Structure and Evolution of the hAT Transposon Superfamily

Eitan Rubin,[1] Gila Lithwick and Avraham A. Levy

Department of Plant Sciences, The Weizmann Institute of Science, Rehovot 76100, Israel

Manuscript received October 20, 2000
Accepted for publication April 9, 2001

ABSTRACT

The maize transposon *Activator* (*Ac*) was the first mobile DNA element to be discovered. Since then, other elements were found that share similarity to *Ac*, suggesting that it belongs to a transposon superfamily named hAT after *hobo* from Drosophila, *Ac* from maize, and Tam3 from snapdragon. We addressed the structure and evolution of hAT elements by developing new tools for transposon mining and searching the public sequence databases for the hallmarks of hAT elements, namely the transposase and short terminal inverted repeats (TIRs) flanked by 8-bp host duplications. We found 147 hAT-related sequences in plants, animals, and fungi. Six conserved blocks could be identified in the transposase of most hAT elements. A total of 41 hAT sequences were flanked by TIRs and 8-bp host duplications and, out of these, 34 sequences had TIRs similar to the consensus determined in this work, suggesting that they are active or recently active transposons. Phylogenetic analysis and clustering of hAT sequences suggest that the hAT superfamily is very ancient, probably predating the plant-fungi-animal separation, and that, unlike previously proposed, there is no evidence that horizontal gene transfer was involved in the evolution of hAT elements.

TRANSPOSABLE elements (TEs) can be divided into two major groups according to their transposition mechanism (Finnegan 1990): (1) retroelements, which transpose via reverse transcription of an RNA intermediate (Xiong and Eickbush 1990), and (2) "DNA-DNA" TEs, which transpose directly via DNA, are flanked by terminal inverted repeats (TIRs), and encode one or several proteins including the transposase that is required for transposition (for reviews, see Saedler 1996). DNA-DNA elements are ubiquitous in the genomes of prokaryotes and eukaryotes. Nevertheless, their origin remains unknown. All DNA-DNA transposons and some retroelements mediate the formation of short host duplications upon insertion.

Most DNA transposons are organized in families of autonomous and nonautonomous elements, characterized by their ability to respond to the same transposase. Transposons of the same family usually share extensive nucleotide similarity at their termini. Superfamilies can also be identified by analysis of the amino acid sequence of the transposase genes, both in eukaryote and prokaryote transposons (*e.g.*, the *Tn3* superfamily; Sherratt 1989). The *Tc1/Mariner*-like superfamily has members in several phyla, including bacteria, vertebrates, invertebrates, and plants (Henikoff 1992). Another example of a superfamily includes elements *Tam1* from *Antirrhinum majus* (snapdragon) and *Tgm* from soybean. These elements were found to be structurally related to the maize *En/Spm* TE (Nacken *et al.* 1991) on the basis of similarity of their transposase as well as their TIRs, which contain the so-called CACTA motif. TIRs and subterminal regions are not strongly conserved among elements of the same superfamily.

The maize element *Activator* (*Ac*), the *Drosophila melanogaster* element *hobo*, and the *A. majus Tam3* elements also form a superfamily of eukaryotic TEs (Calvi *et al.* 1991; Feldmar and Kunze 1991) referred to as the hAT (hobo-*Ac*-Tam3) superfamily. Additional elements, such as Hermes from Drosophila (Warren *et al.* 1994) and Hector (Warren *et al.* 1995), were also shown to belong to the hAT superfamily. The most conserved feature of the transposase of hAT elements is a domain of ~50 amino acids located at the C terminus. This domain was shown recently to be involved in dimerization as well as in additional interaction functions (Essers *et al.* 2000). The TIRs of hAT elements are usually short and their sequence is ill defined. Another conserved feature of active hAT elements (both autonomous and nonautonomous) is that they mediate the formation of 8-bp host duplication (HD) upon insertion. It was proposed that horizontal gene transfer (HGT) could explain the presence of hAT members in such distantly related species as plants and flies (Calvi *et al.* 1991). This proposal, however, should be reexamined given the discovery of hAT elements also in fungi (Kempken *et al.* 1998).

We developed new tools for mining sequence databases for the presence of DNA-DNA transposon-like structures (HD-TIR-transposase-TIR-HD). Using these tools, we carried out a survey of hAT members. New

Corresponding author: Avraham A. Levy, Plant Sciences, Weizmann Institute of Science, Rehovot 76100, Israel.
E-mail: avi.levy@weizmann.ac.il

[1]*Present address:* Compugen Ltd., Pinchas Rozen 72, Tel Aviv 69512, Israel.

hAT sequences were identified, some being candidates for active TEs and most being probably fossil TEs. We found that hAT elements are characterized by six conserved blocks of amino acids and by a weak consensus for TIRs. Our results suggest that the hAT superfamily is very widespread and is probably very ancient, predating the separation of plants, animals, and fungi. Sequence analysis, through clustering and tree-based phylogenetic analysis, showed no evidence for transkingdom horizontal gene transfer.

MATERIALS AND METHODS

A semiautomated system for TE identification and annotation: A system integrating various programs and computer-assisted user annotation was developed for homology-based TE identification and annotation (Figure 1). Information needed by various programs (*e.g.*, BLAST), as well as the parsed results from the programs, was stored within a single Sybase database. Decisions and annotations made by the user were also stored. Iterative searching of various databases (see below) was performed using BLAST (Altschul *et al.* 1997). Each new sequence was reviewed manually and was either accepted or rejected (the rejection criteria are explained below). For a complete list of the rejected sequences, see http://bioinfo.weizmann.ac.il/~lithwick/hAT/Rejected.html. Accepted sequences were then subjected to further searching until no new hits were obtained. After searching was completed, each sequence was subjected to further analysis. A set of tools was used to reduce redundancy by aliasing identical and nearly identical entries (see below). For functional annotation, HD-TIR searching was performed using the Transpolator (see below), and profile searching was performed using the BLOCKS package (see below). Programs clustering the sequences coupled with queries into the annotation of the entries were used to search for phylogenetic inconsistency (see below). In addition, programs that build various tables for the presentation of the results were used.

Semiautomated iterative database searching: Iterative database searching was performed on the National Center for Biotechnology Information databases nt, nr, htg, est_human, and est_others (from Oct 9, 1999; est_others from Oct 3, 1999), beginning with several known members of the hAT superfamily. The blastall program (version 2.0.9) was run (http://genome.nhgri.nih.gov/blastall) using all five available algorithms. A borderline significance threshold of 0.003 was used. Parsing was performed using the bioperl (http://www.bioperl.org) BLAST parser. Since only sequences similar to the transposase were desired, sequences found by similarity to a DNA sequence and not to a protein sequence were rejected. The results were examined manually after each iteration. Clustering of the search results facilitated the decision-making process; sequences similar to the same region of a sequence were clustered together, allowing the manual acceptance or rejection of a group of hits with one decision. Consequently, sequences found by similarity to areas flanking the transposase-like region were clustered together and then rejected together. For each search, the clustering was done transitively on the basis of the coordinates of the query HSP (high-scoring segment pairs, the areas of local alignment produced by BLAST), requiring an overlap of at least 60%.

For the next round of database searching, the sequences used were new protein sequences and DNA sequences of length <10,000 bp. DNA sequences were masked for low complexity regions using RepeatMasker (see http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker). Searching iterations were continued until no new hits were found. Protein segments created by homology-based translation (see below) were run once against the nr protein database using a cutoff of 1*e*-4. The results were examined but were not used for another round of searching because of the inaccuracy of the translations. As an exception, the Morning Glory Tip100 protein sequence (accession no. BAA36225), which was found only by a translated DNA segment, had its DNA sequence (accession no. AB004906) added manually. Expressed sequence tag (EST) hits were not used for additional searches.

Data storage and retrieval: Sequence information as well as search statistics and all subsequently gathered information were stored in the Sybase database. Interfacing with the database was done with the Perl DBI module (http://www.perl.com/CPAN-local/modules/by-module/DBI). Some of the frequently used queries were stored (available upon request), while others were written on a single usage basis.

Sequence aliasing: To reduce redundancy within the database, sequences with a high degree of similarity to another sequence were aliased so that only one sequence was regarded in further analyses. To this end, DNA segments were created, using HSP segments found by a protein-to-DNA comparison with an *e*-value ≤0.01. These segments were then extended by 5000 bp in each direction. For htg sequences, segments were not extended. All DNA segments were compared using blastn. Pairs of sequences with a score >90% (100% for htg segments) of the score obtained by self alignment were considered identical, and one was aliased to the other. In the same method, protein sequences for which the normalized score against each other was 100% of their normalized self-score were aliased. DNA, RNA, and protein triplets or pairs were united on the basis of a high degree of similarity as detected by the appropriate blast algorithm or on the basis of annotation. Some sequences were aliased manually.

Homology-based translation: Standard gene prediction programs were not suitable for our analysis, since they are optimized to find complete genes, and our results involve gene fragments. Furthermore, few programs analyze plant sequences successfully. Homology-based translation was therefore used. For each DNA sequence, segments that were found by protein sequences with an *e*-value ≤0.01 were extended in both directions to the flanking stop codons. Open reading frames (ORFs) occurring within the same frame were merged into a single sequence, allowing skipping of unconserved or untranslated regions (*e.g.*, introns). This method of translation is biased toward sensitivity and with a high degree of confidence includes all of the translated regions at the expense of the inclusion of untranslated regions.

Prediction of functionality. BLOCK analysis and Transpolator: *Identification of conserved blocks:* Blocks were determined on the basis of three programs. Members from each of the six clusters (see below) were chosen. Active members were taken when possible. These were aligned using DIALIGN2 (Morgenstern *et al.* 1998) and BLOCKMAKER (Henikoff *et al.* 1995). In addition, all translated DNA sequences (see above) with an ORF >400 amino acids were used for block searching using MEME (http://meme.sdsc.edu/meme/website/meme.html). Blocks were chosen where at least two of these programs gave the same results and then were adjusted manually. BLIMPS (Wallace and Henikoff 1992) was used to calibrate and run the blocks against the entire sequence collection as defined by the iterative search scheme described above.

Identification of DNA features: In active elements, the transposase must be flanked by TIRs and HDs. To detect such HD-TIR-transposase-TIR-HD structures, DNA segments were created from segments found by proteins with an *e*-value ≤0.01 and

extended by 5000 bp in each direction. Within these segments, flanking the putative transposase area, TIRs of minimum length 8 bp flanked by 8-bp HDs were searched for. For this purpose, the Transpolator computer program was developed (available upon request). TIRs were permitted to have a single imperfect nucleotide at the first base. To avoid simple repeats, TIRs composed of only two nucleotides were rejected. Since some known TEs are present in the database only from the first TIR to the second TIR (*i.e.*, without HD), sequences annotated as having an 8-bp HD were added manually, where a single imperfect nucleotide was permitted in any position.

Tree-based phylogenetic analysis: The core region of the conserved domains described above (Essers *et al.* 2000) was chosen for phylogenetic tree inference. A maximum-likelihood phylogenetic inference package, PROTML (the package and its documentation are available at ftp://sunmh.ism.ac.jp/pub/molphy), was used with the star decomposition heuristics. For visualization, the program TreeView was used, and the results were hand edited to improve readability without changing the tree topology or branch length.

Clustering of protein sequences: We used a technique very similar to the one used by Lander *et al.* (2001) to search the human genome for evidence of trans-kingdom horizontal gene transfer events. In our approach, sequences, including DNA translations (see above), were transitively clustered, on the basis of *e*-values from the protein-to-DNA BLAST search, using a cutoff of 1*e*-20. This *e*-value cutoff was chosen since higher (*i.e.*, more permissive) cutoffs caused the collapse of the clusters into one, moderately lower cutoffs did not change the results of the clustering, and much higher cutoffs were too strict, giving rise to no clusters. Hence, a sequence was put into a cluster if and only if it found, or was found by, another sequence within the cluster, at a significance of at least 1*e*-20 as reported by the BLAST program. Clusters composed of only a single sequence were disregarded.

The best-characterized TE of a cluster was chosen as the cluster representative and was used for naming the cluster.To visualize the clustering (Figure 2), sequences were plotted against each other in a graph, and points were marked where the sequence on the *x*-axis found the sequence on the *y*-axis. For the value of the point, the best *e*-value obtained between the two sequences in any of the BLAST programs was chosen. To keep similar sequences adjacent to each other, ordering within each cluster was achieved by aligning the sequences with CLUSTALW and adopting the resulting order. The image was created using the Perl (http://www.perl.org) version of the GD module (http://stein.cshl.org/WWW/software/GD).

Availability: All programs and the contents of the database are available upon request. More information can also be found at http://bioinfo.weizmann.ac.il/~lithwick/hAT/.

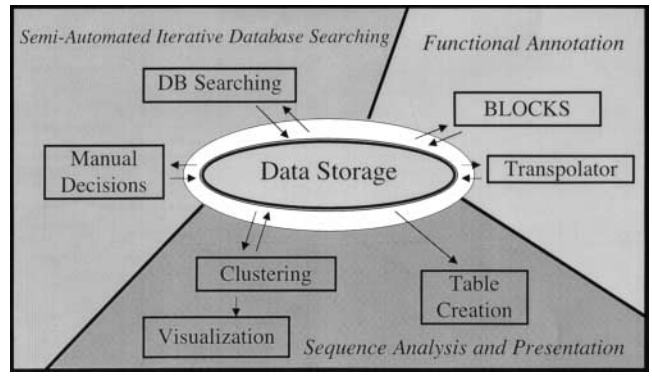

Figure 1.—Scheme used for identification of hAT elements. Information needed for various programs, the parsed program output, and user decisions and annotations were stored into a single database (Data Storage). Iterative searches for sequences similar to the hAT transposase were performed using BLAST, results were reviewed, and sequences were accepted or rejected manually. Each sequence was then functionally annotated by searching for features of active transposons using BLOCKS and Transpolator. Accepted sequences were subjected to further analysis and tables were created.

RESULTS

Abundance of hAT-like sequences in sequence databases: We identified and analyzed 147 nonredundant hAT-related sequences in public databases using the integrative search scheme shown in Figure 1. The search scheme (see details in materials and methods) was developed to allow iterative search with many search tools and in many databases and to integrate all search results with maximal flexibility. Emphasis was put on integrating human reasoning in the search scheme, providing machine support rather than automatic decisions wherever possible. In this work, we used BLAST, BLIMPS, and Transpolator for searching the protein and DNA nonredundant databases (see materials and methods for more details). At the time of the analysis, there were 487,986 nucleotide sequences and 416,691 protein sequences (not including ESTs) in the public databases we searched, totaling $\sim 2 \times 10^9$ bp and 1.3×10^8 amino acids, respectively. The search gave 4111 hits, out of which 258 were accepted as containing similarity to hAT transposases. In addition, 127 hAT-like ESTs were accepted but were not analyzed further. A total of 866 sequences were automatically rejected because they were found only by DNA sequences, and 2749 were automatically rejected because they were found only by irrelevant sequences (*e.g.*, the flanking sequences of TEs). A total of 111 were rejected manually, mainly because they were similar to repetitive areas or were judged as irrelevant, *e.g.*, homology to known genes such as *Starch synthase*. A list of the rejected entries is given in the following site: http://bioinfo.weizmann.ac.il/~lithwick/hAT/Rejected.html. Out of the 258 positive hits, 147 were nonredundant and were used for further analyses. One striking result from the search is the overrepresentation of hAT-related sequences in the *Arabidopsis thaliana* genome. We identified hits in 29 species, including human, fish, nematodes, flies, fungi, and plants. Of these species, only 3 were model species, with advanced genome projects, namely human, *Caenorhabditis elegans*, and *A. thaliana*, with 10, ~100, and ~50% of their genomes available at the time of the analysis. In *A. thaliana*, 75 hits were found, suggesting that there are ~150 hAT-related sequences in the complete genome. In *C. elegans*, only 12 hits were found. In human, 7 hits were found, suggesting that the complete genome contains ~70 hAT members. No hAT-related sequences were found in the fungus *Saccharomyces cerevisiae*, for

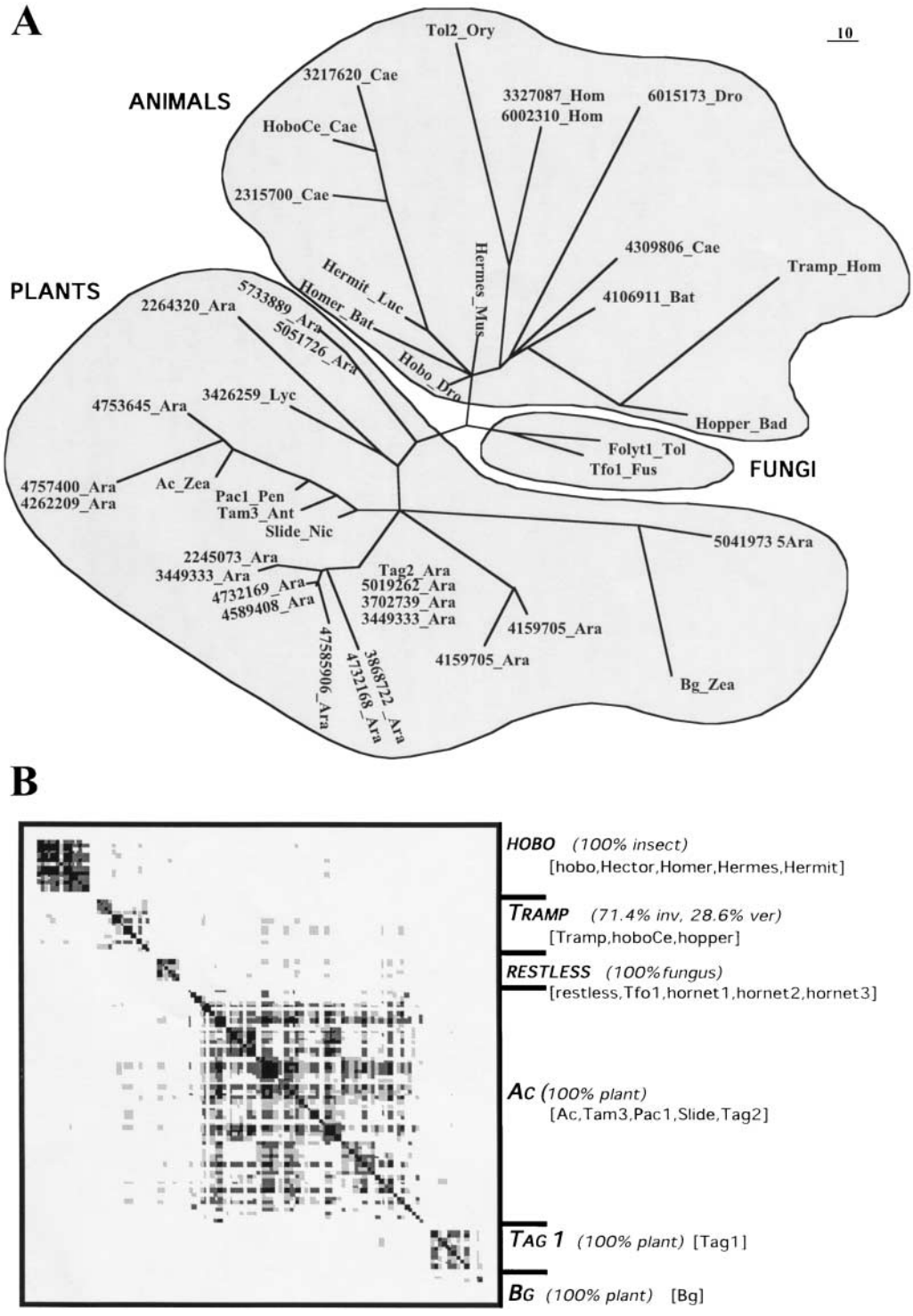

FIGURE 2.—Tree-based phylogeny and sequence clustering of hAT elements. (A) Tree-based phylogeny of the hAT superfamily. The most conserved block in the assembly of sequence collected through iterative database searching (see block E, Figure 4) was used to infer a phylogenetic tree using the maximum-likelihood approach (see MATERIALS AND METHODS). Branch lengths represent the estimated number of substitutions per 100 amino acid sites along the branch (note the scale at the top right corner of A). For each sequence, its identifier (either a name or its global identifier number as found in GenBank) and an abbreviated species name are given (Ant, *Antirrhinum majus*; Ara, *A. thaliana*; Bad, *Bactrocera dorsalis*; Bat, *B. tryoni*; Cae, *C. elegans*; Dro, *D. melanogaster*; Fus, *Fusarium oxysporum*; Hom, *H. sapiens*; Luc, *Lucilia cuprina*; Mus, *Musca domestica*; Nic, *Nicotiana tabacum*; Ory, *Oryzias latipes*; Pen, *Pennisetum glaucum*; Tol, *Tolypocladium inflatum*; Zea, *Z. mays*). Sequences that belong to species from the same kingdom are circled and labeled. For precise branch length and error estimates see http://bioinfo.weizmann.ac.il/~lithwick/hAT/tree.html. (B) Clustering of BLAST results. The degree of similarity of each sequence to all other sequences, as determined by the *e*-value provided by BLAST, is indicated by the point intensity, ranging from 1*e*-10 (light gray) to <1*e*-50 (black). The species distribution of the sequences in each cluster is shown, within brackets, next to the cluster name. Members of each cluster that were previously considered as transposons are mentioned within brackets.

which the complete genome was available at the time of the analysis. The hits we found in *A. thaliana* did not show preference to specific chromosomes. Hits were found on all five chromosomes, and the fraction found on each chromosome was highly similar to that expected on the basis of chromosome length (data not shown).

TABLE 1

Distribution of hAT-related sequences into clusters and Transpolator results

Cluster[a]	Total[b]	NR[c]	TIRs + 8-bp HD[d]
Ac	98	63	17
Bg	3	2	2
hobo	41	14	5
restless	10	6	3
Tag1	15	10	1
Tramp	28	14	5
Unclustered	47	26	6
Low similarity	16	12	2

[a] Clusters are named after a representative transposon, except for accessions that could not be associated with a particular cluster (unclustered) or that had BLAST *e*-values between 0.003 and 1*e*-10 (low similarity).

[b] Total no. of accessions.

[c] No. of nonredundant (NR) accessions out of total.

[d] No. of sequences flanked by TIRs + 8-bp HDs out of NR sequences and found by the Transpolator program.

hAT-like sequences phylogeny and clustering into six groups of homogenous phylogenetic origin: We used the PROTML program to determine the phylogenetic relationships among the various hAT sequences (Figure 2A). A subset of entries that were found to contain a conserved segment of 26 amino acids was selected for the analysis (see BLOCKs analysis below). In total, 43 entries were used to deduce the tree shown in Figure 2A. In this tree, three major branches could be observed, each of which contains sequences from only one kingdom, either from plants, animals, or fungi. Within each kingdom, subgroups were sometimes observed. For example, there was a clear group of hobo-like sequences, all of which were derived from invertebrates. Other branches from the animal kingdom contained a mixture of vertebrate and invertebrate sequences. Among plants, there was a group of Tag2-like sequences that was exclusively from Arabidopsis. Other groups had a mixture of monocot and dicot sequences. We used clustering of blast results to further study the relationship between the whole set of the hAT sequences. This analysis enables us to analyze sequences that could not be included in the tree-based phylogeny due to the large number and diversity of the sequences and because the conserved region was not always large enough. With this method, sequences were grouped on the basis of similarity to each other using a greedy (*i.e.*, transitive) clustering of BLAST results (Lander *et al.* 2001) and with visualization tools using a method originally developed for gene expression analysis (Ben-Dor *et al.* 1999). Clustering of the nonredundant sequences resulted in six groups (Figure 2B) that include 109 nonredundant sequences out of the 147 sequences analyzed. The 38 additional sequences were unclustered (Table 1). The largest group was the *Ac*-like cluster with 63 sequences. This group contained only plant sequences. Similarly, the *Tag1* and *Bg*-like groups contained only plant sequences. The *hobo*-like cluster contained only insect sequences and the *Tramp* cluster contained only animal accessions with sequences from both vertebrates and invertebrates. The *restless* cluster contained only fungal sequences. These results show a strong correlation between the phylogenetic origin of the sequences and their similarity.

Functional analysis of hAT sequences: The function of the hits we obtained was analyzed in three ways: (1) Annotation was used when available to try to understand the function of the protein hit; (2) the Transpolator program was used to find the HD-TIR structure that is expected to flank active transposons; and (3) the presence of conserved domains was used to examine the conservation of the protein.

Of the 147 nonredundant hits we obtained, and of the redundant set of 258 hits found in the databases, all 23 entries that have a known function are from transposable elements. In some cases, the annotation suggested a function unrelated to transposition for the sequence, but the hAT-related fragment was always found to be positioned in an intron or outside the coding region.

A second functional analysis approach was to search for a "HD-TIR-protein hit-TIR-HD" structure, which strongly suggests involvement in transposition. We used the Transpolator program for all hits where genomic DNA (and not mRNA) sequences were available. A segment of 10,000 bp flanking the protein similarity area was used to search for HD-TIR structures, limiting the length of the HD to 8 bp and rejecting all hits likely to have occurred by chance (see materials and methods).

In 41 of the 147 hAT sequences, TIRs and 8-bp HD were found. Alignment of the hAT-flanking TIRs is shown in Figure 3. The consensus (T/C)A(A/G)NG was found at the extremity of the elements. In the *Ac* cluster, there was a preference for TA(A/G)NGNTG, although slight variations were observed. In the hobo group there was a preference for CAGAGA and in restless the consensus was CAGNG. A total of 12 accessions that were not previously described as transposons had TIRs similar to those of well-characterized elements, suggesting that they are active TEs or that they were recently active.

The third functional analysis was to compare the transposase-like regions to identify conserved blocks. Six blocks were found (A–F in Figure 4) of length varying from 10 to 26 amino acids. The relative order of the blocks along the transposase is conserved in all hAT elements. Three blocks (D–F) are clustered in the C terminus of the protein. These blocks contain the dimerization domain of Ac-transposase (Essers *et al.* 2000). They are the most abundant of the six blocks (Table

NAME	GI	ORGANISM	CLUSTER	
Ac	168402	*Z.mays*	Ac	TAGGGATGAAA
Tam3	16063	*A.majus*	Ac	TAAAGATGTGAA
Slide	1617412	*N.tabacum*	Ac	TAATGCTG
*	3868722	*A.thaliana*	Ac	CAGAAAAA
	4262209	*A.thaliana*	Ac	TAGCCCTG
*	4314354	*A.thaliana*	Ac	TAGGGGTGTCAAAA
*	4589408	*A.thaliana*	Ac	GAAACATGA
*	4732169	*A.thaliana*	Ac	GAAACATGA
	4757400	*A.thaliana*	Ac	TGAAGATGC
*	3449333	*A.thaliana*	Ac	TAGGGATGTT
*	3449333	*A.thaliana*	Ac	TAGAAGTGTCAA
	4585906	*A.thaliana*	Ac	TAGGGGTGTCAA
	4732168	*A.thaliana*	Ac	TAGGGGTGTCAAAA
*	4996903	*A.thaliana*	Ac	AAGTTATA
Bg	22493	*Z.mays*	Bg	CAGGG
*	5852170	*O.sativa*	Bg	CAGGGTTCAC
hobo	157606	*D.melanogaster*	hobo	CAGAGAACTGCA
Hermes	514387	*M.domestica*	hobo	CAGAGAAC
Hermit	726316	*L.cuprina*	hobo	CAGAGATGTGCATG
Homer	4106909	*B.tryoni*	hobo	CAGAGATCTGCA
*	4006805	*D.melanogaster*	hobo	GAAAAATA
restless	1542944	*T.inflatum*	restless	CAGAGTGCGTAATC
Tfo1	3410895	*F.oxysporum*	restless	CAGTGTGTCCATCA
	1523779	*A.immersus*	restless	CAGTGGCTCCAACC
Tag1	2935590	*A.thaliana*	Tag1	CAATGTTTTCACGC
hoboCe	733580	*C.elegans*	Tramp	CAGGGGTGTGCGGC
*	1212868	*C.elegans*	Tramp	TAAAATGT
Tol2	1552184	*O.latipes*	Unclustered	CAGAGGTGTAAA
Folyt1	3126915	*F.oxysporum*	Unclustered	TAGAGATGG
	2088712	*C.elegans*	Unclustered	CAGCAATCC
	2088712	*C.elegans*	Unclustered	TGATGGTAA
*	2315700	*C.elegans*	Unclustered	CAGACTTGTGCGGC
Tip100	4063769	*I.purpurea*	Low Similarity	CAGGGGCGGAG
*	2264320	*A.thaliana*	Low Similarity	CAAATATATA

FIGURE 3.—Alignment of terminal inverted repeats (TIRs) flanking hAT-related sequences. TIR sequences identified by Transpolator are shown and are indicated by their GenBank identifier number (GI). A TIR consensus was built on the basis of TIRs of known transposons whose names are indicated. TIRs sharing at least three nucleotides with the consensus were included. Sequences that were not previously annotated in the public databases as similar to hAT elements are marked (*). Bases conserved among all the clusters are highlighted in dark gray, and those that are specific to clusters are highlighted in light gray. The BAC 3449333 has two consecutive hAT-like elements, each flanked by TIRs and 8-bp HDs; both pairs of TIRs are shown. The BAC 2088-712 has two nested sets of TIRs and 8-bp HDs; both pairs of TIRs are shown.

2): Block E is present in all the clusters, block D is present in all clusters except Bg, and block F is present in all clusters except Bg and Tag1-like elements. The region spanned by these blocks was predicted to contain helical structures, which is well in agreement with the suggested role for this region in protein dimerization (KUNZE *et al.* 1993). The other blocks (A–C) are smaller and more dispersed throughout the transposase than the D–F blocks (Figure 4). Blocks A and B are present in all the clusters, and block C is present in all the clusters except in Bg and Tag1-like elements. The function of these blocks is not known.

DISCUSSION

hAT is a diverse and ancient transposon superfamily: We carried out an exhaustive survey of hAT-related sequences in the public sequence databases to estimate the diversity, abundance, and evolution of the hAT superfamily. We found 147 nonredundant sequences (as of October 1999). Some sequences were found upstream or downstream of known genes, others in the introns of known genes, and most were found in genomic fragments from genome projects, mainly bacterial artificial chromosomes (BACs; data not shown). Some of the sequence hits found by our search protocol were previously reported as TE sequences (CALVI *et al.* 1991; BIGOT *et al.* 1996; COATES *et al.* 1996; FRANK *et al.* 1997; COLOT *et al.* 1998; OKUDA *et al.* 1998). For some genomic sequences, automatic annotations were available, suggesting that they are hAT related, *e.g.*, "similar to putative *Ac*-like transposable elements." For the majority of genomic sequences found here (known transposons not included), no annotation was available for the hAT-related regions (see examples in Figure 3). All the sequences with known function corresponded to TEs.

The hAT superfamily was shown to be widely distributed in all eukaryotic kingdoms. hAT sequences could be clustered into subgroups (Figure 2), each of which contained sequences from only one kingdom. This result was obtained using two independent grouping methods: a tree-based phylogeny (Figure 2A) and clustering of BLAST results. Both methods led to the same conclusion, namely that there is no evidence for trans-

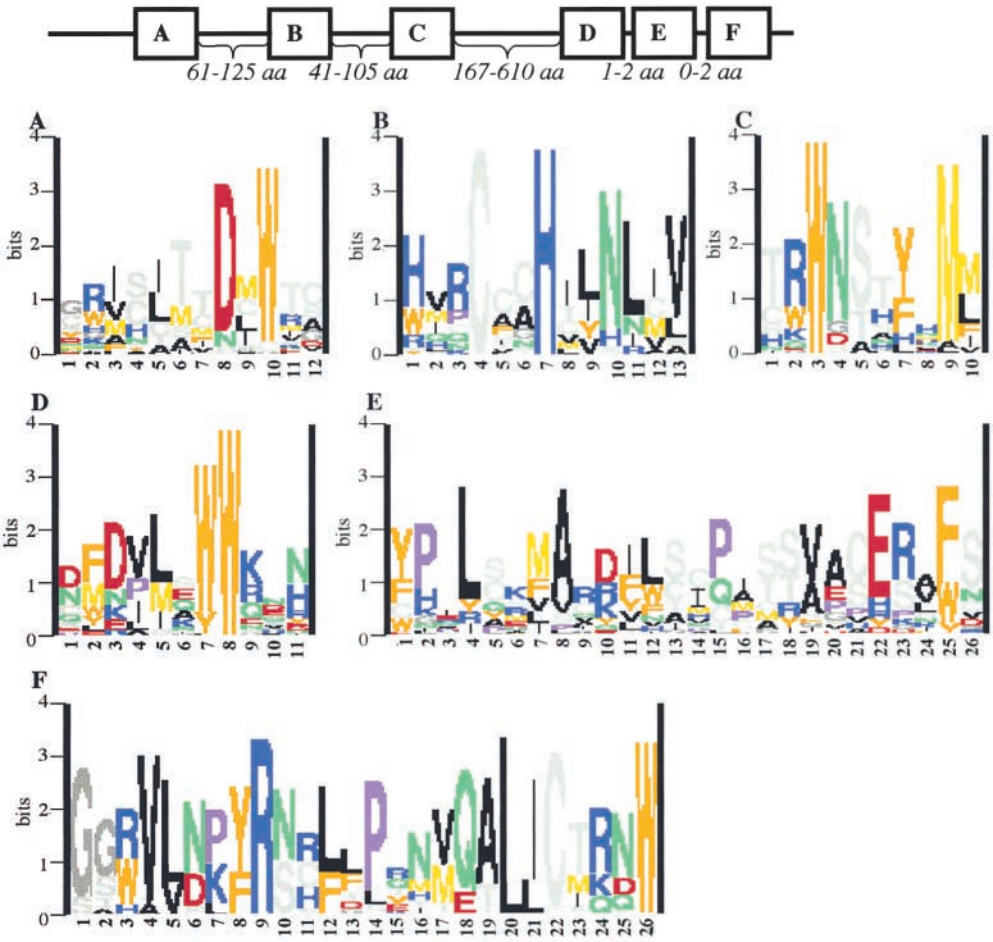

FIGURE 4.—Conserved hAT protein blocks. The six conserved blocks (A–F) are located at the C terminus of the protein. Their linear order and the distances between them are schematically represented by boxes (not to scale) at the top. The blocks are represented below as sequence logos (SCHNEIDER and STEPHENS 1990). A detailed description of the multiple sequence alignment visualization method is given in HENIKOFF *et al.* (1995). Briefly, the multiple sequence alignment is represented by ordered stacks of letters, with the height of each letter in a stack depending on its frequency and the total height of a stack depending on the overall conservation of a specific position. Amino acid coloring is based on physiochemical properties: red for acidic (D and E); blue for basic (H, K, and R); light gray for polar (C, S, and T); green for amide (N and Q); yellow for methionine (M); black for hydrophobic (A, I, L, and V); orange for aromatic (F, W, and Y); purple for proline (P); and gray for glycine (G). Please note that for some of the sequences used in the alignments, there is no experimental proof that they encode for a functional transposase; therefore, additional analysis should be done before using this figure for making functional predictions.

kingdom horizontal gene transfer as was previously proposed (CALVI *et al.* 1991). Another interesting question is whether intrakingdom horizontal transfer of hAT transposons is involved in the evolution of these elements as previously suggested for some hAT members (SIMMONS 1992; KOGA *et al.* 2000). The methodology used here did not enable us to confirm or challenge this possibility because of the small data set that was used and the relatively high error values. Such a question requires more refined tools. Our results also do not completely rule out models for hAT evolution that involve trans-kingdom horizontal gene transfer, such as promiscuity between kingdoms during early eukaryotic evolution. However, these alternative models require additional assumptions, making the "early origin" model the most favorable. In summary, these data suggest that the hAT elements are from an ancient and diverse family that radiated into modern species, mostly via vertical inheritance as was reported for the evolution of the *Tc1/mariner* (KIDWELL 1993). Interestingly, while Mariner is particularly abundant in animals and particularly rare in plants, the opposite is observed here for hAT elements.

TABLE 2

Percentage of sequences containing each hAT-transposase block

	hAT-transposase blocks (%)					
Cluster	A	B	C	D	E	F
Ac	56	44	57	70	75	76
Bg	100	50	0	0	100	0
hobo	57	36	43	57	57	14
restless	67	100	67	33	50	17
Tag1	30	70	0	20	50	0
Tramp	29	36	36	7	71	14
Unclustered	12	4	19	12	50	19
Low similarity	0	0	17	8	42	8

The blocks (A–F) are as described in Figure 4. The number of nonredundant accessions per cluster is shown in Table 1. Block A is contained within Calvi's region 1 (CALVI *et al.* 1991). Blocks B and C correspond to the edges of Calvi's region 2. Blocks D–F are the highly conserved C-terminal region involved in dimerization (ESSERS *et al.* 2000).

Structure of the hAT elements: We analyzed the structure of hAT sequences with respect to both the transposase and inverted repeats. Clustering of the transposase sequences revealed six conserved domains (Figure 4) that were slightly different from those previously re-

ported for *Ac*, *hobo*, and *Tam3* (Calvi *et al.* 1991). The current blocks were obtained using a broad range of sequences from all the hAT clusters and using three different programs (see materials and methods). This makes the new blocks smaller, which may better represent the core of the conserved amino acids. For example, previous works have considered blocks D–F as one block, namely the dimerization domain (Essers *et al.* 2000). The three new conserved domains that were obtained by the block program reflect well the fact that domains D–F are not always present as one unit in all the entries (Figure 4 and Table 2); but different elements can share only two of the three blocks or only one block as with the Bg element. This may suggest that more than one function is encoded in this region, with each block participating in another function. Since this region is involved in dimerization (Essers *et al.* 2000), some of the conserved regions may play a role in regulation of transposition. This is in agreement with the finding that transposase aggregation is involved in regulation of maize transposition (Heinlein *et al.* 1994).

The TIRs of the hAT elements were aligned for 23 elements that are known to be active and the consensus we found, (T/C)A(A/G)NG, is an extension of the "NANNG" consensus previously proposed (Warren *et al.* 1995). The same consensus, sometimes with slight variations, was found in new sequences that were previously annotated as undefined (Figure 3), strengthening the fact that these sequences belong to transposons that are active or that were active recently. The observed similarities suggest some primordial TIR nucleotide sequence present in early eukaryotes that has undergone more or less constant selection. Considering that no hAT-related sequence with a function other than transposase was found, the primordial DNA sequence might have already functioned as a transposon prior to the plant-animal-fungi separation. Another interesting observation is that the TIR consensus that we found resembles the recombination signal sequences (CACAGTG and CACTGTG) that are cleaved by the RAG1 and RAG2 proteins during V(D)J recombination. The RAG genes were previously shown to catalyze cut-and-paste transposition *in vitro* (Hiom *et al.* 1998). The sequence similarity at the breakpoint termini might therefore be interpreted by a common phylogenetic origin for hAT transposition and V(D)J recombination.

The successful identification of TIRs in so many genomic sequences supports the ability of the Transpolator program to identify relevant TIRs. Yet a large number of hAT sequences did not contain any TIRs or contained TIRs unrelated to the consensus. These are probably unable to transpose and are probably fossil transposons. Such elements can be rapidly derived from active transposons through abortive gap repair (Rubin and Levy 1997). They may represent relics from the past, with mutations that did not yet change the sequence beyond recognition. Alternatively, fossilized hAT sequences may play the important role of repressing transposition through transposase dilution or through a negative dominant repression by truncated transposases, as suggested for Mariner elements (Hartl *et al.* 1997).

In summary, the hAT superfamily contains diverse members in all the eukaryotic kingdoms. There is no evidence that trans-kingdom horizontal gene transfer was involved in hAT element evolution. hAT elements are characterized by a transposase containing six conserved blocks and short TIRs with a weak consensus. The lack of evidence for trans-kingdom transfer, the conservation at the amino acid level, and the conservation at the DNA level suggest that hAT was a family of mobile elements prior to or at the early stages of the plant-animal-fungi separation. A large number of fossil hAT sequences (that do not contain TIRs) were identified, representing either relics of the past with no function or active transposition repressors.

We thank J. Prilusky and Irit Orr from the bioinformatics unit of the Weizmann Institute of Science for computational help. We also thank Shmuel Pietrokovsky for his help in constructing the blocks. This work was supported by doctoral and master fellowships from the Feinberg Graduate School to E.R. and G.L., respectively.

LITERATURE CITED

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

Ben-Dor, A., R. Shamir and Z. Yakhini, 1999 Clustering gene expression patterns. J. Comput. Biol. **6:** 281–297.

Bigot, Y., C. Auge-Gouillou and G. Periquet, 1996 Computer analyses reveal a hobo-like element in the nematode Caenorhabditis elegans, which presents a conserved transposase domain common with the Tc1-Mariner transposon family. Gene **174:** 265–271.

Calvi, B. R., T. J. Hong, S. D. Findley and W. M. Gelbart, 1991 Evidence for a common evolutionary origin of inverted repeat transposons in Drosophila and plants: hobo, Activator and Tam3. Cell **66:** 465–471.

Coates, C. J., K. N. Johnson, H. D. Perkins, A. J. Howells, D. A. O'Brochta *et al.*, 1996 The *hermit* transposable element of the Australian blowfly, *Lucilia cuprina*, belongs to the *hAT* family of transposable elements. Genetica **97:** 23–31.

Colot, V., V. Haedens and J. L. Rossignol, 1998 Extensive, nonrandom diversity of excision footprints generated by *Ds*-like transposon *ascot-1* suggests new parallels with V(D)J recombination. Mol. Cell. Biol. **18:** 4337–4346.

Essers, L., R. H. Adolphs and R. Kunze, 2000 A highly conserved domain of the maize activator transposase is involved in dimerization. Plant Cell **12:** 211–224.

Feldmar, S., and R. Kunze, 1991 The ORFa protein, the putative transposase of maize transposable element Ac, has a basic DNA binding domain. EMBO J. **10:** 4003–4010.

Finnegan, D. J., 1990 Transposable elements and DNA transposition in eukaryotes. Curr. Opin. Cell Biol. **2:** 471–477.

Frank, M. J., D. Liu, Y. F. Tsay, C. Ustach and N. M. Crawford, 1997 Tag1 is an autonomous transposable element that shows somatic excision in both Arabidopsis and tobacco. Plant Cell **9:** 1745–1756.

Hartl, D. L., A. R. Lohe and E. R. Lozovskaya, 1997 Modern thoughts on an ancient mariner: function, evolution, regulation. Annu. Rev. Genet. **31:** 337–358.

Heinlein, M., T. Brattig and R. Kunze, 1994 *In vivo* aggregation of maize *Activator* (*Ac*) transposase in nuclei of maize endosperm and *Petunia* protoplasts. Plant J. **5:** 705–714.

Henikoff, S., 1992 Detection of *Caenorhabditis* transposon homologs in diverse organisms. New Biol. **4:** 382–388.

Henikoff, S., J. G. Henikoff, W. J. Alford and S. Pietrokovski, 1995 Automated construction and graphical presentation of protein blocks from unaligned sequences. Gene **163:** GC17–26.

Hiom, K., M. Melek and M. Gellert, 1998 DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations [see comments]. Cell **94:** 463–470.

Kempken, F., S. Jacobsen and U. Kuck, 1998 Distribution of the fungal transposon Restless: full-length and truncated copies in closely related strains. Fungal Genet. Biol. **25:** 110–118.

Kidwell, M. G., 1993 Voyage of an ancient *mariner* (news and views). Nature **362:** 202–203.

Koga, A., A. Shimada, A. Shima, M. Sakaizumi, H. Tachida *et al.*, 2000 Evidence for recent invasion of the medaka fish genome by the Tol2 transposable element. Genetics **155:** 273–281.

Kunze, R., U. Behrens, F. U. Courage, S. Feldmar, S. Kuhn *et al.*, 1993 Dominant transposition-deficient mutants of maize Activator (Ac) transposase. Proc. Natl. Acad. Sci. USA **90:** 7094–7098.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. Nature **409:** 860–921.

Morgenstern, B., K. Frech, A. Dress and T. Werner, 1998 DIALIGN: finding local similarities by multiple sequence alignment. Bioinformatics **14:** 290–294.

Nacken, W. K. F., R. Piotrawiak, H. Saedler and H. Sommer, 1991 The transposable element Tam1 from Antirrhinum majus shows homology to the maize transposon En/Spm and has no sequence specificity of insertion. Mol. Gen. Genet. **228:** 201–208.

Okuda, M., K. Ikeda, F. Namiki, K. Nishi and T. Tsuge, 1998 Tfo1: an Ac-like transposon from the plant pathogenic fungus Fusarium oxysporum. Mol. Gen. Genet. **258:** 599–607.

Rubin, E., and A. A. Levy, 1997 Abortive gap repair: underlying mechanism for *Ds* element formation. Mol. Cell. Biol. **17:** 6294–6302.

Saedler, H., 1996 Transposable elements, pp. 1–229 in *Current Topics in Microbiology and Immunology*, edited by A. E. A. Capron. Springer, Berlin.

Schneider, T. D., and R. M. Stephens, 1990 Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. **18:** 6097–6100.

Sherratt, D., 1989 Tn*3* and related transposable elements: site-specific recombination and transposition, pp. 109–163 in *Mobile DNA*, edited by D. E. Berg and M. M. Howe. American Society for Microbiology, Washington, DC.

Simmons, G. M., 1992 Horizontal transfer of hobo transposable elements within the Drosophila melanogaster species complex: evidence from DNA sequencing. Mol. Biol. Evol. **9:** 1050–1060.

Wallace, J. C., and S. Henikoff, 1992 PATMAT: a searching and extraction program for sequence, pattern and block queries and databases. Comput. Appl. Biosci. **8:** 249–254.

Warren, W. D., P. W. Atkinson and D. A. O'Brochta, 1994 The Hermes transposable element from the house fly, Musca domestica, is a short inverted repeat-type element of the hobo, Ac, and Tam3 (hAT) element family. Genet. Res. **64:** 87–97.

Warren, W. D., P. W. Atkinson and D. A. O'Brochta, 1995 The Australian bushfly Musca-vetustissima contains a sequence related to transposons of the Hobo, AC, and Tam3 family. Gene **154:** 133–134.

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

Communicating editor: S. Henikoff