaliana genome, and the authors considered that elements sharing >95% nucleotide identity could be considered members of the same family and that those sharing <85% identity could be regarded as members of different families (Voytas 1992). Assuming the same criterion, with the exception of AtRE-1 and AtRE-2, the rest of the elements could be considered members of separate families. In a similar way, we propose that there are at least 23 families of copia-like families in the A. thaliana genome. These 23 families are grouped into six major lineages which display high divergence values between them. Interestingly, divergences between members belonging to different lineages are higher than diversities with respect to copia-like elements from other plant species. For example, the copia I lineage includes the Hopscotch element from maize, copia V includes SIRE-I from soybean, and copia VI includes Trt1 from Nicotiana tabacum. It is remarkable that the copia I lineage presents elements from such distant species as the monocotyledonous Z. mays and the dicotyledonous A. thaliana. Our present data suggest the existence of these 23 copia-like families among plant genomes; these families would be grouped into a few main lineages. Four superfamilies, named families G1–G4 by Matsuoka and Tsunewaki (1999), have been described extended to grass species. We think that the lineages copia I–VI described here are related to the G1–G4 superfamilies described in the mentioned work. In fact, copia VI and G2 share some elements analyzed in both works. The presence of sharply diverged lineages between the copia-like fraction of several plant species has been established, for example, in Vicia, Solanum, Gossypium, and Lycopersycyon (VanderWiel, Voytas, and Wendel 1993; Kumar et al. 1997; Yañez et al. 1998). However, to establish the phylogenetic relationships between these interspecific lineages of copia-like elements, a more extensive analysis must be done.

As we discussed above, most of the analyzed members of the corresponding families are potentially active. They tend to maintain the structural characteristics necessary for retrotransposition. The existence in a given superfamily of active elements present in the genome of phylogenetically very distant species suggests the presence of active copies of the ancient element in the ancestral species. Therefore, divergent changes operating in the elements of the different species would have been subjected to functional constraints. Finally, most of the 71 analyzed sequences were defective and probably located in the heterochromatin region. We estimate that approximately 80% of copia-like elements in the A. thaliana genome are defective, which is consistent with the copia-like fraction estimated for other plants, such as rye, with 96% defective elements (Pearce et al. 1997). The presence of such a large fraction of defective elements could constitute a good strategy for the plants to avoid the deleterious effects of the huge number of transposable elements that inhabit their genomes.

Acknowledgments

This work was supported by grants BIO4-CT98-0549 (from EC) and BIO099-1320 (from DGESIC, Ministerio de Educación y Cultura) to M.P.-A. J.T. was supported by the program Reincorporación de Doctores a Grupos de Investigación of the Ministerio de Educación y Cultura.

LITERATURE CITED


Feschotte, C., and C. Mouches. 2000. Evidence that a family of miniature inverted-repeat transposable elements (MITEs)


Heslop-Harrison, J. S., A. Brandes, S. Taketa et al. (15 co-authors). 1997. The chromosomal distributions of Ty1-copia group retrotransposable elements in higher plants and their implications for genome evolution. Genetica 100:197–204.


Nei, M., and R. Chakraborty. 1976. Empirical relationship between the number of nucleotide substitutions and inter-


Pearce, S. R., G. Harrison, D. Li, J. Hesslow-Harrison, A. Kumar, and A. J. Flavell. 1996a. The Ty1-copia group retrotransposons of Allium cepa are distributed throughout the chromosomes but are enriched in the terminal heterochromatin. Chromosome Res. 4:357–364.


Pierre Capy, reviewing editor

Accepted January 23, 2001