Structure and Evolution of Linalool Synthase

Leland Cseke, Natalia Dudareva, and Eran Pichersky

Department of Biology, University of Michigan

Plant terpene synthases constitute a group of evolutionarily related enzymes. Within this group, however, enzymes that employ two different catalytic mechanisms, and their associated unique domains, are known. We investigated the structure of the gene encoding linalool synthase (LIS), an enzyme that uses geranyl pyrophosphate as a substrate and catalyzes the formation of linalool, an acyclic monoterpene found in the floral scents of many plants. Although LIS employs one catalytic mechanism (exemplified by limonene synthase [LMS]), it has sequence motifs indicative of both LMS-type synthases and the terpene synthases employing a different mechanism (exemplified by copalyl diphasate synthase [CPS]). Here, we report that LIS genes analyzed from several species encode proteins that have overall 40%–96% identity to each other and have 11 introns in identical positions. Only the region encoding roughly the last half of the LIS gene (exons 9–12) has a gene structure similar to that of the LMS-type genes. On the other hand, in the first part of the LIS gene (exons 1–8), LIS gene structure is essentially identical to that found in the first half of the gene encoding CPS. In addition, the level of similarity in the coding information of this region between the LIS and CPS genes is also significant, whereas the second half of the LIS protein is most similar to LMS-type synthases. Thus, LIS appears to be a composite gene which might have evolved from a recombination event between two different types of terpene synthases. The combined evolutionary mechanisms of duplication followed by divergence and/or “domain swapping” may explain the extraordinarily large diversity of proteins found in the plant terpene synthase family.

Introduction

Terpenes are among the most widespread and chemically diverse groups of compounds produced by plants. They act in a large number of roles, e.g., hormones, mediators of polysaccharide synthesis, photosynthetic pigments, electron carriers, and membrane components, among other essential functions within the plant (Chappell 1995; McGarvey and Croteau 1995). They also mediate interactions between plants and their environment: some serve as toxic defense compounds to ward off herbivores or fungal attacks (Johnson and Croteau 1987; Lewinsohn, Gijzen, and Croteau 1992; Nebeker, Hodges, and Blance 1993), while others are volatile compounds released from the plants to attract pollinators (Dobson 1993; Knudsen and Tollsten 1993).

Terpenoid compounds (also called isoprenoids) are made from the condensation of one molecule of dimethylallyl pyrophosphate (DMAPP) and one or more molecules of isopentenyl pyrophosphate (IPP), to give geranyl diphasate (GPP), farnesyl diphasate (FPP), or geranylgeranyl diphasate (GGPP) (Gershenzon and Croteau 1993). Practically all other terpenes are derived from these three precursors. The enzymes that use GPP, FPP, or GGPP as substrates are termed terpene synthases or, when the product formed is a cyclic compound, terpene cyclases. The mechanisms of action and of the evolution of these enzymes are of particular interest for an understanding of the many plant processes in which terpenes are involved.

Research on terpenes synthesis has focused on the cyclases, such as limonene synthase (LMS), which catalyzes the formation of limonene, a monoterpenoid defense compound. So far, LMS (a protein with approximately 600 amino acids) is one of the few terpene synthases whose gene has been characterized for diverse species. The deduced amino acid sequences of LMSs from various angiosperms are 40%–65% identical to each other, but share only ~28% amino acid sequence identity with LMS from the gymnosperm Abies grandis (Bohmann, Steel, and Croteau 1997). Significant levels of interspecific sequence identity are also seen with copalyl diphasate synthase (CPS, formerly ent-kaurene synthase A). CPS (a protein with approximately 800 residues), together with the similarly sized ent-kaurene synthase (KS, formerly ent-kaurene synthase B), is responsible for the production of ent-kaurene (the precursor of gibberellin-type plant hormones) from GGPP (Yamaguchi et al. 1996). The deduced amino acid sequences of CPS from A. thaliana, tomato, pea, pumpkin, and corn species are only 45%–50% identical (Sun and Kamiya 1994, 1997; Benson et al. 1995).

Certain sequence motifs shared by all plant terpene synthases suggest that they share a common evolutionary origin (McGarvey and Croteau 1995; Bohmann, Meyer-Gauen, and Croteau 1998). Interestingly, the level of sequence identity between “same” synthases (i.e., enzymes with the same substrate and same product) in different species can be lower than that between two different synthases. This is clearly shown in the phylogenetic analysis of the terpene synthase protein family carried out by Croteau and co-workers (Bohmann, Meyer-Gauen, and Croteau 1998), in which it was found that the same terpene synthases (e.g., LMSs) from different plant species belonged in two different subfamilies, and each of these subfamilies also contained enzymes whose substrates and products were quite different.

Gene structure also points to the common ancestry of many terpene synthases. LMS, 5EAS (5-epi-aristolochene synthase), VES (vetispiradiene synthase), and
AGraceae), as well as in those of many other plant spe-
er the plant terpene synthases seem to fall into two main categories. One class, exemplified by LMS, employs a mechanism involving the ionization of the substrate allylic diphosphate to initiate the reaction (Vogel et al. 1996). A second class, exemplified by CPS, employs a mechanism involving substrate double-bond protonation (Vogel et al. 1996). An important feature of the first class of terpene synthases is the DDXXD motif found in the C-terminal halves of these proteins and shown to be involved in metal cofactor binding and catalysis (Ashby and Edwards 1990; Tarshis et al. 1994; Starks et al. 1997; Lesburg et al. 1997). CPS does not contain such a motif, but instead has an aspartate-rich motif in its N-terminal half that is also believed to be involved in metal cofactor binding (Sun and Kamiya 1994). Thus, the classification based on levels of amino acid sequence similarities among various terpene synthases, which places CPS in a separate subfamily, is consistent with its distinct mechanism.

Although some terpene synthases, such as abietadiene synthase (ABS) and linalool synthase (LIS), are not easily classified based on amino acid sequence comparisons alone. This is so because they have higher sequence similarity to CPS in their N-terminal parts but display higher similarity to LMS-type synthases in their C-terminal parts. Therefore, it has previously been suggested that ABS and LIS (both proteins have similar sizes of 868–886 residues but are placed in different subfamilies based on their amino acid sequences) are of composite origin (Dudareva et al. 1996; Vogel et al. 1996). It is important to note, however, that assessment of the evolutionary relatedness of enzymes with different activities based on amino acid sequence comparisons alone may be misleading, because selective pressures may cause different rates of evolution in corresponding parts of the coding regions in the various gene lineages, depending on the function of each region in the different enzymes.

We have been studying the biosynthesis of linalool, an acyclic monoterpene alcohol found in the floral scents of Clarkia breweri (Pichersky et al. 1994; Pichersky, Lewinsohn, and Croteau 1995; Raguso and Pichersky 1995) and Oenothera spp. (all in the family Onagraceae), as well as in those of many other plant species (Kaiser 1991; Knudsen, Tollsten, and Bergstrom 1993). A cDNA encoding linalool synthase (LIS1, formerly LIS) was previously isolated and characterized from a C. breweri floral cDNA library (Dudareva et al. 1996). As part of our investigation to understand how the LIS gene promoter in C. breweri is differentially regulated in several flower parts during floral development, we isolated the LIS gene from two linalool-scented Onagraceae species, C. breweri and Oenothera arizonica, and from a nonscented one, Clarkia concinna, which has nonetheless been shown to express its LIS gene at a low level in the stigma (Pichersky et al. 1994; Dudareva et al. 1996). In comparing the structures of these LIS genes (and that of Arabidopsis thaliana, which was independently obtained) with the structures of other terpene synthase genes, we discovered that the structure of LIS indicates that it is a composite gene. The results of these comparisons, and their implications for the evolution of the terpene gene family, are presented here.

Materials and Methods
Construction of Partial Genomic Libraries from C. breweri, C. concinna, and O. arizonica

DNAs were digested with EcoRI or EcoRV and separated on 0.7% agarose gels. Fragments in the appropriate range (as previously determined by Southern blot analysis with an LIS1 probe) were cut out and eluted from the gels. In the case of EcoRV-generated fragments, EcoRI/NorI adaptors (Pharmacia Biotech) were ligated to the blunt ends of these fragments. Partial genomic libraries were then constructed and screened using the Lambda ZAP II/EcoRI/CIAP Cloning Kit (Stratagene) according to manufacturer’s instructions. Libraries were screened with LIS1 as a probe under high-stringency conditions (65°C hybridization temperature) for the C. breweri library and under low-stringency conditions (58°C hybridization temperature) for O. arizonica library, with all other conditions being identical to the ones described in Dudareva et al. (1996).

Isolation of C. breweri LIS2 cDNA

A C. breweri floral cDNA library (Dudareva et al. 1996) was screened with a probe derived from LIS1 (a fragment containing the first 2.2 kb of the coding region) using low-stringency conditions (58°C hybridization temperature). Screening and characterization were done as previously described (Dudareva et al. 1996).

GenBank Searches and Sequence Analysis

The A. thaliana LIS gene (accession number AC002294) was found in a BLAST search of GenBank using the LIS1 protein sequence as the query. Introns within the Arabidopsis LIS gene sequence, as well as the LIS genomic sequences of C. breweri, C. concinna, and O. arizonica, were identified using the C. breweri LIS1 cDNA sequence for comparison.

Sequence Comparisons

Sequence alignments were performed using the ClustalW1.7 program (http://dot.imagen.bcm.tmc:9331/multi-align.html).
Northern Blots of *C. breweri* LIS1 and LIS2 mRNAs

Total RNA samples from vegetative and floral organs of *C. breweri* were prepared using the Qiagen RNeasy® mRNA extraction kit and run on vertical agarose gels under the same conditions described by Dudareva et al. (1996). Northern blots were probed with LIS1- and LIS2-specific probes under high-stringency conditions, and results were quantified and normalized to equal amounts of 18S RNA as previously described (Dudareva et al. 1996). The LIS1-specific probe was derived from the 5' region of the LIS1 cDNA (the first 650 nt of the coding region), and had ~70% identity with the corresponding region in LIS2. The LIS2-specific probe was derived from the 5' region of the LIS2 cDNA (the first 500 nt of the coding region) and was also ~70% identical with the corresponding region in LIS1. Test blots with cloned DNAs indicated that the probes did not cross-hybridize under high-stringency conditions.

Antibody Production and Western Blots

An LIS1 cDNA fragment (codons 362–653) and an LIS2 cDNA fragment (codons 394–789) were expressed in *Escherichia coli* using the pET-T7 (Studier et al. 1990) and pTRC99A (Pharmacia) expression systems, respectively. The resulting polypeptides were purified on SDS-PAGE and used to produce antibodies in rabbits (Cocalico Inc., Reamston, Pa.). Western blots with anti-LIS1 and anti-LIS2 antibodies were performed as described by Dudareva et al. (1996) using a 1:4,000 dilution for both sera.

Results

Isolation of Plant LIS Sequences

The isolation of a *C. breweri* LIS1 cDNA (accession number U58314) was previously reported (Dudareva et al. 1996). It was derived from the LIS1 gene, which was shown to be expressed in floral tissues that emit linalool (Dudareva et al. 1996). For this study, we isolated a genomic DNA fragment that corresponds to the first half of the coding region of LIS1 as well as 1.5 kb of the 5' region upstream of the coding region (accession number AF067601). Using the LIS1 cDNA as a probe in low-stringency hybridizations, we isolated an additional cDNA, LIS2 (accession number AF067603), encoding a protein distinct from LIS1. We also isolated a complete LIS genomic sequence from *C. concinna* (accession number AF067602), the proposed progenitor of *C. breweri* (Lewis 1962; MacSwain, Raven, and Thorp 1973; Raguso and Pichersky 1995), using the *C. breweri* LIS1 cDNA as a probe. *Clarkia concinna* flowers are known to express LIS at very low levels only in stigma tissue and to emit 1,000-fold less linalool than *C. breweri* flowers (Pichersky et al. 1994; Dudareva et al. 1996). Another relative of *C. breweri*, *O. arizonica*, produces copious amounts of linalool, and a partial LIS genomic clone from *O. arizonica* (roughly corresponding to the second half of the gene) was isolated and characterized (accession number AF067604), again using the LIS1 cDNA as a probe. Finally, we have identified a genomic sequence with high levels of similarity to the *C. breweri* LIS sequences in the *Arabidopsis thaliana* genome project databank (accession numbers AC002294 and AF067605).

Comparison of LIS1 and LIS2 Expression in *C. breweri*

The tissue-specific, developmentally regulated expression of *C. breweri* LIS1 was reported earlier (Dudareva et al. 1996). Examination of the expression of LIS2 by northern blot analysis using a gene-specific probe indicates that its expression pattern is quite different from and, overall, much lower than that of LIS1 (fig. 1). As previously reported (Dudareva et al. 1996), in freshly opened flowers the highest levels of LIS1 mRNA transcripts are found in the stigma, followed by the petal, style, stamen, and sepal tissues, and no transcripts are seen in vegetative tissues. LIS2, on the other hand, has mRNA levels comparable to those of LIS1 only in petals, with much lower levels detectable in the stigma, styles, sepals, and stamens. As with LIS1, no LIS2 transcripts are found in vegetative tissues.

To test for the presence of LIS2 protein, antibodies against the LIS2 protein that reacted with *E. coli*-expressed LIS2 but not with LIS1 were produced (data not shown). However, the anti-LIS2 antibodies did not detect protein in any floral (or vegetative) tissue, whereas the anti-LIS1 antibodies did detect LIS protein (Dudareva et al. 1996 and additional data not shown). These results suggest that LIS2 mRNA is produced in floral tissues but that its translation results in very low levels of protein or perhaps in a dysfunctional enzyme that is rapidly degraded, consistent with our previous results in which only the LIS1 protein could be isolated and identified from *C. breweri* flowers (Pichersky, Lewinsohn, and Croteau 1995; Dudareva et al. 1996).
Comparison of the Promoter Sequences of *C. breweri* LIS1 and *C. concinna* LIS

The transcriptional start site for *C. breweri* LIS1 was previously located 53 nt upstream of the translational start site (Dudareva et al. 1996). A possible TATA box for this gene is located 25 nt upstream of the transcriptional start site, and a possible CAATT box is located 147 nt upstream of the translational start site (fig. 2). When the upstream sequences of the *C. breweri* and *C. concinna* LIS genes are aligned, comparison indicates that they have high sequence identity with each other (~93% DNA identity in the region illustrated in fig. 2). Compared with the *C. breweri* LIS1 promoter sequence, the *C. concinna* CAATT box has an insertion of four adenine nucleotides, creating the sequence CAAAAAATT (fig. 2), and several other insertions occur within this *C. concinna* promoter region, two of them immediately upstream of the putative TATA box. It is possible that these differences in the LIS promoters from the two related species are responsible for the vastly different expression characteristics of this gene in the two species, but this remains to be verified experimentally.

Structures of LIS Genes from Different Species

The *C. concinna* LIS gene contains 11 introns. The *A. thaliana* gene tentatively identified as LIS also contains 11 introns in identical positions within the gene (illustrated in fig. 3). In addition, the first six introns (I–VI) of the partial *C. breweri* LIS genomic clone are in positions identical to those in the full-length LIS genomic clones of *C. concinna* and *A. thaliana*, and the same is true for the last five introns (VII–XI) of the partial *O. arizonica* LIS genomic clone (fig. 3). Although some variation in the sizes of the corresponding introns in LIS genes from different species is evident (fig. 3), introns that are small (<250 nt) in one species tend to be small in other species as well. The exception is intron XI, which is large (675 nt) in *O. arizonica*. It is noteworthy that the first intron is quite large (440–912 nucleotides) in all LIS genes analyzed.

Comparison of LIS Sequences

The LIS protein sequences from *C. breweri* (LIS1 and LIS2), *C. concinna*, *O. arizonica*, and *A. thaliana* were aligned, and the percentage of identity was calculated for the region encompassing residues 476–795.
Comparison of LIS Sequences with Those of Other Terpene Synthases

The three-dimensional structure of a single plant terpene synthase polypeptide, the one encoded by the tobacco 5EAS gene, has been determined so far (Starks et al. 1997). The analysis of the three-dimensional structure of the 550-residue 5EAS protein (an LMS-type synthase) indicates that residues 36–230, encoded by exon 2 and most of exon 3, form a structure related to glycosylhydrolases that appears not to be involved in terpene biosynthesis. Residues 231–550, encoded by the end of exon 3 and exons 4–7, constitute the terpene synthase portion of the protein involved in initial recognition and the binding of the FPP substrate and Mg\(^{2+}\) cofactor in the active site centered around the DDXXD motif. Identical regions are proposed to be involved in catalysis in other LMS-type synthases (region A in fig. 4).

The LIS protein has significant sequence identity (34%) with 5EAS in the regions encoded by the end of exon 3 and all of exon 4 in the 5EAS gene (corresponding to most of exon 8 and all of exon 9 in LIS) (dark-shaded areas in fig. 4). CPS regions encoded by exons 4, 6, 7, and 10 of CPS retain the highest levels (29%–38%) of identity with the LIS sequences encoded by the corresponding regions (most of exon 1 as well as exons 3, 4, and 7) of LIS (light-shaded areas in fig. 4). The very beginning of the CPS gene encodes a cleavable transit peptide (C. breweri LIS1 numbering). This region corresponds to the C-terminal terpene synthase region of 5EAS and also coincides with the region that has the highest sequence similarity to other terpene synthases such as LMS and CAS. (It should be noted that C. breweri LIS2 encodes a shorter protein [815 residues] than does C. breweri LIS1 [870 residues], and the difference in length is due to earlier termination of the LIS2 protein). In this region, the C. concinna LIS protein is 97.5% identical to LIS1 from C. breweri. Other LIS protein sequences share identity to each other in the 67%–72% range, with the exception of A. thaliana LIS, which is only ~45% identical to the other LIS sequences. When compared over the entire length of the polypeptides, the A. thaliana LIS is ~40% identical to the Onagraceae LIS sequences, and the C. breweri LIS1 and C. concinna LIS proteins are 96% identical. The protein segments encoded by exons 3, 7, 9, and 11 have retained the highest levels of identity among these proteins, with the segment encoded by exon 9 (which includes the important motif DDXXD) being the most conserved (fig. 4).

Comparison of LIS Sequences with Those of Other Terpene Synthases

The three-dimensional structure of a single plant terpene synthase polypeptide, the one encoded by the tobacco 5EAS gene, has been determined so far (Starks et al. 1997). The analysis of the three-dimensional structure of the 550-residue 5EAS protein (an LMS-type synthase) indicates that residues 36–230, encoded by exon 2 and most of exon 3, form a structure related to glycosylhydrolases that appears not to be involved in terpene biosynthesis. Residues 231–550, encoded by the end of exon 3 and exons 4–7, constitute the terpene synthase portion of the protein involved in initial recognition and the binding of the FPP substrate and Mg\(^{2+}\) cofactor in the active site centered around the DDXXD motif. Identical regions are proposed to be involved in catalysis in other LMS-type synthases (region A in fig. 4).

The LIS protein has significant sequence identity (34%) with 5EAS in the regions encoded by the end of exon 3 and all of exon 4 in the 5EAS gene (corresponding to most of exon 8 and all of exon 9 in LIS) (dark-shaded areas in fig. 4), but not elsewhere. Homology modeling (SwissModel) predicts that this region of the LIS protein has a three-dimensional structure similar to that of the corresponding region of 5EAS (data not shown). In contrast, LIS shows the highest identity to CPS in the first part of the protein (~28% compared with 17%–22% for other terpene synthases) (region B in fig. 4). CPS regions encoded by exons 4, 6, 7, and 10 of CPS retain the highest levels (29%–38%) of identity with the LIS sequences encoded by the corresponding regions (most of exon 1 as well as exons 3, 4, and 7) of LIS (light-shaded areas in fig. 4).
that directs its protein to the plastid (Sun and Kamiya 1994; Aach, Bose, and Graeve 1995), but the LIS protein appears to have no cleavable N-terminus of more than eight amino acids (Dudareva et al. 1996). However, the sequence of the N-terminal 60 amino acids of LIS has the characteristics of a plastid-targeting sequence (Keegstra, Olsen, and Thog 1989), being rich in hydroxylated and basic residues.

Comparison of the Gene Structures of LIS with Those of Other Terpene Synthases

It has previously been reported that intron positions in LMS-type terpene cyclase genes are very similar (Back and Chappell 1995). For example, LMS from *Perilla frutescens*, 5EAS from *Nicotiniana tabacum*, and CAS from *Ricinus communis* (encoding a monoterpene, a sesquiterpene, and a diterpene cyclase, respectively) have introns in identical positions, although LMS and CAS have longer first exons which are thought to be encoded, in part, a transit peptide that directs these proteins to the plastids (Facchini and Chappell 1992; Mau and West 1994; Yuba et al. 1996). However, it is noteworthy that the LIS gene conforms to this gene structure only in the region of the LMS-type genes encoding the part of the protein that corresponds to the structure of the terpene synthase (region A in fig. 4). In this region, encoded by exons 9–12 in LIS, the LIS introns occur in the same positions as in the LMS-type terpene cyclase genes. No such correspondence is seen in the N-terminal parts of the protein. Instead, in this region (LIS exons 1–8), exon sizes and intron positions in LIS are essentially identical to those found in the region of CPS encompassed by exons 4–11.

**Discussion**

**Variation Among LIS Protein Sequences**

The levels of variation between LIS protein sequences in species within the same family but not within the same genus (65%–69% in the *Clarkia* vs. *O. arizonica* comparisons) and between LIS protein sequences in angiosperm species not in the same family (40%–45% in the *Oncorhynchus* vs. *A. thaliana* comparisons) are in the same range found for other terpene synthases. These observations and the similarity in gene structure and protein lend support to the identification of *C. breweri* LIS2 and the *O. arizonica* and the *A. thaliana* genes whose sequences are presented here as encoding LIS. However, it should be noted that since amino acid sequence identity of up to 80% by itself does not guarantee the same function in the plant terpene synthase family (Bohlmann, Steele, and Croteau 1997), only a functional study can establish the correct products of the reactions catalyzed by the enzymes encoded by these genes.

Comparisons of Linalool Synthase with Other Terpene Synthases, and the Origin of the LIS Gene

From sequence comparisons and examination of the three-dimensional structure of 5EAS, it was determined that the terpene synthases which employ the LMS-type catalytic mechanism contain the metal-cofactor-binding domain and most of the residues involved in substrate (GPP, FPP, or GGPP) binding in the C-terminal part of the protein. (However, SEAS also has a small N-terminal region [residues 1–35] involved in substrate binding [Starks et al. 1997]). CPS, which utilizes a different mechanism, has a substantially divergent sequence with the metal-cofactor-binding site present in the N-terminal half of the protein (but, since the three-dimensional structure of CPS has not been determined, the location of other residues essential for catalysis is not yet known). In LIS, which appears to employ a simplified LMS-type mechanism (Pichersky, Lewinsohn, and Croteau 1995), the metal-cofactor-binding motif DDDXXD is present in the second part of the protein, and the extensive sequence identity between LIS and other terpene cyclases centered around this motif suggests that the active site of LIS is present in this part of the protein.

We previously reported that the LIS protein shares sequence similarity with LMS-type synthases in its C-terminal half, and with CPS in its N-terminal half (Dudareva et al. 1996). A similar observation was made for ABS (Vogel et al. 1996), suggesting that both of these proteins are of composite origin. In fact, the sequences of a few other terpene synthases also show similarity to both LMS-type synthases and CPS in the respective C- and N-termini, for example, taxadiene synthase (TAS) (Wildung and Croteau 1996) and KS (Yamaguchi et al. 1996, 1998). Interestingly, all of these proteins are of similar size to LIS. Although ABS is a bifunctional enzyme which employs both a CPS-type mechanism and an LMS-type mechanism (Vogel et al. 1996), it is not clear why the LIS protein (or TAS and KS) should display similarities to CPS. Moreover, LIS and KS fall into one subfamily and TAS and ABS fall into another in the plant terpene synthase classification scheme of Croteau and co-workers (Bohlmann, Meyer-Gaun, and Croteau 1998). Prior to the present report, none of the structures of the genes encoding these proteins had been reported. Thus, the segmental similarity in the coding sequences between them and the CPS- and LMS-type synthases was difficult to interpret. For example, their similarity to CPS in the N-terminus could have been due to preferential sequence conservation of a region that may have been present in all ancestral terpene synthases but whose function is conserved only in these enzymes (and not in the other terpene synthases), rather than indicating a more recent common origin.

We now show that the sequence similarity of LIS and CPS in the N-terminal part of the protein is most likely due to a relatively recent common ancestry of LMS (for the first half of the gene) with CPS, as evidenced by the similarity in structure of the genes encoding them. Specifically, introns 1–8 in LIS occur in positions virtually identical to introns 4–11 in CPS, respectively delineating seven pairs of exons of nearly identical sizes (within each pair). However, the rest of the LIS gene shows no structural similarity to CPS. Instead, in the second half of the LIS gene (the region encoded by exons 9–12), introns 8–11 occur in the same positions as introns 3–6 in LMS, 5EAS, and CAS. This observation
suggests that LIS is a composite gene resulting from a discrete recombination event between the first half of a CPS-like gene and the second half of an LMS-type gene. The alternative hypothesis, that LIS gene structure represents the ancestral condition in the terpene synthase family, is unlikely because the LIS protein occupies a nonbasal position in the phylogenetic tree of this family (Bohlmann, Meyer-Gauen, and Croteau 1998) and also because CPS encodes an enzyme in the pathway leading to the production of the gibberellin-type plant hormones which are essential for all higher plants. Thus, CPS is likely to have preceded LIS.

Repeated Evolution

The most recent phylogenetic analysis of the terpene synthase family of proteins, based on comparisons of amino acid sequences only (Bohlmann, Meyer-Gauen, and Croteau 1998), placed all the LMS-type terpene synthases, with the possible exception of LIS and KS, in a monophyletic clade. Although intron positions in most of the genes encoding terpene synthases are not yet known, in the few cases in which gene structure was determined in the LMS-type terpene synthase genes, the intron positions were the same (Back and Chappell 1995). These observations suggest that all of these enzymes originated from a single ancestral gene which has undergone successive cycles of duplications and divergence.

However, within this group, the phylogenetic relationships are not clear. For example, although a priori it would seem most likely that specific enzymes such as LMS have evolved only once in plants, evidence has been presented that the protein sequence of LMS in a gymnosperm species, grand fir (Abies grandis), is more similar to the sequences of other terpene synthases in that same species (64% identical to grand fir pinene cyclase and myrcene synthase) than it is to LMS protein sequences from angiosperms (e.g., 28% identical to LMSs from mint) (Yuba et al. 1996; Bohlmann, Steele, and Croteau 1997). Moreover, grand fir LMS, as well as other terpene synthases from this gymnosperm species, require K+ for optimal activity, while the corresponding enzymes from angiosperms do not (Savage, Hatch, and Croteau 1994). Bohlmann, Steele, and Croteau (1997) initially suggested that perhaps overall lower rates of protein sequence evolution in gymnosperms explain the lower divergence among different gymnosperm terpene synthase sequences. However, their more comprehensive phylogenetic treatment of the plant terpene synthase family clearly indicated that the LMS enzymes in gymnosperms and angiosperms do not form a monophyletic clade distinct from all other LMS-type terpene synthases (Bohlmann, Meyer-Gauen, and Croteau 1998), suggesting that specific LMS enzymatic activities evolved in plants more than once.

Such repeated evolution could be considered “convergent evolution” if LMS genes from gymnosperms and angiosperms evolved from unrelated ancestral genes. However, the sequence similarities that do exist between LMS proteins from gymnosperms and those from angiosperms suggest that although these enzymes evolved independently, each must have evolved from a member of the terpene synthase family, and, more specifically, from within the clade of the LMS-type terpene synthases. The universal presence of a number of related enzymes catalyzing the formation of related terpene compounds which are essential for plant viability (gibberellins, brassinosteroids, etc.) may provide a pool from which new enzymes, catalyzing the formation of “specialized” terpenes, can evolve, sometimes more than once. This may explain the presence of the same specialized terpene in taxonomically diverse species of plants when closely related species in each group do not make such a compound. The evidence presented in this paper shows that in addition to duplication and divergence, additional terpene synthases may be created by a process of recombination between different terpene synthase genes, or “domain swapping” (Gilbert 1980; Doolittle 1995). These two processes—duplication followed by divergence and duplications followed by domain swapping—would increase in frequency as the number of related terpene synthase genes in the genome increases and could explain the extraordinarily large number of diverse terpene synthases found in present-day plants.

Acknowledgments

This research was funded by grant IBN-9417582 from the National Science Foundation to E.P. We thank Dr. R. A. Raguso for collecting the O. arizonica plants and Dr. Peter Kaufman for the use of his facilities for growing the O. arizonica plants.

LITERATURE CITED


Julian P. Adams, reviewing editor

Accepted July 27, 1998