Evolution of Intron/Exon Structure of DEAD Helicase Family Genes in *Arabidopsis*, *Caenorhabditis*, and *Drosophila*

Nathalie Boudet,[1] Sébastien Aubourg,[1,2] Claire Toffano-Nioche,[1] Martin Kreis,[1] and Alain Lecharny[1,3]

[1] *Institut de Biotechnologie des Plantes, Unité Mixte de Recherche-Centre National Recherche Scientifique 8618, Université de Paris-Sud, Bât. 630, F-91405 Orsay Cedex, France*

The DEAD box RNA helicase (RH) proteins are homologs involved in diverse cellular functions in all of the organisms from prokaryotes to eukaryotes. Nevertheless, there is a lack of conservation in the splicing pattern in the 53 *Arabidopsis thaliana* (*AtRHs*), the 32 *Caenorhabditis elegans* (*CeRHs*) and the 29 *Drosophila melanogaster* (*DmRHs*) genes. Of the 153 different observed intron positions, 4 are conserved between *AtRHs*, *CeRHs*, and *DmRHs*, and one position is also found in *RHs* from yeast and human. Of the 27 different *AtRH* structures with introns, 20 have at least one predicted ancient intron in the regions coding for the catalytic domain. In all of the organisms examined, we found at least one gene with most of its intron predicted to be ancient. In *A. thaliana*, the large diversity in *RH* structures suggests that duplications of the ancestral *RH* were followed by a high number of intron deletions and additions. The very high bias toward phase 0 introns is in favor of intron addition, preferentially in phase 0. Results from this comparative study of the same gene family in a plant and in two animals are discussed in terms of the general mechanisms of gene family evolution.

The conservation of the intron–exon organization or gene structure in homologous genes is commonly high enough to show the lineage of introns in evolution (Hardison 1996). When observed, partial departures from the common structure in the duplicated genes may be attributed either to deletions, insertions, or both. The intron early theory suggests that the extant gene structures originated prior to the divergence of prokaryotes and eukaryotes through exon shuffling (Doolittle 1978; Gilbert 1987; Gilbert et al. 1997). In its extreme form, this hypothesis has been used to explain differences in intron distributions between homologous genes by independent intron losses from an ancestral gene containing introns at all of the observed positions in modern genes (Bagavathi and Malathi 1996; Robertson 1998). It has been argued that the diversity in individual intron positions, observed in some of the extensively studied families (Stoltzfus et al. 1997), is rather indicative of the recent origin of introns, that is, the intron late hypothesis (Cavalier-Smith 1985; Logsdon and Palmer 1994; Stoltzfus et al. 1994). Thus, random insertions of introns have been documented (Palmer and Logsdon 1991; Patthy 1996; Cho and Doolittle 1997; O'Neill et al. 1998; Tarrio et al. 1998). Therefore, it has been postulated that duplications of ancestral mosaic genes have been followed by more recent gains and losses of introns (Trotman 1998). The two latter processes are believed to be very slow, as gene structures are often well recognizable between evolutionary distant homologs. However, in some cases, the data indicate drastic steps leading to subgroups of homologous genes, clearly identified by their intron patterns (Gotoh 1998; Paquette et al. 2000; Sanderfoot et al. 2000). The most striking situation is when genes without introns are clear homologs of a family of duplicated genes with a high number of conserved introns (Rzhetsky et al. 1997; Charlesworth et al. 1998; Aubourg et al. 1999; Koch et al. 2000; Paquette et al. 2000; Tavares et al. 2000; Tognolli et al. 2000).

Until now, the question of gene structure evolution was mainly examined either by statistical approaches on the whole set of introns in a given organism or by comparisons of homologous genes from different and often distantly related species. There is one report on small groups of genes belonging to different families of paralogs (for review, see Cho and Doolittle 1997). Therefore, the assumption is made that the evolution of gene structures follows the same rules in all of the organisms and in all of the gene families. However, there are many reasons to suspect the existence of specific evolutionary pressures at these different levels of integration (Robertson 1998). The DEAD box RNA helicase family (RH) presents a number of advantages for studying the evolution of gene structures as follows: (1) a high number of paralogs in higher eukaryotes, (2) a high enough conservation of protein sequences between homologous genes to assign safely the positions of introns, (3) a high number of introns per paralogous families in order to support clear conclusions, (4) a high divergence rate in structures together with a minimum number of shared introns ascertaining the homology, and (5) a discrimination between structures obtained by experimental methods (sequencing of mRNA and gene-mRNA sequence comparisons) of those only predicted. In this work, we present data on *RH* introns from *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster* and compare them with previous data on whole sets of introns from the three organisms.

[2] Present address: Unité de Recherche en Génomique Végétale, INRA, FRE-CNRS, 2 rue Gaston Crémieux, CP 5708, F-91057 Evry Cedex, France.
[3] Corresponding author.
E-MAIL Lecharny@ibp.u-psud.fr; FAX 33-1691-53425.

Article published on-line before print: *Genome Res.,* 10.1101/gr.200801.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.200801.

Helicases are involved in a large number of genetic processes, entailing the unwinding of single-stranded and double-stranded regions of DNA and RNA (Schmid and Linder 1992). The RHs contain a catalytic domain, from 290 to 360 amino acids long, exhibiting 8 specific motifs (Gorbalenya et al. 1993). The most studied RH is the translation initiation factor EIF-4A, known to interact directly with mRNAs and to have an ATP-dependent RNA helicase function (Schmid and Linder 1992). Genomes from prokaryotes contain from one to five *RH*s and there are 26 genes in the *Saccharomyces cerevisiae* nuclear genome (Linder 2000). We previously characterized 32 *RH*s in *A. thaliana* (*AtRH* genes; Aubourg et al. 1999). We now report an exhaustive comparison of the *RH* structures in three phylogenetically distant eukaryotic genomes, namely the genomes of *A. thaliana* (*Arabidopsis* Genome Initiative 2000), *C. elegans* (the *C. elegans* Sequencing Consortium 1998), and *D. melanogaster* (Adams et al. 2000).

The divergence of *RH* structures in *A. thaliana*, *C. elegans*, and *D. melanogaster* are strongly indicative of an evolution of the splicing pattern, independent of the amino acid sequence divergence, massive losses of introns by reverse transcription, and deletion/addition of novel introns. The timing and the relative importance of each of these events in the evolution of *RH* structures are tentatively evaluated.

RESULTS

AtRH, *CeRH*, and *DmRH* Family Organization

The *AtRH* family is composed of 55 different genes (Table 1), of which 2 are disrupted. The structure of the 55 genes was characterized and shown to contain between 0 and 18 introns (Fig. 1A). The *AtRH*s are quite evenly dispersed on the five chromosomes of the *A. thaliana* genome (Table 1), except for the genes *AtRH25* and *AtRH26* that are in tandem and separated by only 450 bp. The 32 *CeRH*s (Table 2A) map to the five chromosomes of *C. elegans* with an even repartition, whereas the 29 *DmRH* genes (Table 2B), are absent from chromosomes 4 and Y. Interestingly, as for *AtRH*s, the *CeRH*s and the *DmRH*s exhibit a relatively large diversity of structures with a number of introns per gene ranging from 1 to 13 for *CeRH*s and from 0 to 11 for *DmRH*s (Fig. 1B, C).

The large majority of both *AtRH*s, *CeRH*s, and *DmRH*s are transcribed, but show large differences in the level of transcription as indicated by the numbers of cognate ESTs found in dbEST (Tables 1 and 2) and by PCR experiments using various cDNA libraries from *A. thaliana* (see Aubourg et al. 1999). Although there is no indication of transcription for 9 *AtRH*s, 6 *CeRH*s, and 2 *DmRH*s, these 17 genes are probably functional, because they all code for at least a complete catalytic domain not interrupted by stop codons or frame-shifts. Two disrupted *RH*s, that is, *AtRH54* and *AtRH55* resulting apparently from a duplication of *AtRH*2 and *AtRH*49, were identified in the complete genomic sequence of *A. thaliana*. Five of the six introns of *AtRH2* are missing in *AtRH54*. Notably, the sequence coding for the conserved PTREL region is altered and the ORF is interrupted twice. In the second putatively nonfunctional *RH*, namely *AtRH55*, the gene structure of *AtRH49* is conserved but three deletions were observed, respectively, of 3, 31, and 291 bp. These deletions do not interrupt the ORF, but some of the conserved amino acids are missing, especially those present in the HRIGR motif, shown to be essential for RNA binding (Schmucker et al. 2000). Hence, *AtRH55* is a nonprocessed pseudogene resulting from a gene duplication, whereas *AtRH54* is a processed pseudogene resulting from a reverse transcription from a mRNA. The two pseudogenes were not used in the present analysis. It is worth noting that the intronless genes *AtRH21*, *AtRH42* and *AtRH47*, and *DmRH16* and *DmRH23* encode a protein with a complete catalytic domain and have cognate ESTs.

RH proteins have long stretches of sequence similarities and many conserved residues (Table 3). Such a level of conservation in protein sequences is strongly in favor of homology (Gogarten and Olendzenski 1999). Nevertheless, the high divergency observed in the structures of the paralogous *AtRH*s, *CeRH*s, and *DmRH*s (Fig. 1) compelled us to consider the possibility that extant DEAD helicases arose through a very unexpected process of convergent evolution (Doolittle 1994). This question was addressed by characterizing all of the *RH* introns, with a particular emphasis on the estimation of the number of introns at identical positions.

Intron Number

The mean number of introns per gene is seven in *AtRH*s, six in *CeRH*s, and three in *DmRH*s (Fig. 2). This is higher than the mean value observed in *A. thaliana*, *C. elegans*, and *D. melanogaster*, with, respectively, 5.2, 4.2, and 2.2 introns per gene (Blumenthal and Spieth 1996; Deutsch and Long 1999; *Arabidopsis* Genome Initiative 2000). Our results reveal that the intron number distribution per *RH* gene is clearly different in the three organisms. In *AtRH*s, the distribution is biphasic with a maximum at 0 to 1 and at 8 to 9 introns per gene, whereas there is only one strong maximum of 4 to 5 introns per gene in *CeRH*s and of 2 to 3 in *DmRH*s (Fig. 2).

Intron Length

The distribution of intron length in *RH*s was not different from that observed for all of the other *A. thaliana* (Goodman et al. 1995; *Arabidopsis* Genome Initiative 2000), *C. elegans* (Blumenthal and Spieth 1996), and *D. melanogaster* introns (Deutsch and Long 1998; data not shown).

Intron Positions

Figures 1A, B, and C present, respectively, the structures of all of the *AtRH*, *CeRH*, and *DmRH* genes and point out the positions used by at least two introns. Altogether, of 345 introns, 244 were found at a strictly identical position in the catalytic region of at least two genes (introns are numbered from 1 to 244 in Fig. 1). After comparison of the structures of the region coding for the catalytic domain, four classes of genes were defined, and based on the following criteria: (1) class I genes exhibit completely identical or partially identical, but not equivocally related, structures, (2) class II genes share at least one intron at an identical position with one other gene of the class, (3) class III genes do not share intron positions with any other gene, and (4) class IV genes are intronless. At the top of Figure 1A and B are illustrated 11 groups of *AtRH*s and four groups of *CeRH*s. They belong to class I, contain from two to five genes, and share a complete or a high degree of structural similarity. There is no gene belonging to class I in *DmRH*s. There is no couple of genes with an identical structure in two different species. Further, for the analysis of the divergence of the structures, each of the 16 groups of class I genes with a similar structure will only be represented by the gene containing the highest number of introns. Hence, the number of different gene structures is reduced to 28 for *AtRH*s, 23 for

Table 1. Summary of the GenBank Relevant Information for the 53 *AtRH* Genes and the two *AtRH* Disrupted Genes

Gene name	Accession no. of the genomic fragment	Chr.	Position of gene in the genomic fragment	Accession no. of associated cDNA	EST no.
AtRH1	Z97339	IV	170118–173400	Y11154	0
AtRH2	AP000417	III	70507–73330	AJ010456	23
AtRH3p	AF058914	V	109702–111767	AJ010457	32
AtRH4	AB019229	III	17945–20999	X65052	116
AtRH5	AC079041	I	6430–9220	AJ010458	8
AtRH6	AC004665	II	37310–40316	AJ010459	3
AtRH7	AB019235	V	1750–5329	X99938	21
AtRH8	AF058919	IV	62389–66000	AJ010460	12
AtRH9	AB022215	III	7628–10500	AJ010461	3
AtRH10	AB008269	V	57004–59719	AJ010462	2
AtRH11	AL137082	III	34109–36962	AJ010463	9
AtRH12	AL137898	III	54356–57749	AJ010464	11
AtRH13	AB028608	III	24793–28802	AJ010465	1
AtRH14	AC009325	III	35013–38140	AB010259	11
AtRH15	AL360314	IV	27738–30548	AJ010466	8
AtRH16	AL161586	IV	86352–89788	AJ010467	5
AtRH17	AC007660	II	19186–21962	AJ010468	1
AtRH18	AB010692	V	63383–66642	AJ010469	6
AtRH19	AC005287	I	31341–33516	X65053	20
AtRH20	AC073944	I	34572–37079	AJ010470	2
AtRH21	U78721	II	20826–23027	—	5
AtRH22	AC005966	I	18043–20689	AJ010471	4
AtRH23	AC010926	I	30447–32447	AJ010472	2
AtRH24	AC002337	II	63359–65893	—	2
AtRH25	AB006697	V	62049–65177	AJ010473	2
AtRH26	AB006697	V	57864–61582	AJ010474	1
AtRH27	AB018108	V	15059–17975	AJ012745	0
AtRH28	Z97341	IV	131027–126677	AJ010475	6
AtRH29	AC002291	I	66898–70477	—	0
AtRH30	AB008265	V	27602–30706	AJ010476	1
AtRH31	AB005234	V	25529–28905	AJ010477	0
AtRH32	AB005232	V	57573–60722	—	1
AtRH33	AC004483	II	53840–58000	—	2
AtRH34	AC006085	I	100551–102558	—	0
AtRH35	AB023044	V	75219–77408	—	0
AtRH36	AC006341	I	23779–25496	—	0
AtRH37	AC007087	II	495–3857	—	2
AtRH38	AL132958	III	12036–14491	—	2
AtRH39	AL049482	IV	34597–37826	—	3
AtRH40	AC011623	III	2630–6599	—	3
AtRH41	AC011664	III	48468–50201	—	1
AtRH42	AC007369	I	59030–62530	—	4
AtRH43	AL035678	IV	32979–34715	—	0
AtRH44	AC021044	I	10042–11981	—	0
AtRH45	AC016661	I	50081–47528	—	0
AtRH46	AL163792	V	48191–51630	—	2
AtRH47	AC025417	I	102230–103986	—	4
AtRH48	AC022355	I	42451–45758	—	0
AtRH49	AC016163	I	76692–78835	—	3
AtRH50	AC016827	III	40624–43921	—	4
AtRH51	AP001303	III	28847–32160	—	1
AtRH52	AL137082	III	50600–53853	—	7
AtRH53	AB022215	III	12887–15391	—	0
AtRH54-pseudo	AC011436	III	79075–80353	—	—
AtRH55-pseudo	AC016162	I	70712–72602	—	—

*CeRH*s, and 24 for *DmRH*s. From these different structures, a scaffold gene, containing all of the different observed intron positions, has been designed for each species (data not shown). The *AtRH*, *CeRH*, and *DmRH* scaffold genes contain, respectively, 83, 66, and 32 introns, of which four intron positions are in common.

As shown in Figure 1, class II is composed of genes with at least one identical intron position. The latter is shared with at least one other gene either in the same or in the other organism. Class II genes are illustrated by the following examples. The first example reveals that the six intron positions of *AtRH10* are independently present in six different genes, namely five *AtRH*s and one *CeRH*. In the second example, the five introns of *DmRH4* are present in five different genes, namely two *AtRH*s, two *CeRH*s, and one human gene, *BAT1*. In *DmRH4*, the positions 231, 233, and 234 are identical to positions 157, 158, and 159 in *AtRH15* and the positions 231–235 correspond to the first, third, fourth, fifth, and sixth in-

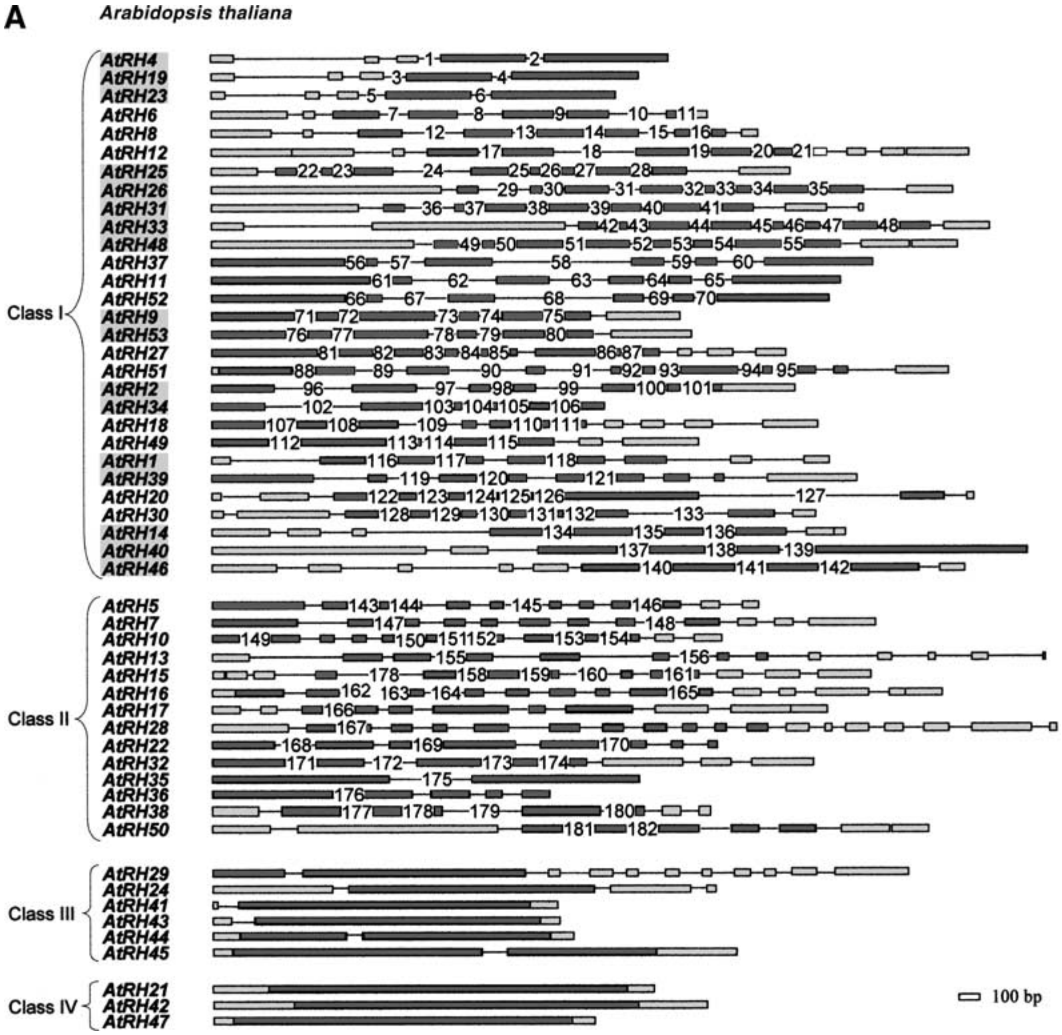

Figure 1 Schematic representation of the *AtRH* (*A*), *CeRH* (*B*), and *DmRH* (*C*) gene structures. Boxes represent exons and lines introns. Lengths are roughly at scale. (Gray) Regions coding for catalytic domains; (white) regions encoding the amino- and carboxy-terminal ends (see Aubourg et al. 1999 for a representation of 32 AtRH protein primary structures). The conserved intron positions are numbered starting from the *top*. Classes based on the conservation of the structures of the genes (see text for details) are separated by an alternative shading. Large *CeRH* and *DmRH* genes are on several lines and broken lines have been used in large introns that are not drawn to scale.

trons in the *BAT1* gene (data not shown). In the third example, the positions 220 and 221 in *CeRH20* are also independently present in two different genes, *AtRH14* and *CeRH24*. In addition, two remarkable intron positions were evidenced in class II. First, the position 116 in *AtRH1* is identical to one position in eight other genes in *A. thaliana* and four in *C. elegans*. Second, one intron is at an identical position in different eukaryotic phyla. This intron is present in *AtRH20* (pos. 127), *AtRH30* (pos. 133), *CeRH26* (pos. 224), *DmRH5* (pos. 236), *DmRH8* (pos. 237), *DmRH25* (pos. 241), in the yeast *DBP2*, and in the human *Hsp68* and *Hsp72*. In *S. cerevisiae* this intron is really unusual both for its position near the 3′ end of the ORF and for its large size (1001 nucleotides) and is unique in the 26 *ScRH*s. A possible role in the autoregulation of *DBP2* transcription has been attributed to this intron by Barta and Iggo (1995).

The intron positions of all of the *RH*s belonging to class II from *A. thaliana*, *C. elegans*, and *D. melanogaster* are detailed in Figure 3. Altogether, in the three sets of nonredundant structures of *RH*s, there are 41 positions occupied by at least two introns from two nonredundant structures. Therefore, ~25% of the positions of the scaffold genes for the three species are occupied by at least two introns from different *RH*s and 17 positions are occupied by three or more introns. Introns are located at position PTREL-48 in 13 of 75 structurally different *RH*s. At three other positions, namely GKT-27, RIV-36, and RIV-45, an intron was identified in five different *RH*s. There are 13 positions that are at least present in two *AtRH*s, two positions in at least two *CeRH*s, one position in two *DmRH*s, 20 in at least two *RH*s, and four in three *RH*s.

Class III groups genes contain all of their introns at a position not used by any other intron from both species. Most of the *AtRH* and *DmRH* genes of class III have no or only one intron in the catalytic region. In contrast, some *CeRH* genes of class III have many introns in this region. Finally, three *AtRH*s, three *DmRH*s, and no *CeRH* are completely devoided of introns in the catalytic region as well as in the amino- and carboxy-terminal extensions (class IV, Fig. 1).

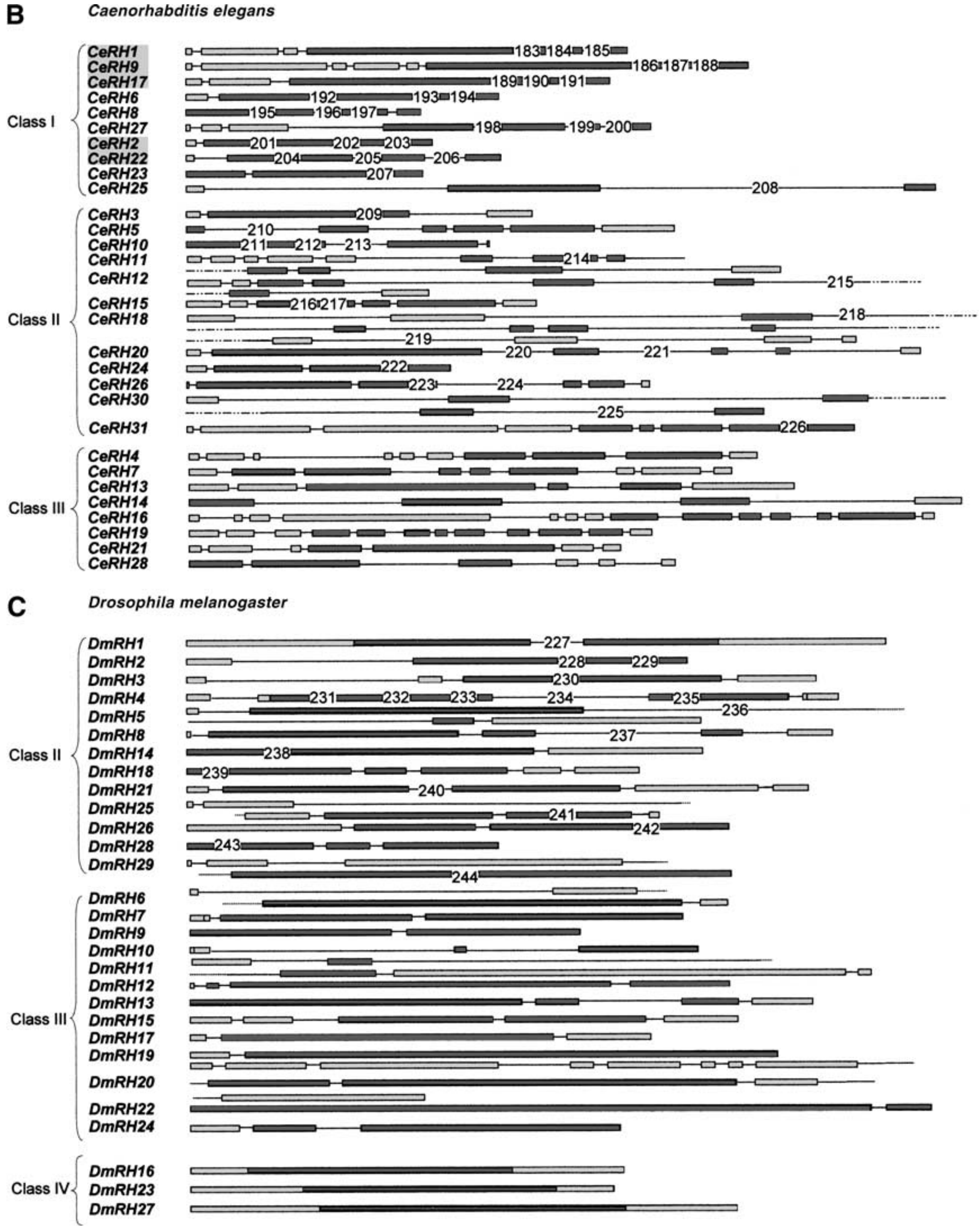

Figure 1 Continued.

Introns at Close Positions

The scaffold gene of 1021 bp built with *AtRHs*, *CeRHs*, and *DmRHs* contains 153 different intron positions. Therefore, it is not surprising that a number of intron positions are separated by only a few nucleotides. For example, three different introns in *AtRH28* are located 2 bp from three different positions; one position is not conserved, whereas the other two are, that is, RI-32 and HRIGR-6, respectively, present in *AtRH7*, *AtRH(49,51)* and *AtRH(10,13,27,51)* (Fig. 3). The third intron, *CeRH28*, is located 1 bp from the position PTREL-48, where an intron is present in 10 *AtRHs* and 3 *CeRHs*. There are two conserved positions, HRIGR-51 and DEAD-14, that are at only 1 bp, respectively, from two other conserved positions,

Table 2. Summary of the GenBank Relevant Information for the 32 *CeRH* (A) and the 29 *DmRH* (B) Genes

Gene name	Accession no. of the genomic fragment	Chr.	Position of gene in the genomic fragment	Accession no. of associated cDNA	EST no.
A. *CeRH1(GLH-1)*	AF000197	I	—	L19948	3
CeRH2	U13876	III	3968–5420	Z12116	9
CeRH3	Z29115	III	667–2700	—	1
CeRH4	L17337	III	72–3427	—	1
CeRH5	Z22177	III	29517–32386	—	3
CeRH6	U53141	V	16276–18110	—	0
CeRH7	U80447	I	1359–4540	—	1
CeRH8	U64840	V	24231–25600	—	0
CeRH9 (GLH-2)	AC006625	I	32092–35382	U60194	5
CeRH10	Z54327	II	7936–9710	U08102	13
CeRH11	Z81449	III	15640–22126	—	7
CeRH12	AF025451	II	5348–10655	—	9
CeRH13	Z81094	V	5559–9108	—	7
CeRH14	AC006661	II	6622–11298	—	0
CeRH15	Z75546	I	19715–21790	—	3
CeRH16	Z50071	II	22054–26452	—	16
CeRH17 (GLH-3)	AF003145	I	20991–23452	AF079509	2
CeRH18	AF045641(1)	IV	244–6828	—	17
	AC024743(2)		6983–11678	—	
CeRH19	AC006665	I	9308–11991	—	7
CeRH20	U13070	III	6571–10890	—	9
CeRH21	AC024810	I	174–2693	—	8
CeRH22	U13876	III	13723–15571	Z12116	0
CeRH23	AF039720	I	33547–34913	—	7
CeRH24	AC006605	III	12520–14052	—	22
CeRH25	AC024844	I	11854–17095	—	0
CeRH26	Z81555	V	24571–27287	—	28
CeRH27	AF125963	V	40937–43739	—	0
CeRH28	AF067608	I	91–2919	—	8
CeRH29 p	AC024830(1)	IV	48090–52906	—	13
	AF100655(2)		25103–27179	—	
CeRH30	AL034488	II	25330–32313	—	5
CeRH31 (GLH-4)	AF039718	I	13970–18076	AF079508	7
CeRH32 p	AF099926	IV	32801–37939	—	0
B. *DmRH1*	AE003442	X	121862–124936	—	6
DmRH2 (EIF-4A)	AE003612	2L	26712–28832	AF0145621	454
DmRH3 (ME31B)	AE003628	2L	2590–5260	M59926	48
DmRH4	AE003610	2L	100868–103158	X79802	107
DmRH5	AE003560	3L	253987–259042	—	56
DmRH6	AE003679	3R	224119–228338	—	21
DmRH7 (DBP73D)	AE003526	3L	134357–136470	—	14
DmRH8 (RM62)	AE003601	3R	36608–40658	X52846	287
DmRH9	AE003588	2L	107096–108768	—	2
DmRH10 (HLC)	AE003568	X	133175–135266	—	55
DmRH11	AE003505	X	177770–185431	—	12
DmRH12 (DDX1)	AE003597	3L	69846–72205	—	13
DmRH13	AE003792	2R	40556–43229	—	0
DmRH14	AE003506	X	85547–88101	—	20
DmRH15 (PIT)	AE003737	3R	80600–83045	—	9
DmRH16	AE003678	3R	68411–70249	—	4
DmRH17	AE003522	3L	263833–265729	—	10
DmRH18 (DBP45A)	AE003834	2R	223722–225584	—	8
DmRH19	AE003659	2L	189794–192322	—	11
DmRH20	AE003468	3L	4016–9725	—	25
DmRH21	AE003838	2R	140941–143499	—	18
DmRH22 (DHH1)	AE003548	3L	97064–100208	—	8
DmRH23 (ABS)	AE003607	3R	33704–35563	—	1
DmRH24	AE003669	2L	132007–133776	—	22
DmRH25 (VAS)	AE003646	2L	69330–75084	—	6
DmRH26	AE003677	3R	151098–153380	—	28
DmRH27	AE003547	3L	9256–11628	—	0
DmRH28	AE003678	3R	113745–115120	—	18
DmRH29	AE003498	X	254832–259105	—	26

Table 3. Sequence Comparisons between the 10 Most Conserved Regions of RH Proteins

A.

	%	GKT	PTREL	TPGR	DEAD	SAT	RI	RII	RIII	RIV	HRIGR
AtRHs	Id.	44–52	24–41	17–66	25–38	24–46	22–22	23–39	33–48	60–65	39–65
	Sim.	68–72	41–62	27–83	53–62	55–55	43–50	41–59	55–55	80–90	61–74
CeRHs	Id.	28–42	21–62	20–46	31–53	30–39	29–29	15–31	26–43	40–85	48–74
	Sim.	54–58	38–83	39–63	53–69	55–64	36–36	41–51	52–57	60–90	65–82
DmRHs	Id.	35–82	—	20–31	24–33	19–19	—	24–38	40–70	—	61–67
	Sim.	73–96	—	65–67	79–79	58–66	—	64–70	85–95	—	78–89
AtRHs/	Id.	39–59	37–53	20–37	19–78	21–48	29–43	23–71	36–64	48–86	65–65
CeRHs	Sim.	51–80	53–63	27–46	41–84	55–78	43–50	40–89	50–81	57–90	65–83
AtRHs/	Id.	28–58	71–71	34–49	24–48	26–36	—	17–43	23–23	50–55	61–83
DmRHs	Sim.	72–82	93–93	52–78	70–79	71–73	—	52–67	46–46	90–90	89–100
CeRHs/	Id.	36–56	50–50	20–56	30–45	18–79	—	12–31	22–22	40–90	53–61
DmRHs	Sim.	64–80	75–75	57–88	70–79	59–94	—	44–52	43–43	85–100	74–84

B.

```
GKT    At   G α e r P T  p I Q A a A β P β β β x ( ) G r ( ) D β β G a A r T G S G K T L A F l P β β e x β x x x x x
       Ce   g i x t P T  p I Q a a α I P x β β e ( ) G r ( ) D β β G x A x T G S G K T L A F β P β β x x β l δ x x x
       Dm   G ε x x P T  p I Q α x α I P β β L x ( ) G r ( ) D β β g x A x T G S G K T l A F l P β l x x β x x x x x

PTREL  At   a p r A L I  x α P T R E L A x Q v x ( ) x x x  x x β α k x x
       Ce   g l q A V l  β v P T R E L A x Q I f ( ) k E f  l k l g d y l
       Dm   x x x A L v  β α P T R E L A x Q I x ( ) x x x  x x x x x x x

TPGR   At   g β r v x v  β G G α x x p x Q x R x L x r G ( ) p δ I β V α T P G R β x D h β E x α
       Ce   N β k v x c  a I G G g k I d E q i a d l k G ( ) a e β V V α T P G R β i D β β q k g
       Dm   x β r x α β  β G G α x x x q x x x x l x x g ( ) x d i β β a T P G R l β D β β x x x

DEAD   At   L d n L k γ  L V β D E A D R β L d x x ( ) G ( ) F e d q β x x β β q x β P
       Ce   l x x β r γ  L V β D E A D R R M l d x ( ) g ( ) F E d q β x x β x n x β P
       Dm   l x x β x γ  L V L D E A D R m L d β x ( ) G ( ) F e x x β x x i x x x β x

SAT    At   p x R Q T l  L F S A T x p s e V x ( ) x L x L a r f k δ P v k i x x v
       Ce   x q k Q T β  L F S A T F P r e β q ( ) x f A K k x β d δ P β e V m V g
       Dm   x x r q t β  β f S A T β p x x v x ( ) x l a x x x L x δ p β x β x β α

RI     At   x t x x g ( ) β x Q e f v v x x
       Ce   k p t e r ( ) V e Q v v y m V P
       Dm   x α α x x ( ) β x q x β x x β x

RII    At   x e k k x x  L l x l L β x ( ) K x β I F c x T K r x v d x L x x
       Ce   d e K k a k  β β e l L k n ( ) K v β I F c q T K r d V D a β A e
       Dm   δ x x k x x  x l β x l β x ( ) x β x i F c δ t k x x β d x l α x

RIII   At   l L x x l G ( ) β k A x x β H G δ β t Q s x R l k a L x x F R a G x x x x L β A T D V A
       Ce   β β R s g G ( ) β p α β s β H G d q δ Q e e R d x α L n q F K s G k y q β β β A T D V A
       Dm   x x α β x G ( ) β x x x x β H G δ β x Q x e R δ x x β x x F r δ g x x x β L β A T D V A

RIV    At   A R G β D v ( ) P x V x l V V Q Y δ l P n d
       Ce   α R G I D V ( ) q D V x L V I N Y D β P n N
       Dm   a R G L D β ( ) β x v x x V I N Y D x P x x

HRIGR  At   s E d Y β H  R V G R α G R a G r k G
       Ce   I E D Y I H  R I G R T G R α G k K G
       Dm   x e δ Y b H  R β G R T G R β G x x G
```

Minimum and maximum identities and similarities are given. The consensus sequences of the 10 regions have been obtained from multiple alignments of the proteins translated from the 3 nonredundant sets of genes with intron in the 10 regions. The accepted conservation is: α, A/T/G/S: β, I/L/M/V; δ, D/E/N; γ, F/Y/W. The letter x denotes any amino acid. Brackets indicate gaps inserted in the alignment to build up the consensus. A dash denotes the absence of a sequence with intron in these conserved regions.

namely HRIGR-52 and DEAD-15. Of the 153 different intron positions in the scaffold genes, 15 positions are at +/− 1 bp, 23 at +/− 2 bp, 30 at +/− 3 bp, and 54 at +/− 5 bp from an other intron position.

Intron Phases and Exon Types

The percentages of intron phases were determined in the complete *RH* sequences, (Table 4). The weak differences observed between data from validated and predicted introns indicates that our predictions are essentially correct.

Phase 0 is highly over-represented in *AtRHs* (69%), that is, two times the expected value for an addition of intron with an equal probability for the three codon sites. Phase 0 dominates in *CeRHs* with 41%. In *DmRHs*, the numbers of introns in the three phases are close to one-third, and no conclusion can be deduced from small differences due to the small number of introns. Distributions of intron phases and exon types

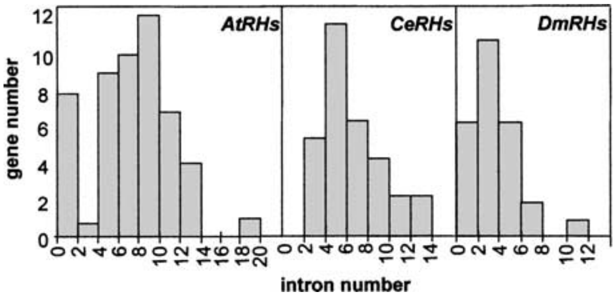

Figure 2 Repartitions of *AtRH*, *CeRH*, and *DmRH* intron number.

have been reported previously in *A. thaliana*, *C. elegans*, and *D. melanogaster* (Long et al. 1998) (values between brackets in Table 4). The above results show a larger bias toward phase 0 for the *AtRH* introns compared with the *A. thaliana* introns in general. In contrast, there is no difference in the representation of intron phases between *CeRH*s and all of the other genes from *C. elegans*. In *A. thaliana*, the observed bias of intron phases was even higher when only the introns from the catalytic domain of *AtRH*s were considered (data not shown), because >76% of introns were in phase 0. Conversely, the bias toward phase 0 in the amino- and carboxy- terminal regions was slightly less than for *A. thaliana* introns in general. Although the same tendency was observed in *CeRH*s, the differences were only a small fraction of what was observed in *AtRH*s. If the number of positions at a given phase is considered instead of the number of introns, the data are only slightly changed, and the overall bias observed above is about the same.

As a consequence of the high over-representation of intron phase 0 in *AtRH* genes, symmetrical exons of type 0–0, representing 48%, are largely in excess compared with symmetrical exons of type 1–1, (2%), or 2–2, (3%). In the case of the *CeRH* genes, the percentage of symmetrical exons of type 0–0 and 1–1 represent, respectively, 18% and 15%, approximately double the symmetrical exons of type 2–2. In the case of *DmRH* genes, the percentage of symmetrical exons of both type 0–0 and 1–1 represent 11%, and 9% for symmetrical exons of type 2–2. Another direct consequence of these percentages of symetrical exons is that a majority (54%) of *AtRH* exons are 3N-bp long (N is an integer), only 20% 3N + 1-bp long and 26% 3N + 2-bp long. Again, this excess of 3N-bp long exons in *AtRH*s is higher than in *A. thaliana* introns in general (3N = 0.44, 3N + 1 = 0.29, 3N + 2 = 0.30). In *C. elegans*, however, no difference has been observed between the repartition of *CeRH* exons and *C. elegans* exons in the three length classes (3N = 0.42 and 0.40, respectively; 3N + 1, 0.29, and 0.32; 3N + 2, 0.29, and 0.28). Furthermore, the bias toward 3N exons is less pronounced than in *A. thaliana*. In *D. melanogaster*, no significant difference has been observed between the repartition of *DmRH* exons in the three length classes (3N = 0.35 and 0.40, respectively; 3N + 1, 0.32, and 0.32; 3N + 2, 0.33, and 0.28).

Intron Sequences

Each *RH* intron sequence has been compared by BLAST against all other *RH* intron sequences. Only alignments with both an e-value less than e-10 and a percent of identity >90% were considered. As expected, the sequence comparisons of *AtRH* introns did not allow validation of the hypothesis of an ancestral relationship between close intron positions in different genes. The evolution of intron sequences is rapid and a significant conservation can only be observed in genes resulting from recent duplications. Therefore, only significant identities between introns at identical positions were observed in three pairs of genes with similar structures. This type of sequence conservation may help to track down the timing of the duplication events in groups of genes with a similar structure. In three groups of genes, high-sequence identity was detected between conserved introns. For instance, the sequences of the *AtRH33* introns are identical to those of *AtRH48*, except for the first intron, which is absent from *AtRH48* and for the third intron of *AtRH33* (72-bp long), which shares only 40 identical basepairs with the 105 bp of the second intron of *AtRH48*. Although these two genes belong to the same group of duplication as *AtRH25–AtRH26* and *AtRH31*, no intron sequence conservation has been observed between the latter three and *AtRH33* and *AtRH48*. These data iindicate that the duplication event between *AtRH33* and *AtRH48* is the latest duplication that occurred in this group. In the 5′ untranslated region of *AtRH19* and *AtRH4*, the first intron of both genes contains a conserved sequence of 49 bp. The first intron of *AtRH4* is 117-bp longer than the first *AtRH19*, and the region of similarity is shifted by 117 bp from the beginning of introns. Therefore, either a deletion or an insertion event after the duplication of the genes is highly likely. *AtRH23* belongs to the same structural group as *AtRH4* and *AtRH19*, but no significant conservation has been detected between the first intron of *AtRH23* and the first intron of *AtRH4* and *AtRH19*. The lack of conservation could be due to the fact that the duplication between *AtRH19* and *AtRH4* is more recent than the duplication between *AtRH23* and *AtRH4–AtRH19*. Moreover, an identical sequence of 72 bp is present in the third intron of *AtRH14*, from 300 to 372 bp, and in the third intron of *AtRH46* from 1 to 73 bp. No sequence conservation was detected in the third intron of *AtRH40*, which is the third gene of the duplication group. This latter result suggests that the duplication between *AtRH14* and *AtRH46* is more recent than the duplication between *AtRH40* and *AtRH14–AtRH46*. There are no significant similarities in intron sequences between the different *CeRH*s and between the different *DmRH*s. No significant stretches of conserved sequences have been found in introns at identical positions in both species and especially in the intron common to *A. thaliana*, *C. elegans*, *D. melanogaster*, human, and yeast.

DISCUSSION

Gene Structure Divergence

AtRH, *CeRH*, and *DmRH* genes exhibit a large diversity of structures, although protein sequences in the catalytic domain are well conserved. To our knowledge, families of genes with a relative number of different structures as large as the one described in this work for the *RH* family have not been reported previously. Even though a very conservative evaluation of this diversity estimates 28, 23, and 24 different structures for *AtRH*s, *CeRH*s, and *DmRH*s, respectively. Interestingly, these figures are similar to the number of yeast *RH*s, 26 genes with evidence for one recent duplication. This suggests a link between the number of essential functions and the number of different gene structures in the *RH* family. The number of introns in the present-day genes ranges from 0 to 18 per gene. Results showing large differences in the structures of paralogs have been published recently for two other

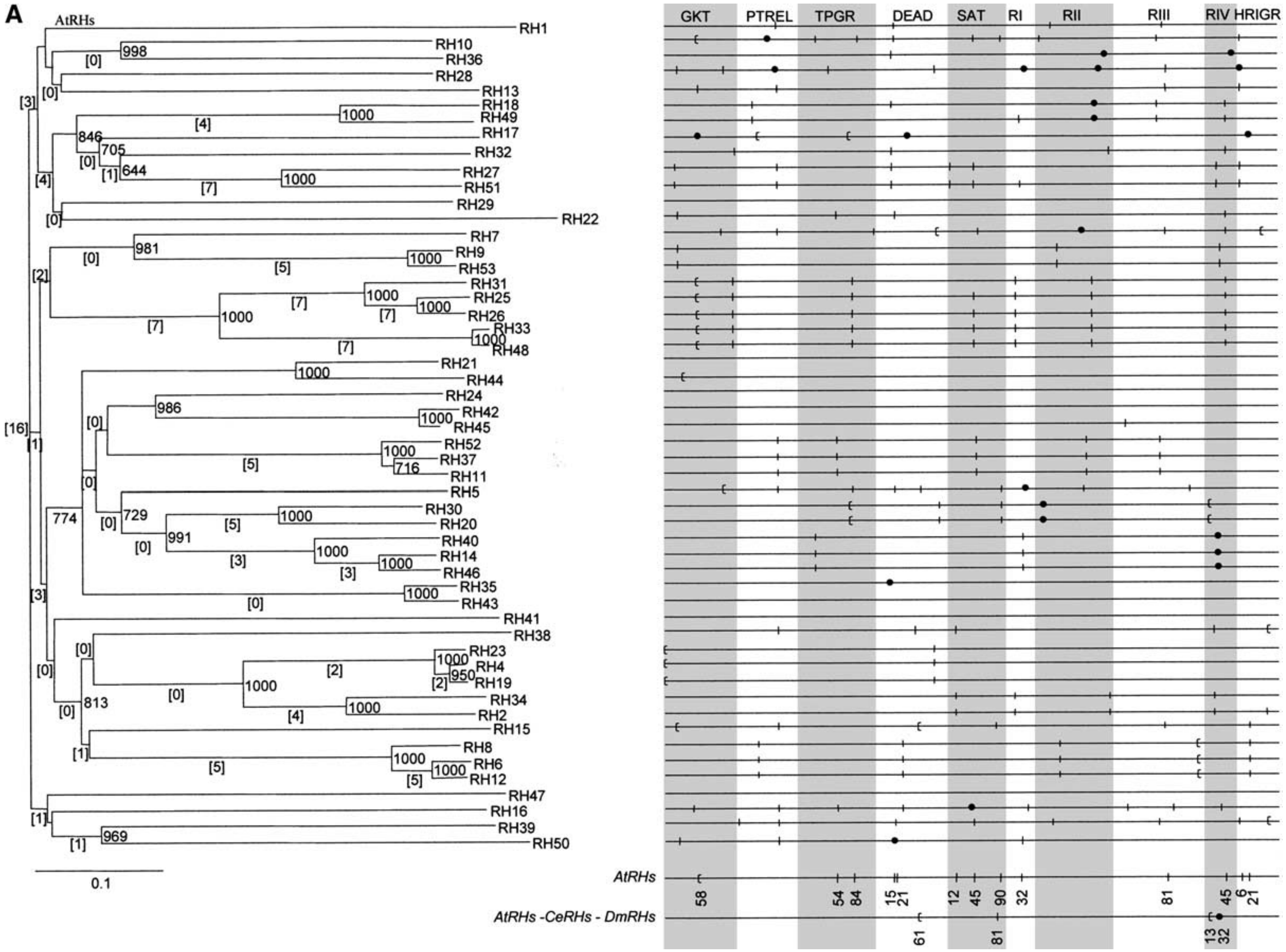

Figure 3 Neighbor-joining trees of AtRH (*A*), CeRH (*B*), and DmRH (*C*) proteins and intron positions in the 10 most conserved regions of the catalytic domain for *AtRH*, *CeRH*, and *DmRH* genes. (square brackets) The numbers of introns conserved between the genes in a branch. The bootstrap values for 1000 trials are indicated at each fork. The horizontal bar represents 0.1 substitution per nucleotide. Intron positions in the overall alignments of sequences are as follows: (|) phase 0 introns; ([) phase 1 introns, and (•) phase 2 introns. Conserved protein regions are named as in Table 1. Conserved intron positions in *AtRHs*, *CeRHs*, and *DmRHs*, between *AtRHs* and *CeRHs*, *AtRH* and *DmRHs*, and *CeRHs* and DmRHs, as well as between the three organisms, are represented on the scaffold structures at the *bottom* of (*A*), (*B*), and (*C*). Numbers below the intron position are the positions of introns in base pairs in the considered conserved protein region. (Figure continues on following page.)

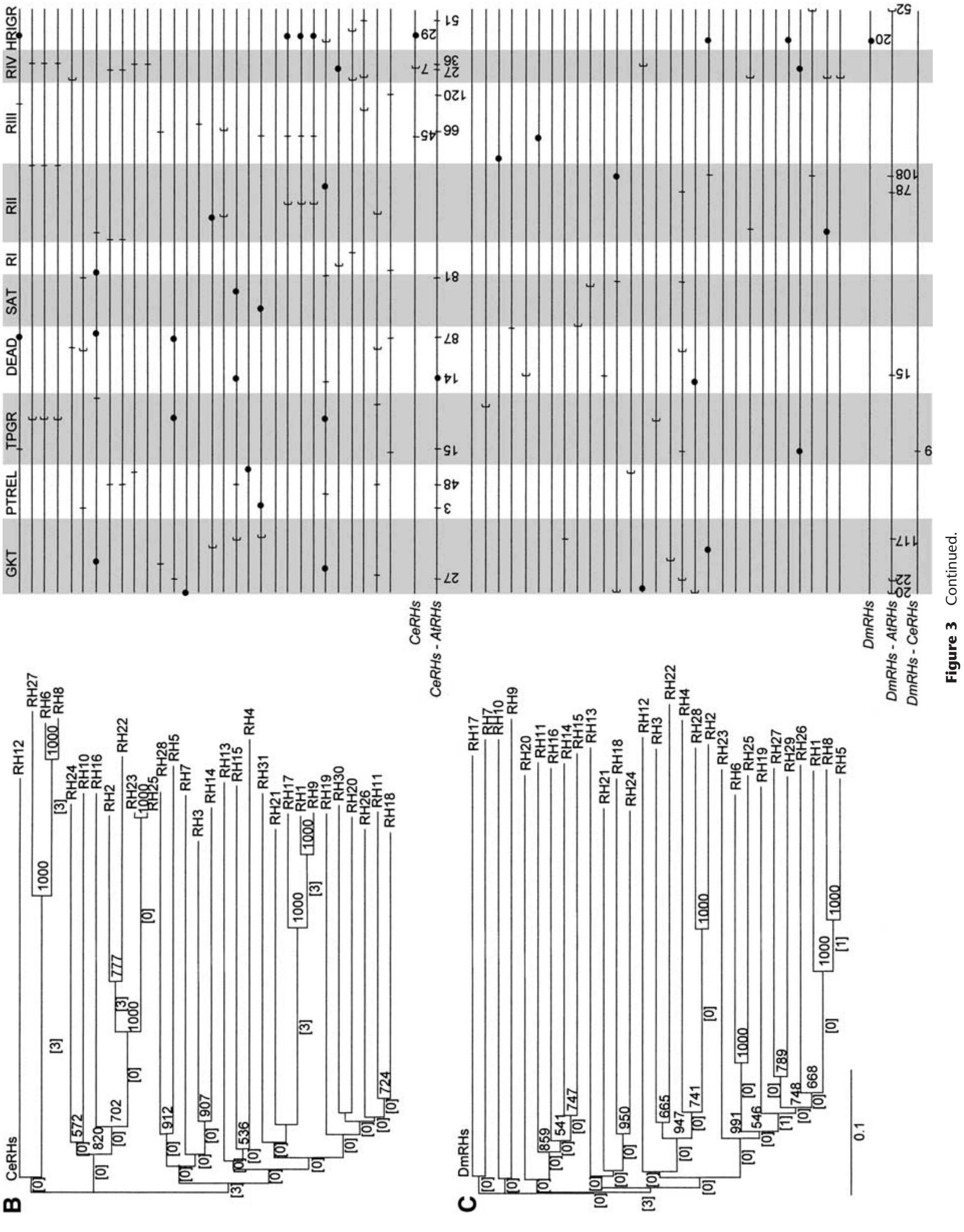

Figure 3 Continued.

Table 4. Repartitions of Intron Phases in *AtRHs*, *CeRHs*, and *DmRHs*

		Verified introns			Predicted introns			Mean value		
	Phase	0	1	2	0	1	2	0	1	2
AtRHs	%	67,5	12,5	20	71	13,5	15,5	69 {56}	13 {23}	18 {21}
	intron nb	144	27	42	106	20	23			
CeRHs	%	40,5	36,5	23	41,5	31,5	27	41 {47}	34 {29}	25 {24}
	intron nb	41	37	23	37	28	24			
DmRHs	%	32	36	32	48	24	28	37 {46}	32 {31}	31 {24}
	intron nb	20	22	20	12	6	7			

Numbers in brackets are values representing the complete sets of *Arabidopsis thaliana,Caenorhabditis elegans,* and *Drosophila melanogaster* genes (Deutsch and Long 1998).

gene families of *A. thaliana*. A recent study of the 135 *A. thaliana* cytochrome P450 (*AtCYPs*) showed that *AtCYPs* of the A type may be classified into four different structural groups with 0–6 introns and *AtCYPs* of the non-A type into six structural groups with 0–13 introns (Paquette et al. 2000). The 24 *A. thaliana* syntaxins (*AtSYPs*), containing between 0 and 12 introns, are classified in 10 groups with different splicing patterns (Sanderfoot et al. 2000). There is no apparent common characteristic between RHs, CYPs, and SYPs. Our working hypothesis proposes that the divergence of gene structure of large families is associated with the expansion of the number of the members of the family from an ancient paralog.

Ancient Introns

Despite the lack of conserved structures in the *AtRH*, *CeRH*, or *DmRH* genes, 4 identical intron positions have been identified between homologs from *A. thaliana*, *C. elegans*, and *D. melanogaster* and 20 intron positions are identical between homologs from two organisms. Introns whose positions have been shown to be maintained in genes from organisms phylogenetically distant are generally considered as ancient introns. Following this criterium, in 74 *RH* structures with introns, 43 have at least one predicted ancient intron in their catalytic region. This is a conservative evaluation of the number of positions with a common origin, as there are indications that intron sliding by one or two bases might well be a real, although rare, phenomenon (Jellie et al. 1996; Stoltzfus et al. 1997; Rogozin et al. 2000). Sliding can be defined as the movement of the intron–exon boundaries over short distances. Thus, in the 21 positions observed at +/− 2 bp from another position, at least some may well be due to exons having slid from an ancient position. Two other data are in favor of ancient introns in *RHs*. First, the three intron positions in *AtRH14–AtRH40–AtRH46*, as well as three of four in *AtRH50* and the five intron positions in *DmRH4* are also present in genes from another species. In *DmRH4*, three positions are present in three different organisms and two in four. Furthermore, the five *DmRH4* positions are also observed in the *BAT1* human gene. Second, one intron present in *AtRH20*, *AtRH30*, *CeRH26*, *DmRH5*, *DmRH8*, and *DmRH25* is also present strictly at the same position, in one DEAD-box RNA helicase gene from *S. cerevisiae*, and *Homo sapiens*. The conservation of this intron between unicellular and pluricellular eukaryotes strongly suggests that it is an ancient intron, thus present in an ancestor gene containing at least one intron. This intron might have been maintained in one paralog in each organism because of its regulatory role. Even if it may be argued that some of the introns at identical positions in only two species may have been inserted by chance and by independent events (see discussion below), the four conserved intron positions observed in three or more than three species are strong evidence for ancient introns. Therefore, our data confirm the assumption, on the basis of sequence comparisons, that the RH family has been formed by duplications of ancient genes containing introns, followed by divergence of the copies and not by formation of similar genes by convergence events. Thus, the remaining question is how and why did the *RH* structures diverge so drastically?

Recent Gene Duplications and Reverse-Transcribed Genes

An overall comparison of the gene structures of *AtRHs*, *CeRHs*, and *DmRHs* suggests that the same events could explain the evolution of the gene structure of the family in these organisms. Of course, some of these events predate the separation of the plant phyla from the animal phyla. Nevertheless, the presence of some completely and largely conserved structures between paralogs and the possibility of finding some sequence conservation in introns indicate the occurrence of relatively recent events of duplication in both organisms. Consistently, the three distant trees (Fig. 3A,–C) show that the genes with the same structures (except for *AtRH1* and *AtRH39*) are grouped in the same terminal branches, supported by very high bootstrap values, although the length and sequence of introns have diverged drastically. However, the bootstrap values drop drastically in the inner branches when the conservation of the gene structures are no longer observed. Some exceptions are remarkable because they indicate that a high conservation in the protein sequences is not correlated with a conservation of the gene structures. For instance, this is the case for *AtRH10* and *AtRH36* or *AtRH20* and *AtRH14*. Moreover, the distance trees show that genes without intron or with no intron in the catalytic domain (*AtRH21*, *AtRH24*, *AtRH42*, *AtRH43*, *AtRH41*, *AtRH47*) are generally not related to any other gene in the tree except for the genes with only one intron. For these genes, it is not possible to safely design by sequence comparisons the paralog that would have been generated by the same event of duplication. This is possible only in the case of the two disrupted genes, as sequence identity is very high between *AtRH2* and the disrupted *AtRH54* and between *AtRH49* and the disrupted *AtRH55*. Therefore, the genes at the origin of the creation of the genes without introns either have largely diverged from their coduplicated gene or the latter has been deleted.

An hypothesis that might help to organize all of the apparently conflicting data assumes two different mechanisms of evolution for the above gene families. The first mechanism involves the duplication of genes with the expected conservation of the structures, whereas the second would be an event of reverse transcription of mRNAs with recombination of the synthesized cDNA in the genome. The latter mechanism, formerly proposed by Lewin (1983), Fink (1987), and Martinez et al. (1989), would explain the formation of genes without intron from paralogous genes with introns. This hypothesis has been discussed more recently by Liaud et al. (1992), Frugoli et al. (1998), and Charlesworth et al. (1998). Reverse transcriptase may have two different origins in the cells, retrotransposons, and retroviruses. The sequence of the nuclear genome of *A. thaliana* contains several hundred sequences, indicating the existence of present or past sequences potentially coding for a reverse transcriptase. The impossibility to establish a clear relation between *AtRH*s without introns and any other *AtRH* except for the disrupted genes suggests that only events of reverse transcription followed by a homologous reinsertion were efficient. Hence, homologous reinsertions place the cDNA always downstream of the promoter of the cognate gene but only rarely in the case of a heterologous reinsertion. The reverse-transcribed *AtRH*s, except for *AtRH54* and *AtRH55*, are not what is generally called processed pseudogenes; they are not subject to obvious disablements relative to their functioning homologs.

Intron Deletions

The data presented in this study shed new light on the contribution of a reverse-transcription mechanism in the formation of large gene families. The biphasic characteristic of the repartition of the number of introns in *AtRH*s is an indication of the existence of two different populations of genes. First, the genes with a relatively high number of introns are most likely genes duplicated from an ancient gene formed by shuffling of small exons (Gilbert et al. 1997), and thus contain many introns. Thereafter, the genes have undergone individual additions or deletions of introns. Second, the eight *AtRH* genes with 0 or 1 intron have been generated by reverse transcription and recombination, affecting either the whole transcript or only a part of it. Deletions of different introns in different paralogs, bringing about a concomitant increase of exon size and number of intron positions in the scaffold genes, might have been at the origin of a large part of the presently observed diversity in *RH* structures. In compact genomes, the ratio between DNA loss and gain determines the size of the genome (Kirik et al. 2000; Petrov et al. 2000). In *A. thaliana*, *C. elegans*, and *D. melanogaster*, the intron length repartitions indicate a tendency to reduce the size of introns, which could explain a number of successful complete deletions. Recombinations and conversions between reverse-transcribed genes and genes with introns might also be a mechanism for intron deletions (Clegg et al. 1997).

Intron Additions

New insertions of introns could have occurred through two possible mechanisms. First, intron insertions may result from a duplication of a pre-existing intron. Recent insertions have been observed in the *Xdh* gene in *Ceratitis capitata*, *C. willistoni*, and *C. saltans* groups of *Drosophila* (Tarrio et al. 1998). Second, intron insertions may also have occurred via transposon insertions (Nouaud et al. 1999). Thus, in maize, ~5% of the transcripts of the *Sh2* gene with a Ds insertion were correctly spliced (Giroux et al. 1994). Using sequence comparisons, we did not observe, either in *AtRH*s, *CeRH*s, or *DmRH*s, any indication of a new addition of an intron resulting from a duplication or from a transposon insertion event. It should be noted, however, that intron sequences diverge rapidly as indicated by the deletions/insertions of introns from duplicated genes with a very high identity in their exon sequences. Nevertheless, in the *AtRH* distance tree, the terminal branch containing *AtRH24*, *AtRH42*, and *AtRH45* (Fig. 3A) indicates that the latter three genes derived from a reverse-transcribed gene without introns and that *AtRH45* recently acquired an intron that is absent from both *AtRH42* and *AtRH24*. The size distributions of introns from *AtRH*s, *CeRH*s, and *DmRH*s are similar to those observed in all of the other genes from their respective organisms. Hence, there is no general indication of a specific kind of intron linked to the rapid divergence of the gene structures.

In *AtRH*s, the bias observed toward introns in phase 0 is higher in *AtRH*s than in all of the other *A. thaliana* genes, 69% and 56% respectively. This difference is even increased if only introns in the catalytic region are considered (76.5% introns in phase 0), but does not exist in the amino- or carboxy-terminal extensions (51.5% introns in phase 0). This indicates, as far as introns are considered, that the two regions of the *AtRH*s have not been submitted to the same type of evolution. In this family of genes, it is expected that the pressure from the protein against intron addition (Fichant 1992) is high in the catalytic domain and low or absent in the extensions that are characterized more by a general composition in amino acids than by a sequence (data not shown). Therefore, the difference in the relative number of phase 0 introns between the catalytic domain and the extensions might be due to a higher success of intron addition in any phase in the latter.

The four intron positions conserved in the three species are in phase 2 for two and in phase 0 and phase 1 for the other two. Other intron positions in common between *CeRH*s and *AtRH*s or between *DmRH*s and *AtRH*s are mainly in phase 0. Whereas *D. melanogaster* and *C. elegans* are phylogenetically closer than *A. thaliana* and *D. melanogaster* or *C. elegans* (Baldauf et al. 2000), the numbers of intron positions in common are higher between *A. thaliana* and *D. melanogaster* (seven) or *C. elegans* (four) than between *D. melanogaster* and *C. elegans* (one). Therefore, some of the common positions observed in only two species are probably not ancient but are rather insertion at the same position independently in *AtRH*s and in one of the two other species.

Conclusion

A comparison of *AtRH*, *CeRH*, and *DmRH* families suggests that the same mechanisms of intron gain or loss have been used during the formation of this family in the three organisms. The existence of reverse transcription in *CeRH*s is not as evident as in *AtRH*s or *DmRH*s, due to the absence of an intronless gene in the nematode. The relative number of genes without introns, ~20%, is nevertheless about the same in both *A. thaliana* and *D. melanogaster*. Therefore, we suggest that in *CeRH*s, evolution of gene structures by reverse transcription stopped earlier than in *AtRH*s and *DmRH*s. During evolution, there was a different balance among species between massive deletions of introns through reverse transcription on one hand and duplication of genes followed by deletions and additions of introns on the other hand. Thus, yeast genes lost almost all of their ancient introns by reverse transcription (Fink 1987), and intron addition has not played an active role thereafter. In *C. elegans*, *D. melanogaster*, and *A. thaliana*, reverse transcription was a less efficient, but nevertheless important, mechanism involved in the evolution of gene families. The presence of regulatory elements in introns is now a well-documented fact and, therefore, evolution of introns might have a role in the evolution of functions in paralogs.

METHODS

Data Mining and Sequence Analysis

Different programs were used to search and analyze genomic sequences, transcripts, and proteins. An extensive screening of databases (dbEST, GenBank, ACEDB, HTGS) was performed using the different BLAST algorithms (Altschul et al. 1997). The positions of the *AtRH* genes on the five chromosomes of *A. thaliana* were established using the TAIR map viewer server (http://www.arabidopsis.org/servlets/mapper). An extensive screening identified 55 *AtRH*, 32 *CeRH*, and 29 *DmRH* genes. When available, sequence database annotations of gene structures were considered, but a systematic re-evaluation of the predictions was achieved by use of the most efficient prediction tools (Pavy et al. 1999). The structures of the *AtRH*s were predicted by use of the NetPlantGene or NetGene2 (Hebsgaard et al. 1996) and Genemark.hmm (Lukashin and Borodovsky 1998) software programs, especially trained for *A. thaliana*. In the case of the *CeRH*s and *DmRH*s, the putative splicing sites and the potential coding regions in anonymous genomic sequences were predicted by Netgene2 (Hebsgaard et al. 1996), trained for *C. elegans* and Genemark.hmm (Lukashin and Borodovsky 1998), trained for *D. melanogaster*. The comparison of *AtRH* and *CeRH* intron sequences has been realized with the BLASTN program (Altschul et al. 1997). EST sequences, together with all of the mRNA/cDNA sequences available (full-length or partial), were used to validate the positions of introns by alignments with the genomic sequences. The number of ESTs corresponding to each *RH* was counted in order to have an idea of the transcriptional expression of the genes. Our analysis also included one *RH* from *S. cerevisiae*, *DBP2* (accession no. L11574), three *RH*s from *Homo sapiens* [*BAT1*, *p68*, and *p72* (accession nos. Z37166, X15729, and U59321, respectively)].

Intron Positions

The positions of introns were obtained from nucleotide sequence alignments derived from the protein alignments. Only introns in the catalytic domain were exploited in this work. Regions outside of the catalytic domain were not taken into consideration because they code for protein extensions that are variable in RHs. The consensus alignment of the catalytic domain of the proteins AtRHs, CeRHs ,and DmRHs was obtained by running CLUSTALW (Thomson et al. 1994) and after manual modifications to correct obvious mispairings. The 10 most conserved regions with essentially unambiguous alignments have been extracted, namely the GKT, PTREL, TPGR, DEAD, SAT, and HRIGR regions, as defined previously by Schmid and Linder (1992) and the RI, RII, RIII, and RIV regions as defined in this work (Table 3). All together, these regions cover >80% of the complete catalytic domain. A total of 382 introns interrupt the nucleotide sequences coding for the catalytic domains of *AtRH*s, *CeRH*s, and *DmRH*s, but only the 345 introns present in the 10 conserved regions have been used. The positions of 217 introns have been validated by alignments with either cDNAs or ESTs. An identical position for two introns in two different genes may or may not reflect a common origin. Therefore, a distinction has been made between introns and their positions. In this latter case, each position is considered only once, despite the number of introns found at this position in different genes. An intron can be located between two codons (phase 0) or within a codon, lying either after the first or after the second base pair (phase 1 and phase 2, respectively). Intron positions that are apart by one or more than one base pair were considered as not identical even if it is not excluded that these introns may have the same ancestor. The phases of the two introns surrounding an exon define the exon type. Exon types are either symmetrical when flanked by introns of identical phases, 0–0, 1–1, and 2–2, or asymmetrical when bordered by introns of different phases, for instance 1–0.

Neighbor-Joining Trees

AtRH, CeRH, and DmRH neighbor-joining trees were constructed from CLUSTALW alignments of the RH catalytic domain and consisted of 1000 trials with bootstrap. Exclusions for positions with gaps and corrections for multiple sequences were both set to off.

ACKNOWLEDGMENTS

We thank O. Richard and I. Gy for helpful discussions. We are grateful to B. Chalhoub for comments on the manuscript. N.B. was supported by a doctoral fellowship from the French Ministère de l'Education Nationale, de la Recherche et de la Technologie.

NOTE ADDED IN PROOF

Two new genes were characterized since the search of family members was performed: *AtRH56*, accession no. AL360314 at position 29936–33720 and *AtRH57*, accession no. AC016661 at position 78858–81927. *AtRH56* is closely related to *AtRH15*.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

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

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815.

Aubourg, S., Kreis, M., and Lecharny, A. 1999. The DEAD box RNA helicase family in *Arabidopsis thaliana*. *Nucleic Acids Res.* **27:** 628–636.

Bagavathi, S. and Malathi, R. 1996. Introns and protein revolution – an analysis of the exon/intron organisation of actin genes. *FEBS Lett.* **392:** 63–65.

Baldauf, S.L., Roger, A.J., Wenk-Siefert, I., and Doolittle, W.F. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290:** 972–977.

Barta, I. and Iggo, R. 1995. Autoregulation of expression of the yeast Dbp2p 'DEAD-box' protein is mediated by sequences in the conserved DBP2 intron. *EMBO J.* **14:** 3800–3808.

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

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*. A platform for investigating biology. *Science* **282:** 2012–2027.

Cavalier-Smith, T. 1985. Selfish DNA and the origin of introns. *Nature* **315:** 283–284.

Charlesworth, D., Liu, F.L., and Zhang, L. 1998. The evolution of the alcohol dehydrogenase gene family by loss of introns in plants of the genus *Laevenworthia* (Brassicaceae). *Mol. Biol. Evol.* **15:** 552–559.

Cho, G. and Doolittle, R.F. 1997. Intron distribution in ancient paralogs supports random insertion and not random loss. *J. Mol. Evol.* **44:** 573–584.

Clegg, M.T., Cummings, M.P., and Durbin, M.L. 1997. The evolution of plant nuclear genes. *Proc. Natl. Acad. Sci.* **94:** 7791–7798.

Deutsch, M. and Long, M. 1999. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.* **27:** 3219–3228.

Doolittle, R.F. 1994. Convergent evolution: The need to be explicit. *Trends Biochem. Sci.* **19:** 15–18.

Doolittle, W.F. 1978. Genes in piece: Were they ever together? *Nature (London)* **272:** 581–582.

Fichant, G.A. 1992. Constraints acting on the exon positions of the splice site sequences and local amino acid composition of the protein. *Hum. Mol. Genet.* **1:** 259–267.

Fink, G.R. 1987. Pseudogenes in yeast? *Cell* **49:** 5–6.

Frugoli, J.A., McPeek, M.A., Thomas, T.L., and McClung, C.R. 1998. Intron loss and gain during evolution of the catalase gene family in angiosperms. *Genetics* **149:** 355–365.

Gilbert, W. 1987. The exon theory of genes. *Cold Spring Harbor Symp. Quant. Biol.* **52:** 901–905.

Gilbert, W., de Souza, S.J., and Long, M. 1997. Origin of genes. *Proc. Natl. Acad. Sci.* **94:** 7698–7703.

Giroux, M.J., Clancy, M., Baier, J., Ingham, L., McCarty, D., and Hannah, L.C. 1994. De novo synthesis of an intron by the maize transposable element Dissociation. *Proc. Natl. Acad. Sci.* **91:** 12150–12154.

Gogarten, J.P. and Olendzenski, L. 1999. Orthologs, paralogs and genome comparisons. *Curr. Opin. Genet. Dev.* **9:** 630–636.

Goodman, H.M., Ecker, J.R., and Dean, C. 1995. The genome of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci.* **92:** 10831–10835.

Gorbalenya, A.E. and Koonin, E.V. 1993. Helicases: Amino acid sequence comparisons and structure-function relationship. *Curr. Opin. Struct. Biol.* **3:** 419–429.

Gotoh, O. 1998. Divergent structures of *Caenorhabditis elegans* cytochrome P450 genes suggest the frequent loss and gain of introns during the evolution of Nematodes. *Mol. Biol. Evol.* **15:** 1447–1459.

Hardison, R.S. 1996. A brief history of hemoglobins : Plant, animal, protist, and bacteria. *Proc. Natl. Acad. Sci.* **93:** 5675–5679.

Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbretch, J., Rouze, P., and Brunak, S. 1996. Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.* **24:** 3439–3452.

Jellie, A.M., Tate, W.P., and Trotman, C.N. 1996. Evolutionary history of introns in a multidomain globin gene. *J. Mol. Evol.* **42:** 641–647.

Kirik, A., Salomon, S., and Puchta, H. 2000. Species-specific double-strand break repair and genome evolution in plants. *EMBO J.* **19:** 5562–5566.

Koch, M.A., Haubold, B., and Mitchell-Olds, T. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Mol. Biol. Evol.* **17:** 1483–1498.

Lewin, R. 1983. How mammalian RNA returns to its genome. *Science* **219:** 1052–1054.

Liaud, M-F., Brinkmann, H., and Cerff, R. 1992. The B-tubulin gene family of pea: Primary structures, genomic organization and intron-dependent evolution of genes. *Plant. Mol. Biol.* **18:** 639–651.

Linder, P. 2000. DEAD-box proteins. *Curr. Biol.* **10:** R887.

Logsdon, Jr., J.M. and Palmer, J.D. 1994. Origin of introns—early or late? *Nature* **369:** 526–527.

Long, M., de Souza, S.J., Rosenberg, C., and Gilbert, W. 1998. Relationship between "proto-splice sites" and intron phases: evidence from dicodon analysis. *Proc. Natl. Acad. Sci.* **95:** 219–223.

Lukashin, A.V. and Borodovsky, M. 1998. GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Res.* **26:** 1107–1115.

Martinez, P., Martin, W., and Cerff, R. 1989. Structure, evolution and anaerobic regulation of a nuclear gene encoding cytosolic glyceraldehyde-3-phosphate dehydrogenase from maize. *J. Mol. Biol.* **208:** 551–565.

Nouaud, D., Boëda, B., Levy, L., and Anxolabehere, D. 1999. A P element has induced intron formation in *Drosophila*. *Mol. Biol. Evol.* **16:** 1503–1510.

O'Neill, R.J., Brennan, F.E., Delbridge, M.L., Crozier, R.H., and Graves, J.A.M. 1998. De novo insertion of an intron into the mammalian sex determining gene, SRY. *Proc. Natl. Acad. Sci.* **95:** 1653–1657.

Paquette, S.M., Bak, S., and Feyereisen, R. 2000. Intron-exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of *Arabidopsis thaliana*. *DNA Cell Biol.* **19:** 307–317.

Palmer, J.D. and Logsdon, J.M. 1991. The recent origins of introns. *Curr . Opin. Genet. Dev.* **1:** 470–477.

Patthy, L. 1996. Exon shuffling and other ways of module exchange. *Matrix Biol.* **15:** 301–310.

Pavy, N., Rombauts, S., Dehais, P., Mathe, C., Ramana, D.V., Leroy, P., and Rouze, P. 1999. Evaluation of gene prediction software using a genomic data set: Application to *Arabidopsis thaliana* sequences. *Bioinformatics* **15:** 887–899.

Petrov, D., Sangster, T.A., Johnston, J.S., Hartl, D.L., and Shaw, K.L. 2000. Evidence for DNA loss as determinant of genome size. *Science* **287:** 1060–1062.

Robertson, H.M. 1998. Two large families of chemoreceptor genes in the Nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. *Genome Res.* **8:** 449–463.

Rogozin, I.B., Lyons-Weiler, J., and Koonin, E.V. 2000. Intron sliding in conserved gene families. *Trends Genet.* **16:** 430–432.

Rzhetsky, A., Ayala, F.J., Hsu, L.C., Chang, C., and Yoshida, A. 1997. Exon/intron structure of aldehyde dehydrogenase genes supports the "intron-late" theory. *Proc. Natl. Acad. Sci.* **94:** 6820–6825.

Sanderfoot, A.A., Assaad, F.F., and Raikhel, N.V. 2000. The *Arabidopsis* genome. An abundance of soluble N-ethylmaleimide-sensitive factor adaptor protein receptors. *Plant Physiol.* **124:** 1558–1569.

Schmid, S.R. and Linder, P. 1992. D-E-A-D protein family of putative RNA helicases. *Mol. Microbiol.* **6:** 283–292.

Schmucker, D., Vorbruggen, G., Yeghiayan, P., Fan, H.Q., Jackle, H., and Gaul, U. 2000. The *Drosophila* gene abstrakt, required for visual system development, encodes a putative RNA helicase of the DEAD box protein family. *Mech. Dev.* **91:** 189–196.

Stoltzfus, A., Spencer, D.F., Zuker, M., Logsdon, Jr., J.M., and Doolittle, W.F. 1994 Testing the exon theory of genes: The evidence from protein structure. *Science* **265:** 202–207.

Stoltzfus, A., Logsdon, Jr., J.M., Palmer, J.D., and Doolittle, W.F. 1997. Intron "sliding" and the diversity of intron positions. *Proc. Natl. Acad. Sci.* **94:** 10739–10744.

Tarrio, R., Rodriguez-Trelles, F., and Ayala, F.J. 1998. New *Drosophila* introns originate by duplication. *Proc. Natl. Acad. Sci.* **95:** 1658–1662.

Tavares, R., Aubourg, S., Lecharny, A., and Kreis, M. 2000. Organization and structural evolution of four multigene families in *Arabidopsis thaliana* : AtLCAD, AtLGT, AtMYST and AtHD-GL2. *Plant Mol. Biol.* **42:** 703–717.

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

Tognolli, M., Overney, S., Penel, C., Greppin, H., and Simon, P. 2000. A genetic and enzymatic survey of *Arabidopsis thaliana* peroxidases. *Plant Perox. Newslett.* **14:** 3–12.

Trotman, C.N.A. 1998. Introns-early: Slipping lately? *Trends Genet.* **14:** 132–134.

Received June 14, 2001; accepted in revised form September 12, 2001.