628–636 Nucleic Acids Research, 1999, Vol. 27, No. 2

The DEAD box RNA helicase family in *Arabidopsis thaliana*

Sébastien Aubourg, Martin Kreis and Alain Lecharny*

Institut de Biotechnologie des Plantes, Laboratoire de Biologie du Développement des Plantes, Bâtiment 630, Université de Paris-Sud–ERS/CNRS 569, F-91405 Orsay Cedex, France

Received August 19, 1998; Revised and Accepted November 25, 1998

DDBJ/EMBL/GenBank accession nos+

ABSTRACT

The numerous genomic sequences and ESTs released by the *Arabidopsis thaliana* Genome Initiative (AGI) have allowed a systematic and functional study of the DEAD box RNA helicase family. Sequencing and *in silico* analysis led to the characterization of 28 novel *A.thaliana* DEAD box RNA helicases forming a family of 32 members, named AtRH. Fourteen *AtRH* genes with an unexpected heterogeneous mosaic structure are described and compared bringing new information about the genesis of the gene family. The mapping of the *AtRH* genes shows their repartition on the five chromosomes without clustering and therefore *AtRH*s have been estimated to 60 genes per *A.thaliana* haploid genome. Sequence comparisons revealed a very conserved catalytic central domain flanked or not by four classes of extensions in the N- and/or C- extremities. The global amino acid composition of the extensions are tentatively correlated to specific functions such as targeting, protein interaction or RNA binding. The expression of the 32 *AtRH* genes has been recorded in different tissues. Separate patterns of expression and alternative polyadenylation sites have been shown. Based on the integration of all this information, we propose a classification of the AtRH proteins into subfamilies with associated functions.

INTRODUCTION

A large number of genetic processes demand the unwinding from double-stranded or base-paired regions of DNA/DNA, RNA/RNA or RNA/DNA hybrids to single-stranded polynucleotides. These complex reactions require the intervention of several types of proteins including helicases (1).

Despite the diversity of their biological functions and the wide range of organisms in which these proteins have been identified, a high sequence conservation has been maintained in the large group of helicases, suggesting that all the helicase genes evolved from a common ancestor. Hence, signature sequences can be used efficiently for the detection and the prediction of new helicases in the genome databases. A sequence based classification has led to the definition of three superfamilies of helicases, namely SF1, SF2 and SF3 (2). To date, SF2 is the best characterized superfamily which includes the protein families SNF2 and the DEAH and DEAD box helicases. Each protein contains eight conserved motifs named I, Ia, Ib and from II to VI (3). These conserved motifs contain the amino acid residues most important for the function of a helicase, specifically those involved in catalysis and in substrate binding.

The DEAD box RNA helicase family has been defined by Linder *et al.* (4) and named according to the highly conserved residues, Asp-Glu-Ala-Asp, in motif II. The eukaryotic initiation factor eIF-4A is the prototype and the best biochemically characterized member of the family (5). Although a large number of DEAD box proteins has been identified as 'putative computer-predicted helicases', for only a few of them (e.g. human p68, mouse eIF-4A, *Xenopus* An3 and xp54, *Drosophila* VASA and *Arabidopsis thaliana* DRH1) the ATP-dependent RNA helicase activity has been demonstrated *in vitro*. Despite their shared biochemical function (i.e. RNA unwinding) the DEAD box helicases are involved in a number of different molecular mechanisms such as RNA splicing, ribosome assembly and initiation of translation. They are also important cellular factors for regulatory events, in particular during organ maturation and cellular growth and differentiation (1,6).

Even if the eight characteristic motifs are well conserved between all the helicases, the DEAD box proteins may be specifically sorted out using peculiarities in their motifs. Figure 1 presents the residues best conserved in the eight motifs of the DEAD family. Their involvement in the biochemical functions and their interactions with substrates have been demonstrated by site-directed mutagenesis (5,7). The BLOCKS database has also defined blocks characteristic of the DEAD box family (8) and corresponding to the best conserved regions (BL00039A–F). The relationship between motifs and blocks is shown in Figure 1. Recently, X-ray crystallographic studies suggested that the different conserved helicase motifs are closely associated in the tertiary structure of the protein and that they may form a large functional domain rather than seven individual ones with strictly independent functions (9,10). Despite sequence similarities in common regions between the DEAH and the DEAD box helicases, these two families are functionally different. In the PROSITE database, different signatures have been defined for the two families (11). The DEAH members show conserved motifs not present in DEAD box proteins, both differ markedly in blocks B, C and E and DEAH box helicases are all significantly larger

*To whom correspondence should be addressed. Tel: +33 1 69 33 63 93; Fax: +33 1 69 33 64 25; Email: lecharny@ibp.u-psud.fr

+AJ010456–AJ010477 and AJ012745

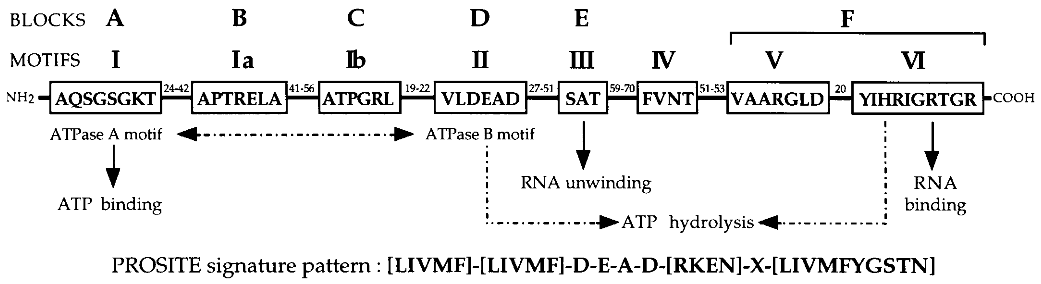


Figure 1. Organization and involvement in the catalytic functions of the best conserved residues of the DEAD box RNA helicases. The relationships between the eight characteristic motifs of a helicase (3) and the six blocks defined in the BLOCKS database (8) for the DEAD box RNA helicase family are indicated. The known physical interactions between the motifs in the tertiary structure are represented by dotted lines. The numbers indicated between the motifs are the typical range of amino acid residues. In the N- and C-termini, extension lengths respectively before the first motif and after the last one are from 20 to 400 amino acids.

than the DEAD ones. Furthermore, the mutation from Asp to His in a DEAD protein reduced strongly the helicase activity (7).

The complete genomes of prokaryotes contain from one to five DEAD box helicases and in the genome of *Saccharomyces cerevisiae*, 26 different DEAD box proteins have been found (12). All the published data on plant DEAD box RNA helicases concern *eIF-4A* genes from *Oryza sativa*, *Triticum aestivum*, *Nicotiana plumbaginifolia* and the *PRH75* gene from *Spinacia oleracea*. Ten different eIF-4A cDNAs have also been characterized in *Nicotiana tabacum* (13) and *sylvestris* (14). In *A.thaliana*, only five DEAD box RNA helicases have been studied including the two highly similar eIF-4A factors (15), PRH75 (16), AtRH1 (17) and AtDRH1 for which the helicase activity has recently been proven (18). Furthermore, a preliminary study using a recurrent and complete search of ESTs together with an assembly of overlapping tag sequences indicated that *A.thaliana* has a minimum of 18 different expressed DEAD box genes (17).

The *Arabidopsis* Genome Initiative (AGI) aims, through an international effort at sequencing the five chromosomes (120 Mb) of *A.thaliana*. To date, 25% of the genomic sequence and >36 000 ESTs (19–21) have been released in the databases. All these sequences are a highly valuable source of information and we have extensively screened the *A.thaliana* database (A*t*DB: http://genome-www.stanford.edu) to find all the sequences having significant similarities with DEAD box RNA helicases. Analyses of the sequences available allowed the characterization of 28 novel DEAD box RNA helicases. Sequence comparisons associated with expression study and mapping of 32 *A.thaliana* DEAD box genes provided a new insight into the organization, the evolution and the functions of the DEAD box family in plants.

MATERIALS AND METHODS

Screening of databases and sequence analyses

Different computer programmes have been used in the search and the analyses of the transcript, genomic and protein sequences. The extensive screening of databases (i.e. dbEST, HTGS, GenBank/EMBL/DDBJ) has been done using the different BLAST algorithms (22) to detect similarities with known DEAD box RNA helicase sequences. The ESTs detected have been aligned in several groups as described in Aubourg *et al.* (17). The corresponding clones have been sequenced by Dye Terminator reactions (Perkin-Elmer/Applied Biosystems) after DNA preparation using the QIAwell Plus Plasmid Kit (Qiagen). The alignments of overlapping sequences and the consensus sequences have been obtained using the GDE software (Genetic Data Environment). The putative splicing sites and the potential coding regions in anonymous genomic sequences were predicted by the NETPLANTGENE software (23) especially realized for *A.thaliana*. The search of DEAD blocks in the deduced amino acid sequences were done by BLOCKS SEARCHER (24). Protein alignments and relationship trees were realized by the CLUSTAL W 1.5 programme (25) using the neighbor-joining distance method (26) conjugated to a bootstrap analysis of 100 replicates (27). A parsimony analysis was conducted using the heuristic search algorithm of PAUP 3.1 (28). The PSORT programme (29), for which an algorithm specific to plants is available, allowed the prediction of the subcellular targeting.

Expression study

Expression patterns were studied using a PCR-based method (30,31). For each predicted gene or cDNA, a set of specific oligonucleotides was chosen and PCR amplifications were carried out using as a template 15 ng of DNA, extracted from nine different cDNA libraries including Columbia dry seeds (Raynal, Perpignan, France), Columbia roots cultured in NO_3^- liquid medium (Forde and Zhang, Rothamsted, UK), roots of 13-day-old Columbia plantlets grown in liquid MS medium (Salanoubat, Gif/Yvette, France), 2-week-old Columbia GH50 leaves (32), Landsberg erecta inflorescences containing flower buds (33), Columbia green siliques (34), pollen (Twell, Leicester, UK), 3-day-old Columbia hypocotyl seedlings (35) and 7-day-old etiolated seedlings (Desprez, Versailles, France). PCR experiments were carried out in a reaction volume of 25 µl containing 0.7 U of *Taq* polymerase (Qiagen) in its associated reaction buffer, 0.1 µM of each primer (Eurogenetec) and 50 µM of each dNTP (Pharmacia). PCR conditions were 5 min at 94°C followed by 40 cycles each of 30 s denaturation at 94°C, 1 min annealing at 55°C and 2 min extension at 72°C. Temperatures ranging from 45 to 60°C during annealing have been tried for each studied gene. The PCR products were sequenced using Dye Terminator reactions after purification with the QIAquick PCR purification kit (Qiagen).

Mapping

Two methods have been used for the mapping of the DEAD box RNA helicase genes. Gene sequences originating from BAC sequences have been mapped using the *A.thaliana* database (A*t*DB). Every BAC sequence was positioned with markers of the genetic map of Lister and Dean (36). Chromosome localizations

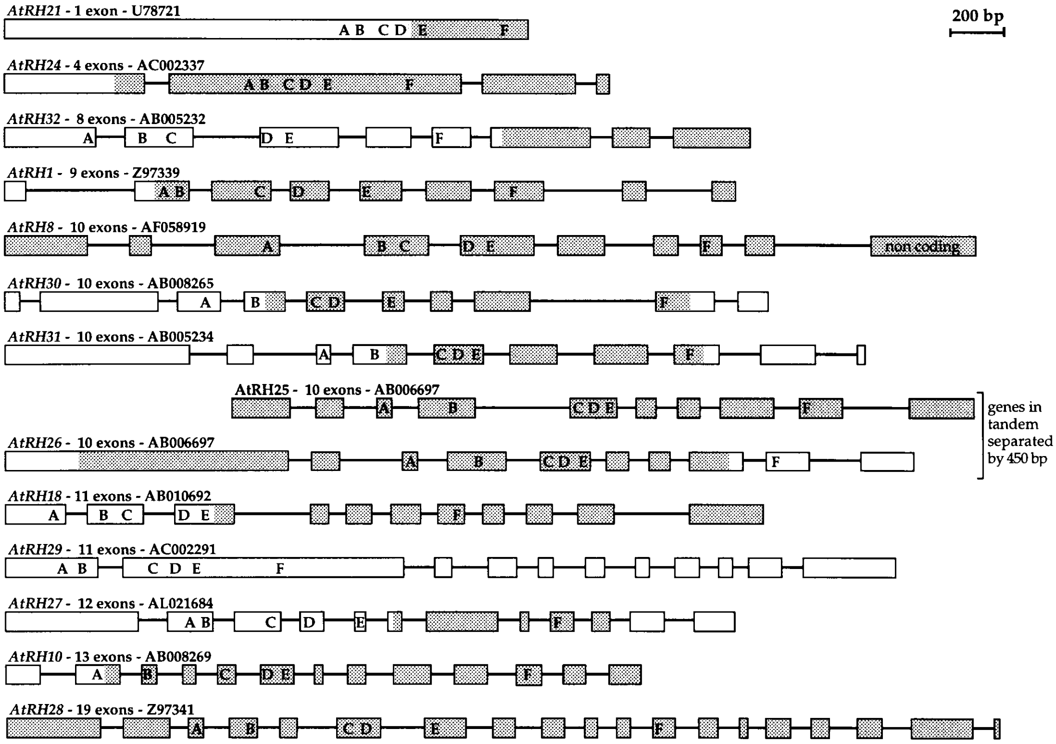

AtRH21 - 1 exon - U78721
A B C D E F
200 bp
AtRH24 - 4 exons - AC002337
A B C D E F
AtRH32 - 8 exons - AB005232
A B C D E F
AtRH1 - 9 exons - Z97339
A B C D E F
AtRH8 - 10 exons - AF058919
A B C D E F
non coding
AtRH30 - 10 exons - AB008265
A B C D E F
AtRH31 - 10 exons - AB005234
A B C D E F
AtRH25 - 10 exons - AB006697
A B C D E F
genes in tandem separated by 450 bp
AtRH26 - 10 exons - AB006697
A B C D E F
AtRH18 - 11 exons - AB010692
A B C D E F
AtRH29 - 11 exons - AC002291
A B C D E F
AtRH27 - 12 exons - AL021684
A B C D E F
AtRH10 - 13 exons - AB008269
A B C D E F
AtRH28 - 19 exons - Z97341
A B C D E F

Figure 2. Schematic representation of the intron–exon structure of 14 genes encoding DEAD box RNA helicases. Exons and introns are respectively represented by boxes and lines. Gene structures have been deduced from the NETPLANTGENE (23) analysis. The gray shaded regions indicate the parts of the genes for which cognate cDNA and/or PCR products have been sequenced to confirm computer predictions. Only the coding regions of the genes are represented except in the case of gene *AtRH8* which has a 3′ non-coding exon. Following the name of a gene, the number of exons and the accession number of the genomic sequence are indicated. Approximate positions of the regions encoding the six blocks typical of the DEAD box proteins (A–F) are shown for reference marks. The scale is respected.

of the different EST clones were obtained using PCR, specific primers and the CIC YAC library (37) as a template. The PCR conditions used are as for the expression study.

RESULTS

Characterization and organization of the AtRH family

Five cDNAs (eIF-4A1, eIF-4A2, AtDRH1, pRH75 and AtRH1), 82 ESTs, 13 BAC (Bacterial Artificial Chromosome of genomic sequence), the ESSA1 contig (38) and eight BAC end sequences with high similarities with DEAD box proteins have been detected in the *A.thaliana* database. All the EST clones have been fully sequenced. After a thorough analysis of all the sequence data, 32 different DEAD box RNA helicases have been obtained and named from AtRH1 to AtRH32. The previously described eIF-4A1, eIF-4A2 (15), PRH75 (16) and AtDRH1 (18) cDNAs correspond respectively to the genes *AtRH4*, *AtRH19*, *AtRH7* and *AtRH14*.

NETPLANTGENE predictions on the genomic sequences and their comparison with the full length or partial cDNA sequences provided the gene structure of 14 *AtRH* genes (Fig. 2). The mosaic structure is different for each member of the *AtRH* family. Not only are position and length of the introns not conserved, their number also is highly variable, from 18 introns in *AtRH28* to none for *AtRH21*. Furthermore, there is no correlation between the six blocks (or eight motifs) of the catalytic domain and the exons since there are examples where motifs are interrupted by introns.

The 14 genes retrieved from the genomic sequences available from AGI and the 18 for which we characterized a cognate cDNA have been mapped on the five chromosomes of the *A.thaliana* genome (Fig. 3). For *AtRH2*, *5*, *6*, *13*, *14*, *22* and *23* no YAC or only YACs without associated markers have been detected in the CIC library. For the mapped genes, there is no evidence of *AtRH* gene clusters except for the two genes *AtRH25* and *26* which are in tandem and separated by only 450 bp. The uneven distribution of the above 14 genes is correlated to the progress of the systematic sequencing available at A*t*DB. These data, together with the results of the mapping of ESTs, strongly indicate that the genes encoding DEAD box RNA helicases have an even distribution on the five chromosomes. At the present time, 25% of the coding genome of *A.thaliana* has been sequenced (29 Mb, June 1998), and 14 *AtRH* genes have been discovered in this significant and rather well distributed part of the genome. Thus the number of *AtRH* may be estimated to be 60 by haploid genome (120/29 × 14 = 58).

The 32 deduced protein sequences, including 11 partial sequences from clones truncated in their 5′ extremity, have been aligned using CLUSTAL W (Fig. 4). The core domain, composed of the six conserved blocks with catalytic functions, is well conserved between the members of the AtRH family. On the other hand, the N- and C-terminal sequences are very variable in length as well as in composition. The proteins have or do not have extension at the N- and C-extremities. The common characteristic of the extensions is their high hydrophilic nature. These sequences may be sorted out into four classes in function of their global amino acid composition (Fig. 4). When new sequence

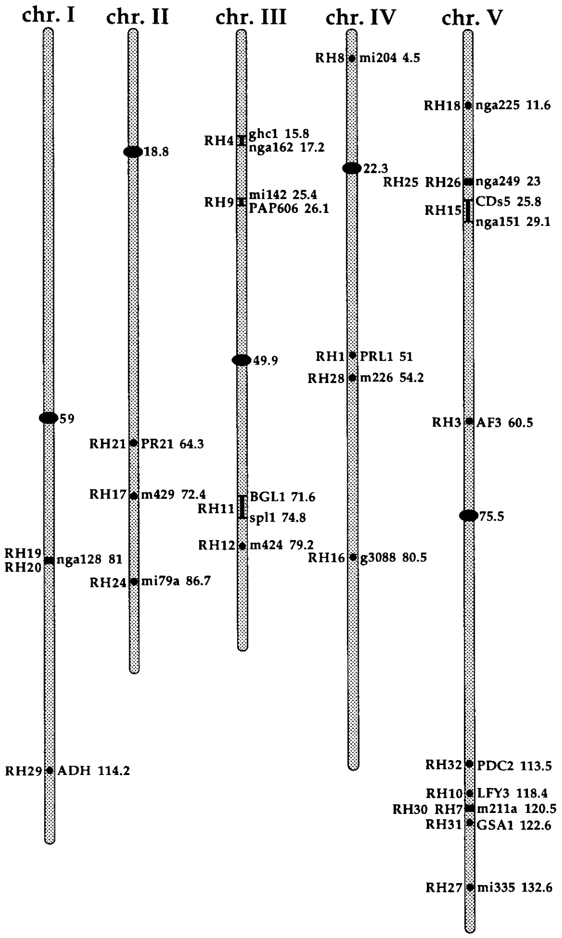


Figure 3. Mapping of the *AtRH* genes on the five chromosomes of *A.thaliana*. Each gene is associated with one or two markers of the genetic map of Lister and Dean. The centromere regions (73) are shown by large black dots. The position of the markers and the centromeres is indicated in cM (according to A*t*DB). The representation is drawn to scale.

extremities are determined, it is possible that other classes of extension will emerge.

AtRH8 and AtRH12 have, in their N-extremity, a short (30–40 amino acids) and hydrophilic extension (Q) composed of 50–60% of Gln. AtRH21, 26 and 31 exhibit an Arg-Ser-Asp-rich N-terminal extension (RSD). The amino acids Asn, Glu and Gly may also be found in high proportion in this second class of extension. The 320 amino acid long extension of AtRH26 contains seven intern repeat sequences (SGSSFRGRxDRNVD with x as S or N). The third class of extension (KE) is mainly composed of Lys and Glu (i.e. from 20 to 70%) with occasional Arg. This extension can be N- (AtRH7, 21 and 27) or C-terminal (AtRH10, 17, 18, 28 and 32). The last class of extension (GRS) was shown to contain 50–80% of Gly, Arg and Ser and is located in the C-extremity (AtRH3, 7, 9, 11, 14, 29 and 30) except for AtRH30 which has also an N-terminal extension.

The core domain of AtRH, composed of eight conserved motifs providing the catalytic function, is very conserved (Figs 4 and 5). The distances between the various blocks are as expected in the typical range (compare with Fig. 1) excepted for the proteins AtRH1, AtRH13 and AtRH21 which have large insertions, not conserved, between the blocks D and E. A typical consensus sequence of the DEAD box RNA helicase has been obtained from sequence alignments of the blocks B, C, D and F (Fig. 5). The protein AtRH22 is very different from the other members of the family. Indeed, AtRH22 does not have the canonical blocks B and C and the motif SAT (block E) is not present despite its known pivotal function in RNA unwinding.

Expression of the *AtRH* genes

Two different and complementary approaches have been used to study the expression of the *AtRH* genes. The first one is based on the screening of dbEST since the transcription level of the genes is thought to be roughly monitored by counting the numbers of matched *A.thaliana* ESTs (39). Indeed, the 36 000 ESTs released (19–21) are a good representation of the highly transcribed genes. The number of cognate ESTs for each *AtRH* gene is reported in the graphic of the Figure 6. One gene, *AtRH4*, matched with 21 different ESTs and is certainly highly expressed. Three other genes (i.e. *AtRH2*, *AtRH3* and *AtRH7*) have more than five cognate ESTs. The analysis of the ESTs gives further information, namely the wide heterogeneity of polyadenylation patterns in mRNAs. The 3′ extremity of the different EST clones cognate to each gene showed up to eight different polyadenylation sites for the gene *AtRH4* (Fig. 6).

The second method used to analyze the expression is the demonstration of the presence of transcripts in different tissues of *A.thaliana*. Using a PCR-based method, the expression of each *AtRH* has been checked in nine different cDNA libraries (Fig. 6). Results show different patterns of expression in function of the genes. For example, the gene *AtRH4* is expressed in all the tissues and conditions tested. This result is in agreement with indications given by the number of EST and confirms the high level of expression of this gene. This method of detection of the expression has the advantage to be fully specific since the amplification products obtained are controlled by sequencing. However, the tissue-specific expression suggested for a small number of *AtRH* genes (Fig. 6), needs to be confirmed using northern blot experiment and/or *in situ* hybridization.

DISCUSSION

The genome sequencing projects do lead to the characterization of large gene families without the drawback of a molecular approach such as the screening of cDNA or genomic libraries using homologous or heterologous probes. Indeed, using these approaches only closely homologous genes with high nucleotidic similarities may be isolated. The AGI (40,41) opened the way to a new and unbiased insight into the organization of genes in plant genomes.

In plants, the family of the ATP-dependent DEAD box RNA helicases is poorly known. With the exception of the eIF-4A subfamily studied in tobacco (14,42), nothing is known about the size, the organization, the expression or the functions of the RNA helicase family. The 32 *A.thaliana* DEAD box RNA helicases, including 28 novel proteins described in this paper bring new data and features about this important family involved in various processes concerning RNA.

The intron–exon structure of the *AtRH* genes

Not all the intron–exon boundaries have been confirmed by the sequencing of the corresponding cDNA. Nevertheless, based on

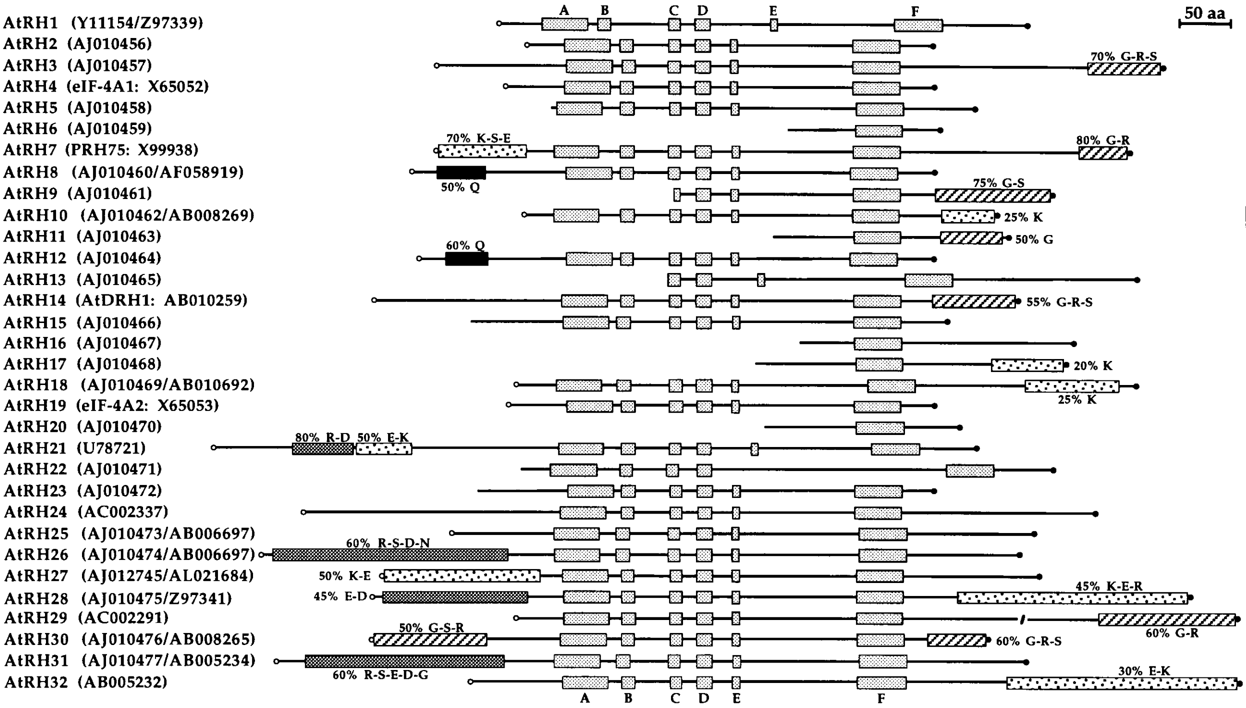


Figure 4. Schematic alignment of 32 DEAD box RNA helicases. The six conserved blocks of the core domains are represented by light grey boxes (A–F). The N- and C-terminal extensions with distinct amino acid content are indicated by boxes with four different patterns: black (Q rich), dark gray (RSD rich), stippled (KE rich) and hatched (GRS rich). Other regions of the sequences, drawn to scale, are represented by lines. If present, the first methionine is represented by a white dot and the last amino acid by a black dot. Accession numbers are specified for each protein.

NETPLANTGENE results together with the high level of conservation between DEAD box proteins and an additional human valuation, the structure of the *AtRH* genes may be predicted with a high degree of confidence. Indeed, when both a gene and its cognate cDNA were available, our interpretation of the NETPLANTGENE output was almost always confirmed by the actual structure.

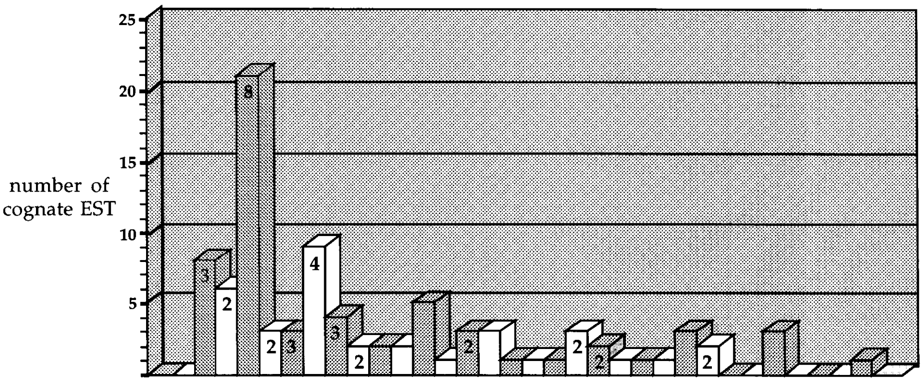
The 14 characterized genes have from 1 to 19 exons; half of them contain 10 or 11 exons (Fig. 2). Each of the identified *AtRH* genes has its unique mosaic structure. At first sight, this result seems contradictory with the relatively high conservation between the protein sequences, strongly suggesting that all the *AtRH* genes are paralogues. The number of *A.thaliana* gene families with a known intron–exon structure is not very large. Generally, members of known gene families exhibit a rather well conserved gene structure (43–45). To the authors' knowledge, the AtRH family is the first one for which a completely non-conserved gene structure between different members is described, based on a high enough number of genes to be significant. There were probably no introns in the putative ancestor gene of the *AtRH* family since none of the present genes has any intron position conserved. Furthermore, the catalytic motifs are occasionally interrupted by an intron and, thus, the insertion of introns is not correlated with protein structure. The exception to the non-conserved *AtRH* structure is observed with the two genes *AtRH25* and *AtRH26* which exhibit a very similar mosaic structure (Fig. 2). Both genes have 10 exons with about the same length except for the first exon. Such an identical feature, contrasting with the general core, could be due to a recent duplication. This explanation is reinforced by the fact that the *AtRH25* and *26* genes have highly similar sequences (92% similarity in overlapping regions) and that they are organized in tandem and only separated by 450 bp.

DEAD box RNA helicases constitute a large gene family in *A.thaliana*

About half of the estimated DEAD box RNA helicases of *A.thaliana* were found by dbEST screening. The other genes are probably expressed at low levels or in specific conditions. Only very few other families with more than 50 genes have been described in plants. The actin gene family (43), the MYB related transcription factors (46), the AtDYW/SNA family (47), the cytochrome P450 family and the cytoplasmic ribosomal family (48) have been estimated to contain more than 100 members and confirmed that gene duplication has been an important factor in the formation of the *A.thaliana* genome (49,50). In comparison with the 60 estimated *A.thaliana* DEAD box RNA helicases, the full sequenced prokaryotic genomes of *Synechocystis* sp., *Mycoplasma pneumoniae*, *Haemophilus influenzae* and *Escherichia coli* contain 1, 2, 3 and 5 DEAD box proteins, respectively (TIGR database, 51). The only eukaryote genome completely sequenced, *S.cerevisiae*, has a DEAD box family of 26 members (12). The transition prokaryotes–eukaryotes was thus correlated with a large expansion of the DEAD box RNA helicase family, probably to carry out the novel molecular processes involving RNA and to cope with the different cellular compartments. The apparition of multicellular organisms has required also a new set of more specialized RNA helicases to deal with new appearing processes such as intron splicing.

	B	C	D	F
AtRH1	LVVLPTRDLALQV	DILVATPGRLMD	LVVDETDRLLREAY	ALTRGMDVKGVTNVINYDMPPFAKTFIHRAGRTARAGQAGRCFTL
AtRH2	LILSPTRELATQT	HVVSGTPGRVCD	LILDESDEMLSRGF	VWARGIDVQQVSLVINYDLPNNRELYIHRIGRSGRFGRKGVAINF
AtRH3	LVLAPTRELAKQV	DVVVGTPGRIID	LVLDEADQMLAVGF	VASRGLDIPNVDLVIHYELPNDPETFVHRSGRTGRAGKEGSAILM
AtRH4	LVLAPTRELAQQI	HVVVGTPGRVFD	FVLDEADEMLSRGF	LLARGIDVQQVSLVINFDLPTQPENYLHRIGRSGRFGRKGVAINF
AtRH5	LVLSPTRELAVQI	DIVIGTPGRLRD	VVLDEADRMLDMGF	VAARGLDIPDVEVVINYTFPLTTEDYVHRIGRTGRAGKKGVAHTF
AtRH6				LFTRGIDIQAVNVVINFDFPRTSESYLHRVGRSGRFGHLGLAVNL
AtRH7	LVLLPTRELAKQV	DIVVGTPGRIKD	RVLDEADEMLRMGF	VAARGLDINDVQLIIQCEPPREVEAYIHRSGRTGRAGNTGVPVTL
AtRH8	VIIVPTRELALQT	HLLVGTPGRILD	LVMDEADKLLSQEF	LFTRGIDIQAVNVVINFDFPKNAETYLHRVGRSGRFGHLGLAVNL
AtRH9			VVLDEADQMLQVGF	VAARGLDVPNVDLVIHYELPNNTETFVHRTGRTGRAGKKGSAILI
AtRH10	CVLSPTRELAIQI	HVIVATPGRLWD	LVLDEADRLLNEDF	VASRGLDIPSVDVVINYDIPTNSKDYIHRVGRTARAGRSGVGISL
AtRH11				VAARGLDIPHVAHVVNFDLPNDIDDYVHRIGRTGRAGKSGIATAF
AtRH12	MILVPTRELALQT	HLLVGTPGRILD	LVMDEADKLLSAEF	LFTRGIDIQAVNVVINFDFPRTSESYLHRVGRSGRFGHLGLAVNL
AtRH13		EIVVATPGRLWE	FVLDEADRMVERGH	LVARGIDIKNVRTIIHYKLPHSAEVYVHRCGRTARAFADGCSIAL
AtRH14	LVLSPTRELATQI	DIVVATPGRLND	LVLDEADRMLDMGF	VAARGLDVKDIRAVVNYDFPNGVEDYVHRIGRTGRAGATGQAFTF
AtRH15	LVLCHTRELAYQI	HIVVGTPGRVLA	FILDECDKMLESLD	LVGRGIDIERVNIVINYDMPDSADTYLHRVGRAGRFGTKGLAITF
AtRH16				GVVRGIDFKKVHTVINFDMPQSVTGYIHRIGRTGRAYSSGSSVSL
AtRH17				VAARGLDFPKVRCIIQYDCPGEATEYVHRVGRTARIGEKGEALLF
AtRH18	VIISPTRELSTQI	NVLIGTPGRLSD	LILDEADRLLEMGF	VAARGLDIPGIDYVVQYDPPQDPNMFNHRAGRTARLGRQGRAIVF
AtRH19	LVLAPTRELAQQI	HVVVGTPGRVFD	FVLDEADEMLSRGF	LLARGIDVQQVSLVINFDLPTQPENYLHRIGRSGRFGRKGVAINF
AtRH20				VAARGLDVKDVKYVINYDFPGSLEDYVHRIGRTGRAGAKGTAYTF
AtRH21	VVMAPTRELAQQI	EIVIATPGRLID	VVLDEADRMIDMGF	VVGRGIDIPDVAHVINYDMPKHIEMYTHRIGRTGRAGKSGVATSF
AtRH22	LILCPNVMLCEQV	DIIVSTPAALLN	VVFDEADMLLCGSF	AAARGVDVPNVSHVIQADFASSAVDFLHRIGRTARAGQYGTVTSL
AtRH23	LVLAPTRELAQQI	HVVVGTPGRVFD	FVLDEADEMLSRGF	LLARGIDVQQVSLVINFDLPTQPENYLHRIGRSGRFGRKGVAINF
AtRH24	VICAPTRELAHQI	EIVVATPGRLID	LVLDEADRMFDLGF	VAARGLDIKSLKTVVNYDIAKDMDMHVHRIGRTGRAGDRDGVAYT
AtRH25	LVVCPTRELACQA	QILVATPGRLKD	LVLDEADHLLDMGF	VSARGVDYPDVSLVVQMGLPSDREQYIHRLGRTGRKGKEGEGVLL
AtRH26	LVVCPTRELASQA	QILVATPGRLKD	LVLDEADHLLDMGF	VSARGVDYPDVSLVVQMGLPSDREQYIHRLGRTGRKGKEGEGVLL
AtRH27	LVICPTRELAIQS	NLLVATPGRLLD	LVMDEADRILEQNF	VAARGLDFPHVDNIVQYDPPDNPTDYIHRVGRTARGEGAKGKALL
AtRH28	LILTPTRELAVQI	DIVVATPGRMID	LILDEADRLLQTGF	VAARGLDIIGVQTVINYACPREIDSYVHRVGRTARAGREGYAVTF
AtRH29	LILSPTRDLAEQT	DVIIATPGRLMH	VVFDEADSLFGMGF	IAARGIDIPLLDNVINWDFPPRPKIFVHRVGRAARAGRTGCAYSF
AtRH30	LILAPTRELAVQI	EIVIATPGRLID	LVLDEADRMLDMGF	VAARGLDVKDIKCVVNYDFPNTLEDYIHRIGRTGRAGAKGMAFTF
AtRH31	LVICPTRELANQA	QILVATPGRLKD	LVLDEADHLLDMGF	VSARGVDYPDVTLVLQVGLPKDREQYIHRLGRTGRKGKEGEGILL
AtRH32	IIISPTRELAAQT	NILVCAPGRLLQ	LILDEADRVLDSAF	VLARGLDFDKAVDWVVQDCPEDVASYIHRVGRTARFYTQGKSLLF
	###-PTRELA-Q-	-###-TPGR#-D	-VLDEAD-#L--GF	#-ARG#D#--V--VI----P-----Y#HR#GRTGR-G--G-----

Figure 5. Sequence alignment of the conserved blocks B, C, D and F and the consensus sequence of the *A.thaliana* DEAD box RNA helicase family. The alignment has been generated by the CLUSTAL W programme. Bold letters are for 100% conserved amino acids. Each amino acid of the consensus sequence (at the bottom of the figure) is conserved in at least 70% of the proteins. The # symbol is for the hydrophobic amino acid group: Ile, Leu, Met and Val. The GKT motif of the block A (not shown) is 100% conserved. The SAT motif of block E (not shown) is present in all the proteins with two exceptions: in AtRH14, SAT is replaced by TAT and in AtRH22 no sequence similar to SAT has been identified.



AtRH	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
seeds		+		+	+								+				+		+	+	+									+		
NO3⁻ roots		+		+	+		+					+		+			+							+	+					+		
MS roots	+	+		+	+		+	+		+		+	+	+					+	+		+	+	+	+			+		+		
leaves	+	+		+	+		+	+		+		+	+			+	+		+		+	+	+	+	+	+	+			+		+
flower buds	+	+		+	+		+	+		+			+			+	+			+	+	+	+		+		+	+			+	
siliques	+	+		+	+		+	+				+	+	+	+	+	+		+	+	+	+		+	+	+		+		+		
pollen	+	+		+	+		+	+				+		+			+		+	+	+		+	+	+	+				+	+	
hypocotyl seedlings	+	+		+	+		+	+		+		+				+				+	+	+	+	+	+	+					+	
etiolated seedlings	+	+	+	+		+	+	+			+	+		+	+				+		+		+							+	+	

Figure 6. Expression study: number of cognate ESTs and PCR detection of transcripts in nine different cDNA libraries for 32 *AtRH* genes. The number of cognate ESTs found in dbEST is plotted for each *AtRH* gene. The number of different polyadenylation sites characterized using the 3′ sequences of the EST clones is indicated in the columns. The table at the bottom of the figure summarizes the PCR detection of transcripts for each gene. The nature of the cDNA libraries and the PCR conditions used are described in the Materials and Methods section.

AtRH proteins: a core domain with specialized extensions

The subcellular localization of the proteins AtRH7, 10, 18, 21, 27, 28 and 32 is predicted (PSORT) to be the nucleus, with a significative score of 0.94 or 0.98. Figure 4 shows that all these RNA helicases have a KE extension. These extensions contain several possible candidate sequences, close to KKKEK, for nucleus localization. Furthermore, the protein AtRH7/PRH75 has previously been localized in the nucleus by Lorkovic *et al.* (16). These authors showed that the 81 amino acid N-terminal polypeptide (which has 25% of Lys) was sufficient for nuclear targeting of the protein.

The GRS extension might be a variant of the RGG box, a motif involved in RNA binding (52). Lorkovic *et al.* (16) showed that in AtRH7/pRH75, the C-terminal domain (67% of Gly) has a high affinity to RNA. This domain is similar to the RNA-binding domain of RNA-binding proteins localized in the nucleus. The involvement of this class of extension in RNA binding has also been demonstrated in the yeast Spb4 and *E.coli* SrmB and DbpA DEAD box RNA helicases which bind ribosomal RNA (53–55) and in the p68-like RNA helicases (56). In the DEAD box RNA helicases without the GRS extension, the RNA binding could be stabilized by the intervention of the motif VI on two heterodimers as shown with eIF-4A and eIF-4B (57). Furthermore, in the human RNA helicase II/Gu DEAD box RNA helicase, a 76 residue extension rich in Arg and Gly in its C-extremity has an RNA folding activity that introduces an intramolecular secondary structure in single-stranded RNA (58).

There is no precise function yet for the two other classes of extensions (Q and RSD). They could however be involved in protein–protein interactions which have been previously demonstrated for the RNA helicase A (59) and for the human p68 protein (60).

Alternate polyadenylation of AtRH transcripts

An analysis of the 3′ sequences of all the cognate EST clones for each *AtRH* gene has shown a wide heterogeneity in polyadenylation sites in particular for *AtRH4* where eight different sites have been found (Fig. 6). The data obtained from the *AtRH* family indicate that, for one given gene, there are on average of two to three different polyadenylation sites for four sequenced transcripts. The near-upstream elements, involved in the polyadenylation signal (61) are, like in many other plant genes, not present in the *AtRH* genes. These elements, close to AAUAAA, are less conserved or even absent in plants and this might explain why polyadenylation frequently occurs at multiple sites (62). Alternate polyadenyl-ation is an important post-transcriptional regulatory process and recent studies with human ESTs have shown that differential polyadenylation could be tissue specific (63). In *A.thaliana* no information can be deduced about such a tissue specificity, since ESTs come mainly from a library where mRNAs from different tissues have been pooled (20). The *N.plumbaginifolia* 3′ untranslated region of mRNA encoding the chloroplast RNA-binding protein contains 14 distinct polyadenylation sites including one which occurs in an intron located in the 3′ non-coding part of the gene (64). No polyadenylation site has been observed in the intron located in the 3′ non-coding region of *AtRH8* (Fig. 2).

Expression of the *AtRH* genes

The global expression level of the *AtRH* family is relatively high since 82 ESTs have been found in dbEST. In comparison, the *A.thaliana* DYW/SNA family, for which 58 genes have been characterized, matches only with six reported ESTs (47). Nevertheless, the level of expression of the *AtRH* genes is very heterogenous (Fig. 6). The two most expressed genes are *AtRH4/eIF-4A1* and *AtRH7/PRH75* with 21 and 9 cognate ESTs, respectively. Furthermore, their transcripts have been detected in every cDNA library tested (except in seeds for *AtRH7*). The *AtRH4/eIF-4A1* cDNA has 70% identity with the well-studied mouse eIF-4A DEAD box RNA helicase. This protein is a translation initiation factor facilitating attachment of the 40s ribosomal subunit (7,65). This central function in the cell might explain the high and constitutive expression of its putative plant orthologue *AtRH4*.

The other *AtRH* genes have different expression patterns. Several are expressed in most of the tissues tested and seem to play a role in a basic activity of the plant cell. The transcripts of a few genes such as *AtRH6* or *AtRH32* have been detected in only one cDNA library suggesting a specific expression, either tissue dependent (leaves for *AtRH32*) or developmentally controlled (dark growth for *AtRH6*). Some DEAD box RNA helicase genes showing highly specific conditions of expression, have previously been described. For example, the expression of *N.tabacum eIF-4A8* is anther specific and starts at microspore mitosis (42). The *Drosophila Dbp73D* and *vasa* helicase genes are also specifically expressed in the germ line tissue (66).

Evolution of the *AtRH* genes and hypothetical functional subfamilies

Results from amino acid sequence analyses, using both parsimony and neighbor-joining methods, separate *AtRH* genes into subfamilies and give information about the genesis of the *AtRH* family (Fig. 7). Furthermore, similarities between the members of these subfamilies and known DEAD box RNA helicases from other organims suggest possible functions. The presence of other eukaryotic genes very similar to the *A.thaliana* subfamily members, representing probable orthologues, indicates that the functional specialization occurred before plant speciation. Six subfamilies, named I–VI, have been defined from the 25 protein sequences for which the six catalytic blocks were known (Fig. 7). The proteins AtRH1, 10, 22, 28 and 29 are the most divergent. The particular features of AtRH1 have been described previously (17).

The three members of the subfamily I do not have high similarities with functionally characterized helicases.

The subfamily II contains the *eIF-4A*-like genes. The parsimony tree indicates the most probable order of duplication of these five genes. They are probably orthologues of the mouse *eIF-4A* gene. The proteins of this group do not have an extension.

The subfamily III contains the genes *AtRH8* and *AtRH12*. The partial sequence of *AtRH6*, not used for the construction of the trees, exhibits very high similarity with the latter two genes in overlapping regions (block E) and may also belong to subfamily III. It is interesting to note that the AtRH8 and 12 proteins have a Q extension in their N-extremity suggesting that the duplication event occurred after the acquisition of the extension. The three proteins AtRH6, 8 and 12 show high similarity with the human p54 helicase (oncogene RCK) which has a Q extension in its N-extremity (67), and the *Saccharomyces pombe* STE13 DEAD box RNA helicases which has a crucial role for yeast entry into meiosis (68).

The genes *AtRH3* and *AtRH7/PRH75* of subfamily IV are closest to the human *Gu* gene encoding a DEAD box helicase

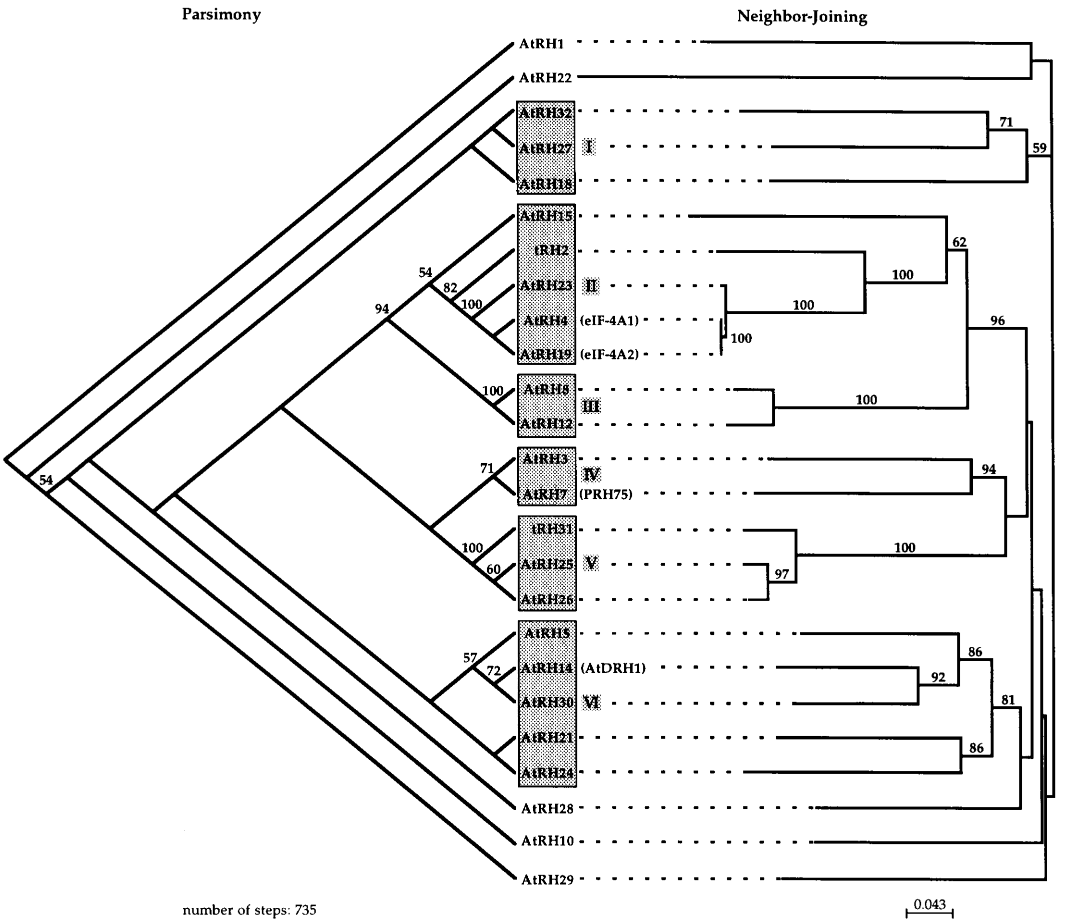


Figure 7. Relationship trees of 25 *A.thaliana* DEAD box RNA helicases. The parsimony and neighbor-joining trees have been generated from an alignment of the six conserved blocks (A–E, i.e. the 131 best conserved residues) using CLUSTAL W and PAUP programmes. Bootstrap values, indicating the number of times a particular node was found in trees generated from 100 replicates of the alignment, are shown on the trees when superior to 50. The proposed subfamilies have been colored in background. The seven partial AtRH proteins for which some blocks are unknown have not been used for the relationship study.

(58). This helicase has a folding activity conferred by the C-terminal RGG extension close to the GRS extension present in AtRH3 and AtRH7.

So far, there is no indication of any function of subfamily V. However, the relationship trees (Fig. 7), the similar gene structure (Fig. 2) and the presence or absence of the RSD extension (Fig. 4) suggest a model for the evolution of the three members of subfamily V. It is probable that the putative ancestor (comprising 10 exons) of the genes *AtRH25*, *26* and *31* already contained the RSD extension. A first duplication led to the *AtRH31* gene and later, a second duplication led to the genes *AtRH25* and *AtRH26*. *AtRH25* has a very reduced first exon compared to *AtRH26* and *31*, and encodes a protein without the RSD extension. This suggests that the loss of this extension occurred during or after the second duplication event. It is interesting to note that this more recent duplication has generated genes organized in tandem (*AtRH25* and *26*), while *AtRH31* is localized at the other extremity of chromosome V (Fig. 3). This situation, where genes in tandem are more similar between them as compared to a third gene localized at another locus, is often encountered. In *A.thaliana*, identical situations have been described in the ubiquitin (69) and nitrilase (70) gene families. These results suggest that physically close genes issued from local duplications tend to remain very similar. Indeed, genes organised in tandem are homogenized both by unequal recombination and gene conversion (71).

The five *AtRH* genes of subfamily VI encode DEAD box RNA helicases which are all very similar to the human and yeast p68. The p68 protein is localized in the nucleoplasm during interphase and translocates to the nucleoli during telophase, suggesting a function in nucleolar assembly (72).

Conclusion

Despite their common RNA unwinding activity and sequence conservation, the DEAD box RNA helicases differ mainly by the addition of N- and C-terminal sequences containing different targeting signals, RNA-binding motifs (sequence specific or not) or regions required for interactions with structural or regulatory proteins. This mechanism of acquisition of different classes of extensions after or before duplication of a core catalytic domain lead to the genesis of a large family with numerous different genes. Furthermore, several mechanisms of regulation both of the level of expression and at the post-transcriptional level explain the wide spectrum of functions involving DEAD box RNA helicases.

ACKNOWLEDGEMENTS

The authors are very grateful to T. Newman, the *Arabidopsis* Biological Resource Center and the stock donor (Ohio State University, USA) for quickly providing the various EST clones. We thank M. Dornelas and R. Tavares for helpful discussions, and I. Gy and F. Chefdor for their excellent sequencing support. M. Krivitsky, D. Bouchez, C. Camilleri and C. Macadre are acknowledged for their help in the mapping work.

REFERENCES

1 Schmid,S.R. and Linder,P. (1992) *Mol. Microbiol*., **6**, 283–292.
2 Gorbalenya,A.E. and Koonin,E.V. (1993) *Curr. Opin. Struct. Biol*., **3**, 419–429.
3 Gorbalenya,A.E., Koonin,E.V., Donchenko,A.P. and Blinov,V.M. (1989) *Nucleic Acids Res*., **17**, 4713–4730.
4 Linder,P., Lasko,P.F., Ashburner,M., Leroy,P., Nielsen,P.J., Nishi,K., Schnier,J. and Slonimski,P.P. (1989) *Nature*, **337**, 121–122.
5 Pause,A., Méthot,N. and Sonenberg,N. (1993) *Mol. Cell. Biol*., **13**, 6789–6798.
6 Stevenson,R.J., Hamilton,S.J., MacCallum,D.E., Hall,P.A. and Fuller-Pace,F.V. (1998) *J. Pathol*., **184**, 351–359.
7 Pause,A. and Sonenberg,N. (1992) *EMBO J*., **11**, 2643–2654.
8 Henikoff,S. and Henikoff,J.G. (1991) *Nucleic Acids Res*., **19**, 6565–6572.
9 Fernandez,A., Guo,H.S., Saenz,P., Simon-Buela,L., Gomez de Cedron,G. and Garcia,J.A. (1997) *Nucleic Acids Res*., **25**, 4474–4480.
10 Yao,N.H., Hesson,T., Cable,M., Hong,Z., Kwong,A.D., Le,H.V. and Weber,P.C. (1997) *Nature Struct. Biol*., **4**, 463–467.
11 Bairoch,A. (1992) *Nucleic Acids Res*., **20**, 2013–2018.
12 Mewes,H.W., Albermann,K., BShr,M., Frishman,D., Gleissner,A., Hani,J., Heumann,K., Kleine,K., Maierl,A., Oliver,S.G., Pfeiffer,P. and Zollner,A. (1997) *Nature*, **387**, 7–65.
13 Owttrim,G.W., Mandel,T., Trachsel,H., Thomas,A.A.M. and Kuhlemeier,C. (1994) *Plant Mol. Biol*., **26**, 1747–1757.
14 Itadani,H., Sugita,M. and Sugiura,M. (1994) *Plant Mol. Biol*., **24**, 249–252.
15 Metz,A.M., Timmer,R.T. and Browning,K.S. (1992) *Gene*, **120**, 313–314.
16 Lorkovic,Z.J., Herrmann,R.G. and Oelmüller,R. (1997) *Mol. Cell. Biol*., **17**, 2257–2265.
17 Aubourg,S., Takvorian,A., Chéron,A., Kreis,M. and Lecharny,A. (1997) *Gene*, **199**, 241–253.
18 Okanami,M., Meshi,T. and Iwabuchi,M. (1998) *Nucleic Acids Res*., **26**, 2638–2643.
19 Höfte,H., Desprez,T., Amselem,J., Chiapello,H., Caboche,M., Moisan,A., Jourjon,M.F., Charpenteau,J.L., Berthomieu,P., Guerrier,D. *et al.* (1993) *The Plant J*., **4**, 1051–1061.
20 Newman,T., de Bruijn,F.J., Green,P., Keegstra,K., Kende,H., McIntosh,L., Ohlrogge,J., Raikhel,N., Somerville,S., Thomashow,M., Retzel,E. and Somerville,C. (1994) *Plant Physiol*., **106**, 1241–1255.
21 Cooke,R., Raynal,M., Laudié,M., Grellet,F., Delseny,M., Morris,P.C., Guerrier,D., Giraudat,J., Quigley,F., Clabault,G. *et al.* (1996) *Plant J*., **9**, 101–124.
22 Altschul,S.F., Stephen,F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol*., **215**, 403–410.
23 Hebsgaard,S.M., Korning,P.G., Tolstrup,N., Engelbrecht,J., Rouzé,P. and Brunak,S. (1996) *Nucleic Acids Res*., **24**, 3439–3452.
24 Henikoff,J.G., Pietrokovski,S. and Henikoff,S. (1997) *Nucleic Acids Res*., **25**, 222–225.
25 Higgins,D.G., Thompson,J.D. and Gibson,T.J. (1996) *Methods Enzymol*., **266**, 383–402.
26 Saitou,N. and Nei,N. (1987) *Mol. Biol. Evol*., **4**, 406–225.
27 Felsenstein,J. (1985) *Evolution*, **39**, 783–791.
28 Swofford,D.L. (1991) *Phylogenetic Analysis Using Parsimony (PAUP) Version 3.1*. Illinois Natural History Survey. Champaign, IL.
29 Nakai,K. and Kanehisa,M. (1992) *Genomics*, **14**, 897–911.
30 Aubourg,S., Chéron,A., Kreis,M. and Lecharny,A. (1998) *Biochim. Biophys. Acta*, **1398**, 225–231.
31 Gy,I., Aubourg,S., Sherson,S., Cobbett,C.S., Chéron,A., Kreis,M. and Lecharny,A. (1998) *Gene*, **209**, 201–210.
32 Bianchi,M.W., Guivarc'h,D., Thomas,M., Woodgett,J.R. and Kreis,M. (1994) *Mol. Gen. Genet*., **242**, 337–345.
33 Weigel,D. and Meyerowitz,E.M. (1993) *Science*, **261**, 1723–1726.
34 Giraudat,J., Hauge,B.M., Valon,C., Smalle,J., Parcy,F. and Goodman,H.M. (1992) *The Plant Cell*, **4**, 1251–1261.
35 Kieber,J.J., Rothenberg,M., Roman,G., Feldmann,K.A. and Ecker,J.R. (1993) *Cell*, **72**, 427–441.
36 Liu,Y.-G., Mitsukawa,N., Lister,C., Dean,C. and Whittier,R.F. (1996) *Plant J*., **10**, 733–736.
37 Creusot,F., Fouilloux,E., Dron,M., Lafleuriel,J., Picard,G., Billaut,A., Le Paslier,D., Cohen,D., Chaboute,M.E., Durr,A. *et al.* (1995) *Plant J*., **8**, 763–770.
38 Bevan,M., Bancroft,I., Bent,E., Love,K., Goodman,H., Dean,C., Bergkamp,R., Dirkse,W., Van Steveren,M., Stiekema,W. *et al.* (1998) *Nature*, **391**, 485–488.
39 Sato,S., Kotani,H., Nakamura,Y., Kaneko,T., Asamizu,E., Fukami,M., Miyajima,N. and Tabata,S. (1997) *DNA Res*., **4**, 215–230.
40 Goodman,H.M., Ecker,J.R. and Dean,C. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 10831–10835.
41 Bevan,M., Ecker,J., Theologis,S., Federspiel,N., Davis,R., McCombie,D., Martienssen,R., Chen,E.,Waterston,B., Wilson,R. *et al.* (1997) *Plant Cell*, **9**, 476–478.
42 Brander,K.A. and Kuhlemeier,C. (1995) *Plant Mol. Biol*., **27**, 637–649.
43 McDowell,J.M., Huang,S., McKinney,E.C., An,Y.-Q. and Meagher,R.B. (1996) *Genetics*, **142**, 587–602.
44 Haouazine-Takvorian,N., Tymowska-Lalanne,Z., Takvorian,A., Tregear,J., Lejeune,B., Lecharny,A. and Kreis,M. (1997) *Gene*, **197**, 239–251.
45 Dornelas,M.C., Lejeune,B., Dron,M. and Kreis,M. (1998) *Gene*, **212**, 249–257.
46 Romero,A., Fuertes,A., Benito,M.J., Malpica,J.M., Leyva,A. and Paz-Ares,J. (1998) *Plant J*., **14**, 273–284.
47 Aubourg,S., Boudet,N., Kreis,M. and Lecharny,A. (1998) submitted.
48 Cooke,R., Raynal,M., Laudié,M. and Delseny,M. (1997) *Plant J*., **11**, 1127–1140.
49 McGrath,J.M., Jancso,M.M. and Pichersky,E. (1993) *Theor. Appl. Genet*., **86**, 880–888.
50 Clegg,M.T., Cummings,M.P. and Durbin,M.L. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 7791–7798.
51 Kalman,M., Murphy,H. and Cashel,M. (1991) *New Biol*., **3**, 886–895.
52 Kiledjian,M. and Dreyfuss,G. (1992) *EMBO J*., **11**, 2655–2664.
53 Nishi,K., Morel-Deville,F., Hershey,J.W., Leighton,T. and Schnier,J. (1988) *Nature*, **336**, 496–498.
54 Sachs,A.B. and Davis,R.W. (1990) *Science*, **247**, 1077–1079.
55 Fuller-Pace,F.V., Nicol,S.M., Reid,A.D. and Lane,D.P. (1993) *EMBO J*., **12**, 3619–3626.
56 Ford,M.J., Anton,I.A. and Lane,D.P. (1988) *Nature*, **332**, 736–738.
57 Rozen,F., Edery,I., Meerovitch,K., Dever,T.E., Merrick,W.C. and Sonenberg,N. (1990) *Mol. Cell. Biol*., **10**, 1134–1144.
58 Valdez,B.C., Henning,D., Perumal,K. and Busch,H. (1997) *Eur. J. Biochem*., **250**, 800–807.
59 Nakajima,T., Uchida,C., Anderson,S.F., Lee,C.-G., Hurwitz,J., Parvin,J.D. and Montminy,M. (1997) *Cell*, **90**, 1107–1112.
60 Buelt,M.K., Glidden,B.J. and Storm,D.R. (1994) *J. Biol. Chem*., **269**, 29367–29370.
61 Li,Q. and Hunt,A.G. (1995) *Plant Mol. Biol*., **28**, 927–934.
62 Hunt,A.G. (1994) *Annu. Rev. Plant Physiol. Plant Mol. Biol*., **45**, 47–60.
63 Gautheret,D., Poirot,O., Lopez,F., Audic,S. and Claverie,J.-M. (1998) *Genome Res*., **8**, 524–530.
64 Klahre,U., Hemmings-Mieszczak,M. and Filipowicz,W. (1995) *Plant Mol. Biol*., **28**, 569–574.
65 Thach,R.E. (1992) *Cell*, **68**, 177–180.
66 Patterson,L.F., Harvey,M. and Lasko,P.F. (1992) *Nucleic Acids Res*., **20**, 3063–3067.
67 Lu,D. and Yunis,J.J. (1992) *Nucleic Acids Res*., **20**, 1967–1972.
68 Maekawa,H., Nakagawa,T., Uno,Y., Kitamura,K. and Shimoda,C. (1994) *Mol. Gen. Genet*., **244**, 456–464.
69 Callis,J., Carpenter,T., Sun,C.-W. and Vierstra,R.D. (1995) *Genetics*, **139**, 921–939.
70 Bartel,B. and Fink,G.R. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 6649–6653.
71 Graham,G.J. (1995) *J. Theor. Biol*., **175**, 71–87.
72 Iggo,R.D., Jamieson,D.J., MacNeill,S.A., Southgate,J., McPheat,J. and Lane,D.P. (1991) *Mol. Cell. Biol*., **3**, 1326–1333.
73 Round,E.K., Flowers,S.K. and Richards,E.J. (1997) *Genome Res*., **7**, 1045–1053.