Characterization of a Single-Copy *Arabidopsis* Gene Encoding a Protein Showing Limited Similarity to the N-terminus of the Mammalian Clathrin-Assembly Protein AP180

Rajeev Gupta and John C. Gray*†

*Department of Plant Sciences and Cambridge Centre for Molecular Recognition, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK*

(Received 14 January 1999)

**Abstract**

The *Arabidopsis* 194 gene encoding a protein containing sequence similarity to an N-terminal region of the clathrin-assembly protein AP180 has been identified in a 4.9-kb region of genomic DNA upstream of the gene encoding the high mobility group protein HMG-I/Y. The gene consists of 12 exons and 11 introns, identified by comparison with partial cDNAs and using the NetPlantGene programme, and encodes a protein of 584 amino acid residues. The C-terminal region of the protein contains 8 tandem repeats of a 17-amino-acid-residue sequence. Southern blot analysis of genomic DNA from Columbia and Landsberg ecotypes of *Arabidopsis* indicates the presence of a single copy of the 194 gene. The 194 gene is expressed in all organs of *Arabidopsis* including roots, stems, leaves, flowers, and developing siliques, as determined by northern blot analysis.

**Key words:** *Arabidopsis thaliana*; clathrin; clathrin assembly protein AP180; high mobility group protein HMG-I/Y; *Mycoplasma pulmonis* variable surface antigen

Clathrin-coated vesicles are involved in pathways of receptor-mediated intracellular transport. These include the movement of proteins from the trans-Golgi network to secretory vesicles for regulated secretion, the transfer of lysosomal hydrolases from trans-Golgi network to lysosomes, receptor-mediated endocytosis, and the biogenesis and recycling of synaptic vesicles. The soluble form of clathrin, the major coat protein of clathrin-coated vesicles, is a triskelion consisting of three identical heavy chains (192 kDa) and three clathrin light chains (22–28 kDa). Triskelions can associate with one another to form icosahedral cages. When these structures form at the surface of a cell membrane, the membrane invaginates into a coated pit, followed by separation from the parent membrane as a coated vesicle. Coated vesicles contain one or more of assembly proteins AP1, AP-2, AP-3, or auxilin. These assembly proteins all share the property that they promote the assembly of clathrin triskelions into a homogenous population of clathrin cages in solution. In vivo, the assembly proteins are believed to sit between the cell membrane and the clathrin cage. The assembly proteins are thought to be involved in directing clathrin cages to a particular cell membrane through an interaction with a membrane receptor. AP-3 was independently discovered in a number of laboratories and has also been known as AP180, pp155, NP185 and F1-20. However, there have been no reports of plant genes encoding these proteins. In the course of sequencing a region of *Arabidopsis* genomic DNA upstream of the single-copy gene encoding high-mobility-group protein HMG-I/Y, we detected a gene encoding a protein showing limited similarity to the N-terminal region of the mammalian AP180 protein. This paper describes the characterisation of the single-copy gene and shows that it is expressed in all organs of *Arabidopsis*.

Previously we had isolated and characterised the single-copy gene encoding high-mobility-group protein HMG-I/Y on a 6.3-kb EcoRI fragment of genomic DNA from *Arabidopsis thaliana* ecotype Landsberg erecta. In order to examine if there were other gene(s) present upstream of the HMG-I/Y gene, the EST database was searched using the Blast programme with ~ 2 kb re-
Figure 1. Nucleotide and deduced amino acid sequences of a 4.9-kb region of Arabidopsis genomic DNA containing the 194 gene. The region is part of a 6.3-kb EcoRI fragment containing the HMG-I/Y gene; overlapping sequence has been presented previously [8; GenBank accession number Y10836]. The 3' ends of the 137 cDNA and the 194 cDNA are shown as arrow (↓, 4719) and underlined arrow (|l, 4854). The introns identified by comparison of cDNA and genomic sequences are in lower case, whereas introns identified using the NetPlantGene programme11 are in lower case and underlined. The sequence is available as GenBank accession number Y10986.
region upstream of the Arabidopsis HMG-I/Y gene as a query sequence. The search resulted in the identification of two ESTs which showed similarity to a region ~1 kb upstream of the HMG-I/Y gene. The two identified clones 137H15T7 and 194P10T7 were obtained from the Arabidopsis Biological Resource Center (Columbus, Ohio, USA) and nucleotide sequences were determined as described earlier for both the 137 (GenBank accession number Y10863) and 194 cDNAs (GenBank accession number Y10864). A comparison of the sequences revealed that cDNA 137 was identical to the 3' end of cDNA 194 but contained a different poly(A) addition site. The 3' UTR was 130 bp longer in the 194 cDNA than that of the 137 cDNA. Analysis of the 1476 bp sequence of cDNA 194 revealed that it contains 1254 bp of coding sequence and 216 bp of 3' UTR excluding a 16-bp poly(A) tail. The cDNA contains in frame ATG codons at positions 72-74, 84-86 and 102-104 which may represent translation start codons. However the open reading frame continues to the 5' end of the cDNA indicating that the cDNA may be incomplete at the 5' end. The cDNA 137 consisted of 524 bp, containing 417 bp of coding region and 85 bp of 3' UTR excluding a 23-bp poly(A) tail. The protein predicted from both cDNAs contained eight repeats of a 17-amino-acid-residue sequence in the C-terminal region. A database search using the Blast program with the 194 protein sequence did not detect any similar protein sequences in the database.

Northern blot analysis (see Fig. 4b) indicated that the sequenced 194 cDNA was incomplete and suggested that ~500 bp was missing from the 5' end. It was therefore decided to determine the complete nucleotide sequence of the 6.3-kb EcoRI genomic fragment. The 6.3-kb Arabidopsis DNA fragment was sub-cloned in smaller fragments to allow double-stranded overlapping sequence as described earlier. Two primers were also designed to cover gaps in the sequence. The primers were P22, 5' GCC AGC CTG CAT TGA TAA CAC TAA CCC 3' (nucleotides 2870-2845 in Fig. 1) and P24, 5' CCC CTG AAT GTA GCA GTA AGG ACT CAC C 3' (nucleotides 1866-1839 in Fig. 1). The nucleotide sequence of a 4.9-kb region upstream of the gene encoding HMG-I/Y and the deduced amino acid sequence are shown in Fig. 1 (GenBank accession number Y10986). The sequence of the HMG-I/Y gene and the 5' end of the 6.3-kb EcoRI fragment have been presented previously. The introns in the region downstream of nucleotide position 2900 were identified by comparing the sequences of the cDNAs and the genomic DNA. In the region upstream of nucleotide position 2900, for which only the genomic sequence was determined, introns were identified using the NetPlantGene programme. Splice junctions for the region upstream of position 2900 were identified by consideration of intron donor/acceptor splice sites which were predicted with high probabilities by NetPlantGene and which allowed in-frame splicing to give a continuous open reading frame (Fig. 1). Six introns were identified in the 5' region of the 194 gene upstream of position 2900 (Fig. 1). When combined with the 5 introns identified in the 3' region of the 194 gene, this suggests that the coding region is interrupted by 11 introns. Most of the splice junctions conform to the plant consensus of AG^GTAAG at the 5' end and TGCAG^GT at the 3' end of the introns. Recent analysis of introns in genes from Arabidopsis has shown the same consensus intron splice junctions. The average length of the introns and exons, present in the 4.9-kb genomic fragment, is 94 bp and 169 bp, respectively. The introns on average contain more A/T bases (61%) than exons, which contain 53% A/T bases on average.

The predicted open reading frame of the 194 gene is 1752 bp which, together with the 3' UTR, is similar to the size of the transcript (~2 kb) detected in the northern blot analysis (Fig. 4). The Arabidopsis 194 gene encodes a putative protein of 584 amino acid residues with a calculated molecular mass of 63.1 kDa and an isoelectric point of 4.1. The predicted protein is rich in leucine and proline which together account for ~20% of the total amino acids. A Blast search using the complete Arabidopsis 194 protein sequence resulted in identification of the rat clathrin assembly protein AP180 and human CALM (Clathrin Assembly Lymphoid Myeloid leukemia gene product) which is a homologue of AP180. The alignment of the rat and human AP180 proteins and Arabidopsis 194 protein over a 28-amino-acid region is shown in Fig. 2a. The Arabidopsis 194 protein shows 42% identity to rat AP180 protein over this region. The N-terminal region of AP180 has been shown to be responsible for clathrin binding, although the particular amino-acid residues responsible for clathrin binding have not yet been identified (Dr. S. A. Morris, Department of Pharmacology, Cambridge, personal communication). However, as the Arabidopsis 194 protein shows sequence similarity to AP180 over a short region only, it is possible that this region has a conserved function in clathrin binding.

Another potentially interesting feature of the Arabidopsis 194 protein is the presence of eight 17-amino-acid repeats in the C-terminal region (amino acid residues 423–557, Fig. 2b). These repeats contain the sequence TTGxG which is present in repeats of the variable-Vl surface antigen A (VsaA) encoded by the vsa gene in the murine pathogen Mycoplasma pulmonis (Fig. 2c). A total of 32 repeats are present in this protein. V1 is a major M. pulmonis surface antigen consisting of proteins that have a repetitive subunit structure. Surface antigens having tandem repetitive domains seem to be prevalent in mycoplasmas and their phylogenetic relatives such as streptococci and clostridia. Proteins with tandem repetitive domains are frequently involved in cell-cell interactions and, in many cases, the repetitive units are thought to be ligand-binding domains. Extracellular repetitive domains have
bands of approximately 8 kb, 5 kb, 9 kb, 7 kb, 6.5 kb, band of ~ 23 kb seen with the I digestion. Single
Xba detected in the I digest in contrast to the largest Ara-
thesis (see later).9 The Southern blot probed with the 32 P-labelled
genome. The smallest band of ~ 1.6 kb was
nylon membrane as described
puted onto GeneScreen Plus nylon mem-
olved by electrophoresis on a 0.7% agarose gel and blot-
Arabidopsis as described earlier. 9 The DNA was digested
solved by electrophoresis on a 0.7% agarose gel and blot-
ecotypes Landsberg erecta (L) and Columbia (C) was digested with various restriction enzymes at 37°C overnight. The fragments were resolved by electrophoresis in a 0.7% agarose gel, transferred to GeneScreen Plus nylon mem-
AR180 clathrin-assembly proteins from rat12 and human.13
The numbers indicate the positions of the amino acid residues. (b) Alignment of 17-amino-acid repeats from the C-terminal region of the Arabidopsis 194 protein. Numbers indicate the positions of the amino acid residues. The most abundant amino acid present at each position is shown as the consensus.
(c) Comparison of consensus repeats from the Arabidopsis 194 protein and the Mycoplasma pulmo-
sis VsaAl protein.14

Also been detected in plant proteins. Leucine-rich repeats have been reported from several disease-resistance genes in plants.17^-20 By comparison to the VsaAl protein and the proteins containing leucine-rich repeats, it is possible that the 17-amino-acid-residue repeats in the 194 gene product may be extracellular and involved in cell-cell or ligand-receptor interactions. However there are no obvious hydrophobic membrane-spanning regions in the predicted 194 protein.

Figure 2. Comparisons of regions of the Arabidopsis 194 protein. Alignments of sequences were done using the ClustalW programme.9 Identical and conserved amino acids are indicated by asterisks (*) and dots ( ), respectively. (a) Similarity to AP180 clathrin-assembly proteins from rat12 and human.13

In order to determine the copy number of the 194 gene in the Arabidopsis genome, genomic DNA was extracted from leaves of A. thaliana ecotypes Landsberg erecta and Columbia as described earlier.9 The DNA was digested with 13 different restriction enzymes, fragments were resolved by electrophoresis on a 0.7% agarose gel and blotted onto GeneScreen Plus nylon membrane as described earlier.9 The Southern blot probed with the 32P-labelled Arabidopsis 137 cDNA demonstrated the presence of only a single hybridising band in each lane (Fig. 3) indicating the presence of a single copy of the 194 gene in the Arabidopsis genome. The smallest band of ~ 1.6 kb was detected in the Nde I digest in contrast to the largest band of ~ 23 kb seen with the Xho I digestion. Single bands of approximately 8 kb, 5 kb, 9 kb, 7 kb, 6.5 kb, 12 kb, 4 kb, 2.3 kb, 2 kb and 3 kb were detected on digestion with BamHI, HindIII, EcoRI, Sty I, EcoRI, Cla I, Xho I, Xba I, Sal I, Dra I, Bgl II, Sac I, 12, Sac I, 13, and Nde I. The positions of DNA size markers are shown at the side.

Acknowledgements: We are extremely grateful to Mandy Walker for help with library screening, to John Lester for DNA sequencing, and to Martyn Seekings for growing plants. Rajeev Gupta was financially supported by the Cambridge Commonwealth Trust, the Cambridge Hinduja Trust, the Downing Miller Trust, the Cambridge Philosophical Society and an Overseas Research Studentship. This work was supported by a grant from the BBSRC, UK.
References


