

# A conserved intron in the V-ATPase *A* subunit genes of plants and algae

Thomas Starke and Johann Peter Gogarten

*University of Connecticut, Department of Molecular and Cell Biology, 75 North Eagleville Rd, Storrs, CT 06269-3044, USA*

Received 11 November 1992; revised version received 30 November 1992

Amplification and sequencing of part of the coding regions of the catalytic V-type ATPase subunit shows the presence of (at least) two genes in all land plants as well as the conservative insertion of a noncoding sequence. The two genes exhibit a coding region of the same length but differ in the number of nucleotides present in the intron. The latter is surprisingly conserved suggesting the presence of functional constraints on the intron sequences. The findings presented in this report indicate that introns from plants and animals are characterized by different internal structural elements.

V-type ATPase; Intron evolution; Gene duplication; Isoform; Gene expression; Plant

## 1. INTRODUCTION

On average more than 90% of the nucleotide sequence in genes from higher eukaryotes does not encode for the final protein product. These stretches of noncoding DNA inserted in the coding region of exons are termed introns. Their origin is still in dispute; the two most popular theories being: (i) Introns are remnants of the original genetic material and are therefore ancient [1]. They separate nucleotides that encode blocks of amino acids that represent structural units (domains); the introns facilitate recombination of the protein 'words' into new functional units [2]. (ii) Introns invaded the coding regions at a later stage of evolution [3]. Most intervening sequences are not translated into an amino acid sequence but are excised from the primary transcript. Many studies using fungal and higher animals as model systems have elucidated the minimum requirement for the correct and efficient removal of the noncoding sequences from the exons. These include conserved 3' and 5' splice sites, an internal branch point and a pyrimidine rich region [4]. Study of plant genes has revealed that they do follow the same general principal but also have some quite different characteristics. It has been shown that the requirements for the splicing of nuclear pre mRNA in plants are substantially different from animal systems. Introns in dicots in order to undergo splicing are required to be rich in the nucleotides

A and U but do not require the internal branch point nor the pyrimidine rich region [5,6]. Recent studies also have demonstrated that not all intervening sequences are selfish 'junk' without functional significance; for example introns were found to influence the level of gene expression [7,8].

In this paper we characterize a surprisingly conserved plant intron present in the gene of the catalytic subunit of the vacuolar-type ATPase. Like the F-type ATPases, V-type ATPases are multimeric enzymes, and they can be dissociated into a water soluble and a membrane portion. Besides several single copy subunits the water soluble part of the V-type ATPase contains three copies each of the two major polypeptides, a catalytic subunit of about 70 kDa binding the ATP that is hydrolyzed during the catalytic cycle (subunit A) and a noncatalytic subunit of about 60 kDa in size (for a recent discussion about function, structure and evolution of V-type ATPases see [9] and other articles in the same volume).

Sequence comparisons established that V-type ATPases are homologous to the bacterial coupling factor ATPases; proton pumping ATPases were shown to be valuable marker molecules for the evolution of organisms and they were successfully used to infer and root a universal tree of life [10]. Building on this information, we discuss the invasion of the catalytic subunit by an intron as well as a gene duplication event in the plant lineage. The finding of surprisingly conserved sequence motifs adds support to an increasing number of reports [11,12] suggesting that plant and animal introns are different and that some noncoding sequences are important for proper gene function. The characterization of these motifs by inter-species comparisons presented in this paper should be a valuable asset to guide future biochemical and genetic investigations of these motifs.

*Correspondence address:* J.P. Gogarten, University of Connecticut, Department of Molecular and Cell Biology, 75 North Eagleville Rd, Storrs, CT 06269-3044, USA. Fax: (1) (203) 486 4331.

## 2. MATERIALS AND METHODS

### 2.1. Materials

All chemicals are from Pharmacia, BRL and Merck. [ $\alpha$ - $^{35}$ S]dATP was purchased from Du Pont NEN Products. Plant material was obtained from the University of Connecticut Biological Sciences greenhouse. Algae (*Coleochaete*, *Zygnema* and *Euglena*) were purchased from Carolina Biological Supply Company (Burlington, NC) together with the corresponding culture media. After sufficient growth under sterile conditions the algae were harvested and stored at  $-90^{\circ}\text{C}$  until used. Oligonucleotide primers were synthesized by the University of Connecticut Biotechnology Center.

### 2.2. Isolation of genomic plant DNA

4 g of fresh plant tissue (less in the case of the algae) were ground in a precooled mortar in the presence of liquid nitrogen. Whole-cell DNA was extracted and purified as described [12]; the resulting vacuum-dried pellet was resuspended in 25  $\mu\text{l}$  dH $_2$ O. 2  $\mu\text{l}$  were used for the subsequent PCR reaction.

### 2.3. The polymerase chain reaction

A portion of the gene coding for the V-type ATPase was amplified by the polymerase chain reaction (PCR). Using the cDNA sequence of *Methanosarcina barkeri*, *Sulfolobus acidocaldarius*, *Neurospora crassa* and *Daucus carota*, two primers were constructed (see Fig. 1). The upstream 5' primer (termed primer left) was 32 bp long and the downstream 3' primer (termed primer right) was 33 bp long. The regions on the cDNA were chosen for being sufficiently conserved especially at the 3' end of the primers. A perfect match at this position is necessary for the polymerase to start the chain elongation. The 3' ends of the primers are 90 bp and 87 bp apart on the cDNA of *Daucus* and *Neurospora* respectively. The PCR was carried out under the conditions previously described [12] and the reaction products were analyzed by electrophoresis on an 0.8% agarose gel. A piece of agarose containing the amplified DNA band was cut out from the gel, placed in an Eppendorf vial and soaked for 12 h at  $4^{\circ}\text{C}$  in 50  $\mu\text{l}$  dH $_2$ O.

### 2.4. The sequencing of PCR amplified DNA

After 12 h of soaking, when sufficient DNA had diffused out of the agarose, 1  $\mu\text{l}$  of the solution was used as template for a second PCR. The conditions for the reaction were the same as in the first, but only 25 cycles were run with a rise in the annealing temperature to  $55^{\circ}\text{C}$  and a reduction in the polymerization time at  $72^{\circ}\text{C}$  to 1 min. 40  $\mu\text{l}$  out of a 50  $\mu\text{l}$  reaction were loaded on a 0.8% agarose gel and run for 2 h with 10 V/cm. The band was localized under UV light and the DNA purified from the gel with activated DEAE cellulose paper with the procedure described in [12]. The resulting eluates were extracted with phenol, then precipitated with ethanol at  $-20^{\circ}\text{C}$  for 12 h, and the pellet was vacuum-dried and resuspended in 25  $\mu\text{l}$  dH $_2$ O.

The nucleotide sequence was determined using the  $^{32}\text{P}$  Sequencing Kit (Pharmacia LKB Biotechnology) and [ $\alpha$ - $^{35}$ S]dATP (1000 Ci/mmol) employing the dideoxynucleotide chain termination method [13]. In some cases, the primers used for sequencing were two shorter versions (18 bp) of the PCR primers, lacking the redundant region (Fig. 1). Annealing of the primers to the template was achieved by denaturing 11.5  $\mu\text{l}$  of the purified double stranded DNA at  $95^{\circ}\text{C}$  for 3 min in the presence of 2  $\mu\text{l}$  annealing buffer and 0.5  $\mu\text{l}$  (50 pmol) of the primer and immediately frozen in an ethanol/dry ice bath. After thawing at room temperature, the sequencing procedure was continued as described in the kit manual. The samples were electrophoresed in a 8% denaturing polyacrylamide gel containing 8 M urea and analyzed by autoradiography. Using the left and right PCR primer for sequencing, the nucleotide sequences of the amplified DNA fragments could be obtained from both directions without any subcloning.

## 3. RESULTS

### 3.1. Intron insertion and gene duplication

The left and right primers (see Fig. 1) were used in the PCR [15] to amplify the corresponding fragments from the isolated genomic DNA from different organisms. The products were longer than the expected 155 bp (90 bp of coding region plus 65 nucleotides of primers) meaning that the coding sequence must be interrupted by an intervening noncoding stretch of DNA. The use of the same PCR primers with DNA isolated from *Euglena* and *Zygnema* revealed that no intron in this stretch of DNA is present in these organisms (Starke and Gogarten, unpublished). Sequences obtained from the literature (*Methanosarcina* [16], *Methanococcus* [17], *Sulfolobus* [18], *Halobacterium* [19] and *Neurospora* [20], *Candida* [21] and *Saccharomyces* [22]) also do not exhibit an insertion of a noncoding sequence at this location. In contrast to the other algae tested, the corresponding fragment of the green algae *Coleochaete* has a very short, 36 bp long intron inserted in the coding region (Fig. 2).

Furthermore, the amplification reaction using genomic DNA from land plants yielded two fragments of different length as visualized on the agarose gel. Sequencing of the fragments and inspection of the coding region revealed that both encode part of the V-type ATPase catalytic subunit. The length variation of the amplified DNA fragments was found to be due to a variation in the number of nucleotides present in the intron. The fragments were termed large and short, respectively. From the amplification reactions with genomic DNA of land plants only the PCR with *Arabidopsis thaliana* produced a single fragment. All other

<i>Methanosarcina barkeri</i> :	TACATCGGTTGTGGTGAGCGTGGAAACGAAAT
<i>Sulfolobus acidocaldarius</i> :	TATGTAAGTTGTGGCGAAAGAGGAAATGAGAT
<i>Neurospora crassa</i> :	TACGTCGGTTGTGGTGACCGCGGTACGAGAT
<i>Daucus carota</i> :	TATGTTGGTTGCGGGGAAAGAGGAAATGAAAT
Primer left:	TATGTCGGTACCGGGGAANGAGGAAANGA <sup>A</sup> GAT
Primer left seq.:	TATGTCGGTACCGGGGAA
<i>Methanosarcina barkeri</i> :	AACACTTCAAACATGCCTGTGGCCGCAACGCGAA
<i>Sulfolobus acidocaldarius</i> :	AATACTAGCAATATGCCAGTAGCAGCTAGAGAA
<i>Neurospora crassa</i> :	AACACCTCTAACATGCCGTCGCGCTCGTGAG
<i>Daucus carota</i> :	AACACTTCAAACATGCCTGTGGCTGCTCGCGAG
Primer complement:	AANACNTCAAACATGCCTGTGGCTGCTCGAGAG
Primer right:	TTNTGNAGTTTGTACGGACACCGACGAGCTCTC
Primer right seq.:	GGACACCGACGAGCTCTC

Fig. 1. Left and right primers that were used to amplify part of the coding region of the V-type ATPase from genomic plant DNA. The strands are shown aligned to the published sequences that were used to construct the primers. Mismatches between the sequences and the primers are emphasized in bold print. The shorter primers used to sequence the amplified DNA are depicted below the primers used for the PCR. In case of the right primer the complementary sequences (bottom lines) were used to prime the synthesis of the strand in the direction of the left primer.

Co	GGG-TA----	CTCGTC	-----ATTAT--ATG-----	-----AGGTCCTTTCTTG----
Cyl	AGG-TGTGTGCGTCG	CTTACCGCTT	-----CGAGTCTCAGATA--CCA	-----TATGATGCTTTTTTGGATG
Cys	AGG-TGTG	CTTATC	-----GTTGTAGATA--TCA	-----TGTGAGGCTTTTTTGTATA
Eq1	GAG-TGAG	ATTACTCTGTGCTTAA	-----TGAGTGTGTGAGATAACACA	-----TGTGAGGCTTTTTTGTGAA
Eqs	GAG-TGTG	ATTATTTTGTCTTA	-----TCTCAGTTGTGAGATA--CCA	-----TGTGAGGCTTTTTTGCATA
Ps1	GAG-TGCAAGTGTAAACCTCTA	CTTAAATGCTTAT	-----GCAGAGTCAGGT--CCT	-----ATGATTTTTTGGCACA
Pss	GAG-TGTGGAGTCC	CTCTCTGCTTA	-----AGAGTTGTGAGATA--CCA	-----TGTGAGGATTTTTTGCATA
Epl	AGG-TGGATC	GTCCCTCCGTGGTC	-----CAATTATGGGATG--CGA	-----CCTGAGGCTCTCTTTGTAGG
Eps	GAG-TGTGCGTAATC	GTTCCTGCT	-----TACGAGATGCGCAATACATCCA	-----ATCAGGTAATTTTTTGTAT
Ar	GAG-TGTGAAAG	GTGAATGGGGT	-----TAAGAGTCTCGG--TCA	-----TTTGACGCTTTTTTTG
Dcl1	CAG-TGCAGC	CTTTGT	-----CGGAGTTCGCGAATAGCTCTAGTTCAATTACAATATTTGT	-----TACTCAGGCTTTTTTGCATA
Dcs	GGG-TGGA	CTTTCT	-----AGGTGTTACGCATAC--CTACTTCTATTTCTTG	-----AATGACGTTTTTTTTG--A
Av1	GAG-TGAGTAGGAACGTGATGTCGCGAATGGCTGAATGCTCTGGTTT	CTTTCT	-----ACCGAGTACTAGATA--GG	-----ACTAATATTTTTTTG--C
Avs	GAG-TGCTTAAAGGCCAGCCAA	CTTTCT	-----AGGTGTTACGCATAC--CTACTTCTATTTCTTG	-----AATGACGTTTTTTTTG--A
Nt1	GAG-TGGATAG	CTTTCT	-----ACCGCTTGC--ATACCATT	-----AATGTTTTTTTTG--G
Nts	GAG-TGCT	CTTTCT	-----CGGAGTTCGAGATAGCTTTAGTTCAAT	-----TCAGGCTTTTTTTGCTAT
Lyl	CGG-TGGTTCGCA	CTTTCT	-----CTTAGCGTATA--ACTTAAGT	-----ATGACATTTTTTTTTG--T
Lys	CGG-TGTGTCGGA	CTTTCT	-----GCCTTCTCG--AACCTCTACTTGGGTAAACGAGGT	-----CCTGATATGTTTTTGGCGGT
Mal	GAG-TGCTTCGGTTCCGGGACACGAGGAAA	CTTTCT	-----GACTCC--CTAGTTGTGTA	-----GCTCATATTTTTTTCGCA
Mas	GAG-TGCTTAAAGCAATTG	CTTTCT	-----GTACTACACGGCGCACACGAGCGGG	-----AATATGGATTTTTTTTTG--C
Hy1	GAG-TGGAGAAAACCAAGTG	CTTTCT	-----AGTGTTCGTAATG--GA	-----AATGACGTTTTTTTTG
Hys	CAG-TGCA	CTTTCT	-----AGTGTTCGTAATG--GA	-----TGTGAGGCTTTTTTTGCACT
C11	GAG-TGCA	CTTTCT	-----AGTGTTCGTAATG--GA	-----GATGTGTTTTTTGCTGT
Cl1	GAG-TGCA	CTTTCT	-----GTGTTCCGATATA--TTGAAT	-----TGAGGTGTTTTTTTTG--G
Cl1s	GAG-TGCA	CTTTCT	-----TCTT--ATACA--TTCTATATGCATGTGAGA	-----T-----TGTTTTTTTGAAC
Ch1	GAG-TGCAGAGGG	CTTTCT	-----TCGTTACCTGACAG--GAAGAGC	-----GAGGAGGA--TTTTTGGGAC
Chs	GAG-TGAA	CTTTCT	-----CCTGTGCTCGATA--TTGAGT	-----TCTCAGGCTTTTTTTTTG--A
Co	-----	-----	-----	-----TGAGCAT-CTT
Cyl	TGAAAGATAGAAATGTAAACTAGGCATTGGATTC	-----	-----	CTTGATAATCCAG-AGT
Cys	GTTAGAGATGCTAAATTTGGATT	-----	-----	CTTCATTATCCAG-ATT
Eq1	TCTGATGAATAATATGGGTGACGTCCGTTATGGAATCTCGATCA	-----	-----	CTTCATTATCCAG-GTA
Eqs	TTGTGAAATGCAAGATTCCGATCA	-----	-----	CTTCATTATCCAG-GIT
Ps1	AGAAGGTTCCCTGTAGACAGAGGATGCCATATGGATTGATATAATTTACTTTGTACGTT	-----	-----	TTTCATTCCAG-GTT
Pss	CTTAAAGATGTTAAATTTGGATCA	-----	-----	CTTCATTATCCAG-GTC
Epl	GCAAGCTAAGTCCAATCTGTCCGTAAGTAGCCAGATCGCACTGCCAGTCCAATC	-----	-----	CTTCGCGCAT-GTA
Eps	-----	-----	-----	TTAGCCCGCCAG-GTT
Ar	-----	-----	-----	TCATCCAG-GTT
Dcl1	CATTGGAACTAAGCTATTGCGGTTAAACGCCCGTTAGCTA	-----	-----	CTTGAAGGCCAT-GTG
Dcs	AATTTGGA	-----	-----	CTTACAGGCCAG-GTA
Av1	GGAT CTGCAATTGTAGCAGCATAGTGTAGCATCCGGTAGCCGTGGTATCTTGACGTAAGGAACT	-----	-----	ATGCTATTGCCAG-GTC
Avs	TATTACCTTTACATGATGTGT	-----	-----	TACTAATGCCAG-GTC
Nt1	TATCGAGCCGTAAGCAATTGCAATCCAGTCCAGTCCGAGTCCGCGG	-----	-----	CTTATAGTCCAG-GTC
Nts	TCTAGGCCATTACGTTAAAGATA	-----	-----	TCATGTGCCAG-GTC
Lyl	TCTGTATAGCGGCTAGACATAAAC	-----	-----	CTTGAATGCCAG-TGA
Lys	-----	-----	-----	CTTCAATTGCCAG-GGA
Mal	TAAATGTA	-----	-----	TCCAAGCCAG-GTT
Mas	-----	-----	-----	TTGATCTCCAG-ATG
Hy1	GAGTAGAAATGTTAAATTTGGATCACTGTCAGTTGATCCAGGTTCTTATGGATTTCACAGCTGACAATGACTTTACCAGAAATGGCCGGGAAGAAGCGTTACATTACGTG	-----	-----	TTACGATCCAG-GTT
Hys	CCCAGCCAGTCCGAGTGCCATCAATATAGCCGCTTGGTTAACTA	-----	-----	TTAGGATCCAG-GTA
C11	GAAACCTTATCACAGGCTCCTTGTGATGGCATTGCAATTACGGTTTAGCAGCTGAACCTCGAGTTCCATT	-----	-----	CTAGGATTTCCAG-GTC
Cl1s	CACAC	-----	-----	TAATATCCAG-GTC
Ch1	TCGACCTCTCTACTTGGTGGCAACCCATTACCGGTGAGCGACAATAATCACCCCTCTCTTAAGCACCGCATGCATCCCTGATGAAATGTGACCTTAAATC	-----	-----	CTTCACCCAG-GTT
Chs	AACATATAGCTCTATAATGAGCCTCCCTTAGGTT	-----	-----	CTGTAGTCTCCAG-CTC

// : AGAAATGGCTAGACGGGAGGCAAAAGAGCAGAGAGGAAAAGAAAGCCGA

Fig. 2. Alignment of the noncoding (intron) sequences. Conserved blocks in the sequence are underlined. Three nucleotides of the adjacent exons are shown to indicate the splice sites. Species are abbreviated by two letters, the third letter indicating the large (l) and small (l) fragment: Co, *Coleochaete scutata*; Cy, *Cyathea crinita*; Eq, *Equisetum arvense*; Ps, *Psilotum nudum*; Ep, *Ephedra altissima*; Av, *Avena sativa*; Ar, *Arabidopsis thaliana*; Da, *Daucus carota*; Ni, *Nicotiana tabacum*; Ly, *Lycopersicon esculentum*; Ma, *Magnolia virginiana*; Hy, *Hydrastis canadensis*; Cl, *Clematis ligusticifolia*; Ch, *Chenopodium rubrum*. The sequence in the large fragment of *Magnolia* indicated by the two slashes is shown at the bottom. Two fragments are present except for *Coleochaete* and *Arabidopsis*. Sequences from *Psilotum* and *Equisetum* are from [13].

higher land plants yielded two fragments demonstrating that there are (at least) two genes coding for the vacuolar type ATPase in the nuclear genome. From *Coleochaete*, the only algal species investigated that had an intron, only one fragment could be recovered from the genomic DNA with the primers used. All fungal sequences of V-type ATPases obtained so far (*Neurospora*, *Candida*, *Saccharomyces*) have revealed the presence of only one gene and confirmed that no intervening sequence is inserted at the corresponding location [20,21,23].

The intron is conservatively inserted at the same location in all sequences (see Fig. 2). Comparing the obtained genomic sequences with the carrot sequence derived from cDNA, the point of insertion of the intron in the coding sequence cannot be pinpointed exactly. The splice site can be shifted one base resulting in either the first 5' base or the 3' base of the intron being a G, the joined exon sequences contain a G in both splice

scenarios (Fig. 2). This makes the exact positioning of the intron arbitrary. The same result, i.e. having two possible insertion points for the intron, is found for all sequences except the large fragment of *Ephedra* and *Avena*. In these two cases the movement of the intron insertion point one nucleotide downstream (3') results in a change in the exon sequence and also in the amino acid encoded by this triplet. In the first case (*Ephedra*) the amino acid would be shifted from serine (AGT) to arginine (AGG), in the second (*Avena*) from histidine (CAT) to glutamine (CAG). Looking at all the deduced amino acid sequences, neither serine nor histidine (as resulting from the upstream insertion position) is present in any other sequence. In contrast, arginine and glutamine are frequently observed at this specific location, suggesting that this downstream positioning of the exon/intron boundary reflects the actual splice sites.

The 5' end of the intron is very close to the 3' end of the left primer, starting only 6 bases following the end

of the primer. Comparing all obtained sequences the consensus sequence for the splice site is found to be 5'-WGAG/tgnr.....ccag/GTHC-3' (intron sequence in lower case letters and W = A or T; N = A, C, G, or T and H = A, C, or T).

This results in the exon sequence of the 5' splice junction complying in most cases with the AG consensus sequence [24,25], but the intron part does not fit the consensus nucleotide r (R = A or G); the 3' splice junction aligns with the proposed yag/R (Y = C or T; R = A or G) consensus sequence [26]. However, it has been noted that this core consensus sequence is derived from the study of fungal and animal sequences and that the requirement for plant intron processing may differ significantly [5,6].

### 3.2. Sequence and structure of the intron

Inspecting the sequences of the noncoding insertions (introns), it is obvious that the length of the intron varies considerably, the shortest one is only 36 bp long (*Coleochaete*) and the longest one is 183 bp long (*Magnolia*). This makes the alignment of the sequences difficult and the insertion of a significant number of gaps necessary. An alignment of the intron sequences is shown in Fig. 2.

To assess the degree of similarity of the various introns, pairwise comparisons of the intervening sequences were used to calculate  $z$ -values as described [27]. The first and last three nucleotides of the intron were dropped from the calculation.  $z$ -values larger than 3 are usually considered to prove sequence homology, i.e. the observed similarity is not due to convergent evolution or simple chance [28]. Comparisons of the small fragments generally result in  $z$ -values larger than three (average = 3.8). Significant  $z$ -values are also obtained for comparisons among the large fragments, the average being 4.0. The cross comparison (large vs. small) results in significant  $z$ -scores within the group of Pteridophytes (*Cyathea*, *Psilotum* and *Equisetum*) and in several of the other comparisons (average = 2.8). These  $z$ -values show that the introns themselves, even disregarding the residues defining the splice site, contain a significant amount of sequence similarity; this finding again indicates the homology of the intron sequences,

which is already suggested by the identical insertion point in the genes.

A closer inspection of the sequences reveals five conserved regions that are present in all sequences. These more or less conserved blocks are underlined in Fig. 2; a schematic representation of the evolutionary conserved features is given in Fig. 3. Looking at this schematic representation of the intron, the conserved blocks were labeled block 1 through block 5, starting at the 5' end of the intron. The first and last block contain the conserved sequences at the junctions with the exon. Block number two is a block of 6 nucleotides which in most sequences is close to block one. Among the small fragments, only *Avena* and *Magnolia* have a comparatively high number of nucleotides between the first two blocks. The large fragment of *Psilotum*, *Avena* and *Hydrastis* also exhibit a relatively large insertion between these two blocks.

The large fragment of *Magnolia* is exceptional because it has a sequence of 165 nucleotides separating the two blocks (as compared to the second largest being *Avena* with only 26 nucleotides). The distance between blocks two and three ranges from 0 to 20 nucleotides. The most obviously conserved stretch of DNA within the intron sequence is block 4 with a core of thymidines followed by a guanidine. This region (termed TG block) is conserved throughout all sequences, indicating some vital function. All conserved blocks are separated by variable regions that do not exhibit a high degree of sequence similarity. V 4, the variable region between the TG block and the 5' splice site shows the highest degree of heterogeneity with respect to nucleotide sequence and length. The different sizes of the introns are mostly due to the different numbers of nucleotides in this region.

### 3.3. Orthologous and paralogous genes

The finding of two genes (a large and a small fragment) characterizes a gene duplication event. The calculated  $z$ -values suggest that the large fragments and the small fragments from different species are orthologous to each other. To test this hypothesis, Lake's method of evolutionary parsimony or invariants was used [29]. An example of the results employing this method is depicted in Fig. 4. The data used in this example are

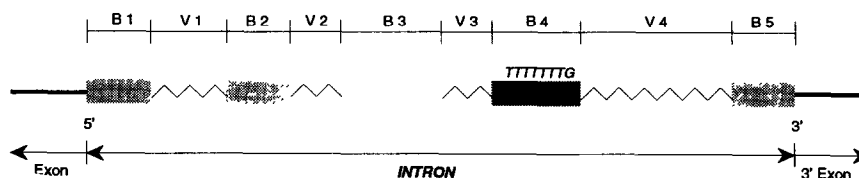


Fig. 3. Schematic representation of the intron structure. The conserved regions are represented as boxes and labeled on the upper axis (B1 to B5), starting from the 5' end of the intron. The conserved stretches are separated by more variable regions (V1 to V4). The most prominent block consisting of 7 thymidines followed by a guanidine is indicated (B4).

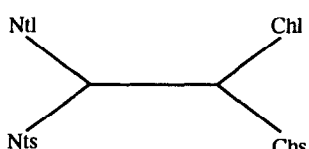
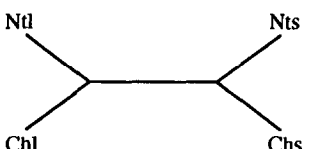
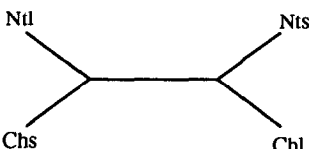
Tree Topology	$\Sigma$	P
	2	0.432
	14	0.003
	1	0.500

Fig. 4. Test of different tree topologies using Lake's method of evolutionary parsimony [29]. The example depicted was performed with the large and small intron of two higher land plants: *Nicotiana* (Nt) and *Chenopodium* (Ch) (see section 3). The sum of the phylogenetic invariants is given in column two ( $\Sigma$ ) and the probability that this or a higher sum arises by chance alone in column three (P). Only tree number two is supported by this analysis: The large and small *Nicotiana* fragments are homologous to the large and small fragments of *Chenopodium*, respectively. The same result (support for the large/small tree) is obtained using different combinations of seed plants with the minimum sum of invariants being 14 and the lowest significance level being 8%.

introns from two angiosperms (*Nicotiana* and *Chenopodium*). The result shows that only one tree topology is supported; the large fragments group with the large fragments and the small fragments with the small ones. The same result is obtained using different combinations of seed plants, demonstrating that the small fragments are orthologous among each other. The same is true for the large fragments; i.e. they also reflect the evolution of the species. In contrast, the two fragments (the large and small) are paralogous; their presence reflects an ancient duplication event that gave rise to the two genes.

#### 4. DISCUSSION

##### 4.1. A gene duplication in the plant lineage

The presence of two genes coding for the A subunit of the V-type ATPase in higher land plants must be due to a duplication of the gene present in the ancestor. *Coleochaete* exhibits the presence of the noncoding sequence but only one gene could be detected. This implies that the immediate ancestor of *Coleochaete* and the higher land plants had the intron and that the duplication took place after the bifurcation separating *Coleochaete* and the higher land plants.

Given the time of divergence (at least 400 million years) it is noteworthy that the insertion of the intron is conservative, i.e. the location did not change during evolution. Although this might be coincidental, it might also imply some functional characteristics of the intron associated with its position within the sequence.

Except for *Arabidopsis*, PCR on the genomic DNA of all higher land plants with the primers yielded two fragments. Both fragments show a coding region of the same length. The length variation of the fragments was only due to the different length of the noncoding sequence. The small fragments are obviously more uniform in their length, the smallest one being 66 bp long, the longest being 96 bp long. Only the intron of *Arabidopsis* is shorter being 60 bases long. In contrast, the length of the large introns is more heterogeneous, the length of the smallest being 108 bp and the length of the largest being 183 bp.

The purpose of the gene duplication of the V-ATPase catalytic subunit cannot be deduced from the available data. One reason might be to compensate for an increased need for the protein product. Such duplications resulting in a higher gene dosage are fairly common [30].

It has long been noted that gene duplication is an important mechanism for the acquisition of new functions and it has even been suggested that this is the only means by which a new gene can arise [31]. The duplication might thus also have been accompanied by a change or optimization in functional features (e.g. with respect to the location of the protein product in the cell). Such a differentiation does not necessarily require a large number of substitutions, a relative small change might be sufficient to create this effect [32].

Genomes of higher organisms often contain two or more genes belonging to one family that have diverged to some extent. For a variety of such isoforms it has been shown that their expression is tissue and/or developmental specific [33]. So far only one cDNA encoding the A subunit in carrot has been characterized [34]. The relevant region of the published carrot cDNA corresponds to the coding region of the small fragment. However, the presence of organelle specific isoforms of the A subunit in carrots was previously indicated by specific inhibition of the V-ATPase isoform at the tonoplast by different antisense mRNAs; an isoform with a more acidic isoelectric point present in Golgi-enriched microsomal fractions was not inhibited [35].

##### 4.2. Intron evolution

After the discovery of split genes it had been proposed that recombination events in introns provide a means for the exchange or reshuffling of coding sequences (exons) and thereby play a significant role in evolution [2]. Furthermore it had been found that in many cases noncoding regions are inserted adjacent to structural or functional domains [36,37]. The intron in the catalytic subunit of the vacuolar ATPase from *Dau-*

*cus* is inserted shortly after the ATP binding site. A lysine residue 28 amino acids upstream of the last amino acid before the insertion of the noncoding sequence marks the putative catalytic site [38]. A cysteine residue 8 amino acids upstream has been proposed to mark the  $Mg^{2+}$  binding region [39]. It might be speculated that the position of insertion of the intron within the coding sequence marks the downstream (3') end of a functional (i.e. ATP binding site) domain. However, mapping the presence or absence of the intron onto the tree of life suggests that the intron invaded the coding region comparatively late in the evolution, i.e. just before the land plants evolved from their green algal ancestor (the intron is present in *Coleochaete* but absent in the related green algae *Zygnema*). At this time the domains and words had long since been assembled into the functional catalytic subunit. The scenario that all the other species have lost the intron and only the plant lineage retained the intron from the time when the subunit evolved by 'exon shuffling' is extremely unlikely. This scenario requires the multiple intron loss in the lineages leading to eubacteria, archaebacteria, fungi, flagellates and green algae. All of these groups branch off independently from the lineage leading from the last common ancestor to the land plants. Therefore, all of these intron losses would have had to occur independently from each other. Clearly, one gain of an intron is a much more parsimonious scenario compared to at least five independent losses of the intervening sequence. Thus, if indeed the intron marks a domain or word boundary in the encoded protein, one has to postulate that this boundary is reflected by other means on the genomic level, and that this unknown feature also guided the insertion of the intron sequence.

#### 4.3. The internal intron structure

Looking at the intervening sequences of the different species it is obvious that they exhibit a conserved internal structure as well as a high degree of sequence similarity. That these features are conserved throughout at least 400 million years of evolution clearly indicates that there are functional constraints on the sequence. The exact reason for the high degree of conservation of the structure within the noncoding sequence remains elusive. It seems likely that the conserved regions at the exon/intron junctions are probably important for the removal (splicing) of the intron [40]. Based upon analysis of introns from animal and fungal sequences, it had been suggested that each intron exhibits 4 structural elements, necessary for the correct splicing of the intron [26]. These are the proposed proto splice sites (see results), an internal pyrimidine rich region and the YTRAY (Y = C or T; R = A or G) box. These requirements do seem to be relaxed or altogether absent from this particular plant sequence. Only one proposed splice site consensus sequence can be found and the YTRAY box is absent from most of the sequences. However, it

had been previously observed in dicots that this branch point as well as the pyrimidine rich region is not needed for intron processing [6], but in order to undergo splicing, they are only required to be rich in the nucleotides A and U [5].

The most prominent region within the intron is block 4 (TG block). This characteristic sequence (as well as the splice junctions) is conserved in all sequences and constitutes a fixed point for the alignment. The presence of this 'landmark' in all sampled species indicates some functional importance. It is noteworthy that the distance between this block and the 3' splice site is comparatively constant as opposed to its distance to the 5' splice junction. In general one can observe a more conserved structure of the intron in the upstream region (3' of the TG block). The structure and sequence are more heterogeneous (with respect to sequence and length) downstream of the TG block.

Dividing the number of nucleotides present in the intron by three always gives an integer result. This is somewhat a surprise considering the different lengths of these sequences. These differences in intron length obviously arose from a number of insertion and deletion events during the evolution. Assuming random single nucleotide insertions/deletions it is very unlikely that the length of all the introns examined in this study turns out to be a multiple of three. In all species examined, at least one intron (either the large or the small one) exhibits an open reading frame that is continuous with the reading frame of the exon. A comparison of the intron sequences with the entries in the genbank using the program FASTA [41] did produce low level sequence similarities to a wide variety of genes but was inconclusive in terms of a dominant score or a common motive among the selected sequences.

Concerning the nucleotide and dinucleotide composition, the introns show the general trends usually associated with noncoding sequences [42,43]. The intron sequences exhibit a low G + C content (41% G + C, 59% A + T), and they are depleted in CG dinucleotides (3.45%) as compared to AT (7.75%) and TG (8.52%). Comparing the introns that have a continuous open reading frame to the ones that contain stop codons reveals no difference between the two groups. This observation suggest that the introns are not translated into protein, and that the presence of open reading frames might be coincidental. However, the preservation of the multiplicity of 3 over the time spans involved without the presence of some evolutionary pressure or the action of some unknown mechanism is hard to explain.

#### 4.4. Conclusion

A gene duplication in the plant lineage gave rise to two genes encoding the V-ATPase *A* subunit. An intron was inserted into this gene before the duplication occurred; the exact location of this intron is maintained in both genes and throughout the plant kingdom. The

intron exhibits a surprisingly conserved internal structure. The described sequence motifs are conserved throughout plant evolution. This finding clearly indicates selection acting to preserve these motifs; it also implies a functional importance of these sequences. The structural requirements for intron excision from the primary transcript differ in the plant and animal kingdom. The availability and study of more intron sequences from plants changes our perception about noncoding sequences and will lead to a better understanding of their importance and evolution.

**Acknowledgements:** This work was supported by a grant from the National Science Foundation (NSF BSR-9020868). We gratefully acknowledge the help of Timothy Linkkila, Patricia Gayda and Jan Thomas Johansson for their assistance in DNA isolation and sequencing.

## REFERENCES

- [1] Darnell, J.E. and Doolittle, W.F. (1986) *Proc. Natl. Acad. Sci. USA* 83, 1271–1275.
- [2] Gilbert, W. (1987) *Cold Spring Harbor Symp. Quant. Biol.* 52, 901–905.
- [3] Roger, J.H. (1989) *Trends Genet.* 5, 213–216.
- [4] Dibb, N.J. (1991) *J. Theor. Biol.* 151, 405–416.
- [5] Goodall, G.J. and W. Filipowicz, W. (1989) *Cell* 58, 473–483.
- [6] Goodall, G.J., Kiss, T. and Filipowicz, W. (1992) *Oxf. Surv. Plant Mol. Biol.*, in press.
- [7] Beaudet, L., Charron, G., Houle, D., Tretjakoff, I., Peterson, A. and Julien, J.P. (1992) *Gene* 116, 205–214.
- [8] Takayanagi, A., Kaneda, S., Ayusawa, D. and Seno, T. (1992) *Nucleic Acids Res.* 20, 15, 4021–4025.
- [9] Kibak, H., Starke, T., Bernasconi, P. and Gogarten, J.P. (1992) *J. Bioenerg. and Biomemb.* 24, 4: 415–424.
- [10] Gogarten, J.P., Kibak, H., Starke, T., Fishman, J. and Taiz, L. (1992) *J. Exp. Biol.*, in press.
- [11] Goodall, G.J. and Filipowicz, W. (1991) *EMBO J.* 10, 2635–2644.
- [12] Van Santen, V.L. and Spritz, R.A. (1987) *Gene* 56, 253–256.
- [13] Starke, T., Linkkila, T.P. and Gogarten, J.P. (1991) *Z. Naturforschung* 46c: 613–620.
- [14] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463–5467.
- [15] Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988) *Science* 239, 487–491.
- [16] Inatomi, K.I., Eya, S., Maeda, M. and Futai, M. (1989) *J. Biol. Chem.* 264, 10954–10959.
- [17] Gogarten, J.P., Rausch, T., Bernasconi, P., Kibak, H. and Taiz, L. (1989) *Z. Naturforsch.* 44c, 641–650.
- [18] Denda, K., Konishi, J., Oshima, T., Date, T. and Yoshida, M. (1988) *J. Biol. Chem.* 264, 10954–10959.
- [19] Mukohata, Y., Ihara, K., Kishino, H., Hasegawa, M., Iwabe, N. and Miyata, T. (1990) *Proc. Jpn. Acad.* 66, 63–67.
- [20] Bowman, E.J., Tenney, K. and Bowmann, B.J. (1988) *J. Biol. Chem.* 263, 13994–14001.
- [21] Schafer, M.P., Howell, M., Xu, J. and Dean, G.E. (1992) *Candida tropicalis* DNA sequence, GenBank entry M64984.
- [22] Hirata, R., Ohsumi, Y., Nakano, A., Kawasaki, H., Suzuki, K. and Anraku, Y. (1990) *J. Biol. Chem.* 265, 6726–6733.
- [23] Kane, P.M., Yamashiro, C.T., Wolczyk, D.F., Neff, N., Gobel, M. and Stevens, T.H. (1990) *Science* 250, 651–657.
- [24] Black, D.L., Chabot, B. and Steitz, J.A. (1985) *Cell* 42, 737–750.
- [25] Jacob, M. and Gallinaro, H. (1989) *Nucleic Acids Res.* 17, 1259–1280.
- [26] Keller, E.B. and Noon, W.A. (1985) *Nucleic Acids Res.* 13, 4971–4981.
- [27] Needleman, S.B. and Wunsch, C.D. (1970) *J. Mol. Evol.* 48, 443–453.
- [28] Lipman, D.J. and Pearson, W.R. (1985) *Science* 227, 1435–1441.
- [29] Lake, J.A. (1987) *Mol. Biol. Evol.* 4, 167–191.
- [30] Ohta, T. (1988) *Proc. Natl. Acad. Sci. USA* 85, 3509–3512.
- [31] Ohta, T. (1991) *J. Mol. Evol.* 33, 34–41.
- [32] Betz, J.L., Brown, P.R., Smyth, M.J. and Clarke, P.H. (1974) *Nature* 247, 261–264.
- [33] Doyle, J.J. (1991) *Mol. Biol. Evol.* 8, 366–377.
- [34] Zimniak, L., Dittrich, P., Gogarten, J.P., Kibak, H. and Taiz, L. (1988) *J. Biol. Chem.* 263, 9102–9112.
- [35] Gogarten, J.P., Fichman, J., Braun, Y., Morgan, L., Styles, P., Taiz, S.L., DeLapp, K. and Taiz, L. (1992) *The Plant Cell* 4, 851–864.
- [36] Go, M. (1981) *Nature* 291, 90–92.
- [37] Liaud, M.F., Brinkmann, H. and Cerff, R. (1992) *Plant Mol. Biol.* 18, 639–651.
- [38] Andrews, W.W., Hill, F.C. and Allison, W.S. (1984) *J. Biol. Chem.* 259, 14378–14382.
- [39] Yoshida, M., Poser, J.W., Allison, W.S. and Esch, F.S. (1981) *J. Biol. Chem.* 256, 148–153.
- [40] Rosbash, M. and Seraphin, B. (1991) *Trends Biochem. Sci.* 16, 45–48.
- [41] Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448.
- [42] Li, W.H., Wu, C.I. and Luo, C.C. (1984) *J. Mol. Evol.* 21, 58–71.
- [43] Razin, A. and Riggs, A.D. (1980) *Science* 210, 604–610.