Genomic Organization of Plant Terpene Synthases and Molecular Evolutionary Implications

Susan C. Trapp and Rodney B. Croteau

Institute of Biological Chemistry, Washington State University, Pullman, Washington 99164-6340

Manuscript received November 30, 2000
Accepted for publication March 1, 2001

ABSTRACT

Terpenoids are the largest, most diverse class of plant natural products and they play numerous functional roles in primary metabolism and in ecological interactions. The first committed step in the formation of the various terpenoid classes is the transformation of the prenyl diphosphate precursors, geranyl diphosphate, farnesyl diphosphate, and geranylgeranyl diphosphate, to the parent structures of each type catalyzed by the respective monoterpene (C_{10}), sesquiterpene (C_{15}), and diterpene synthases (C_{20}). Over 30 cDNAs encoding plant terpenoid synthases involved in primary and secondary metabolism have been cloned and characterized. Here we describe the isolation and analysis of six genomic clones encoding terpene synthases of conifers, [(−)-pinene (C_{10}), (−)-limonene (C_{10}), (*E*)-α-bisabolene (C_{15}), δ-selinene (C_{15}), and abietadiene synthase (C_{20}) from *Abies grandis* and taxadiene synthase (C_{20}) from *Taxus brevifolia*], all of which are involved in natural products biosynthesis. Genome organization (intron number, size, placement and phase, and exon size) of these gymnosperm terpene synthases was compared to eight previously characterized angiosperm terpene synthase genes and to six putative terpene synthase genomic sequences from *Arabidopsis thaliana*. Three distinct classes of terpene synthase genes were discerned, from which assumed patterns of sequential intron loss and the loss of an unusual internal sequence element suggest that the ancestral terpenoid synthase gene resembled a contemporary conifer diterpene synthase gene in containing at least 12 introns and 13 exons of conserved size. A model presented for the evolutionary history of plant terpene synthases suggests that this superfamily of genes responsible for natural products biosynthesis derived from terpene synthase genes involved in primary metabolism by duplication and divergence in structural and functional specialization. This novel molecular evolutionary approach focused on genes of secondary metabolism may have broad implications for the origins of natural products and for plant phylogenetics in general.

THE terpenoids compose the largest and most diverse family of natural products. Of the more than 30,000 individual terpenoids now identified (Buckingham 1998), at least half are synthesized by plants. A relatively small, but quantitatively significant, number of terpenoids are involved in primary plant metabolism including, for example, the phytol side chain of chlorophyll, the carotenoid pigments, the phytosterols of cellular membranes, and the gibberellin plant hormones. However, the vast majority of terpenoids are classified as secondary metabolites, compounds not required for plant growth and development but presumed to have an ecological function in communication or defense (Harborne 1991). Mixtures of terpenoids, such as the aromatic essential oils, turpentines, and resins, form the basis of a range of a commercially useful products (Zinkel and Russell 1989; Dawson 1994), and several terpenoids are of pharmacological significance, including the monoterpenoid (C_{10}) dietary anticarcinogen limonene (Crowell and Gould 1994), the sesquiterpenoid (C_{15}) antimalarial artemisinin (Van Geldre *et al.* 1997), and the diterpenoid anticancer drug Taxol (Holmes *et al.* 1995; Figure 1).

All terpenoids are derived from isopentenyl diphosphate (Figure 2). In plants, this central precursor is synthesized in the cytosol via the classical acetate/mevalonate pathway (Qureshi and Porter 1981; Newman and Chappell 1999), by which the sesquiterpenes (C_{15}) and triterpenes (C_{30}) are formed, and in plastids via the alternative, pyruvate/glyceraldehyde-3-phosphate pathway (Eisenreich *et al.* 1998; Lichtenthaler 1999), by which the monoterpenes (C_{10}), diterpenes (C_{20}), and tetraterpenes (C_{40}) are formed. Following the isomerization of isopentenyl diphosphate to dimethylallyl diphosphate, by the action of isopentenyl diphosphate isomerase, the latter is condensed with one, two, or three units of isopentenyl diphosphate, by the action of prenyltransferases, to give geranyl diphosphate (C_{10}), farnesyl diphosphate (C_{15}), and geranylgeranyl diphosphate (C_{20}), respectively (Ramos-Valdivia *et al.* 1997; Ogura and Koyama 1998; Koyama and Ogura 1999; Figure 2). These three acyclic prenyl diphosphates serve as the immediate precursors of the corresponding monoterpenoid (C_{10}), sesquiterpenoid (C_{15}), and diterpenoid

Corresponding author: Rodney Croteau, Institute of Biological Chemistry, Washington State University, Pullman, WA 99164-6340.
E-mail: croteau@mail.wsu.edu

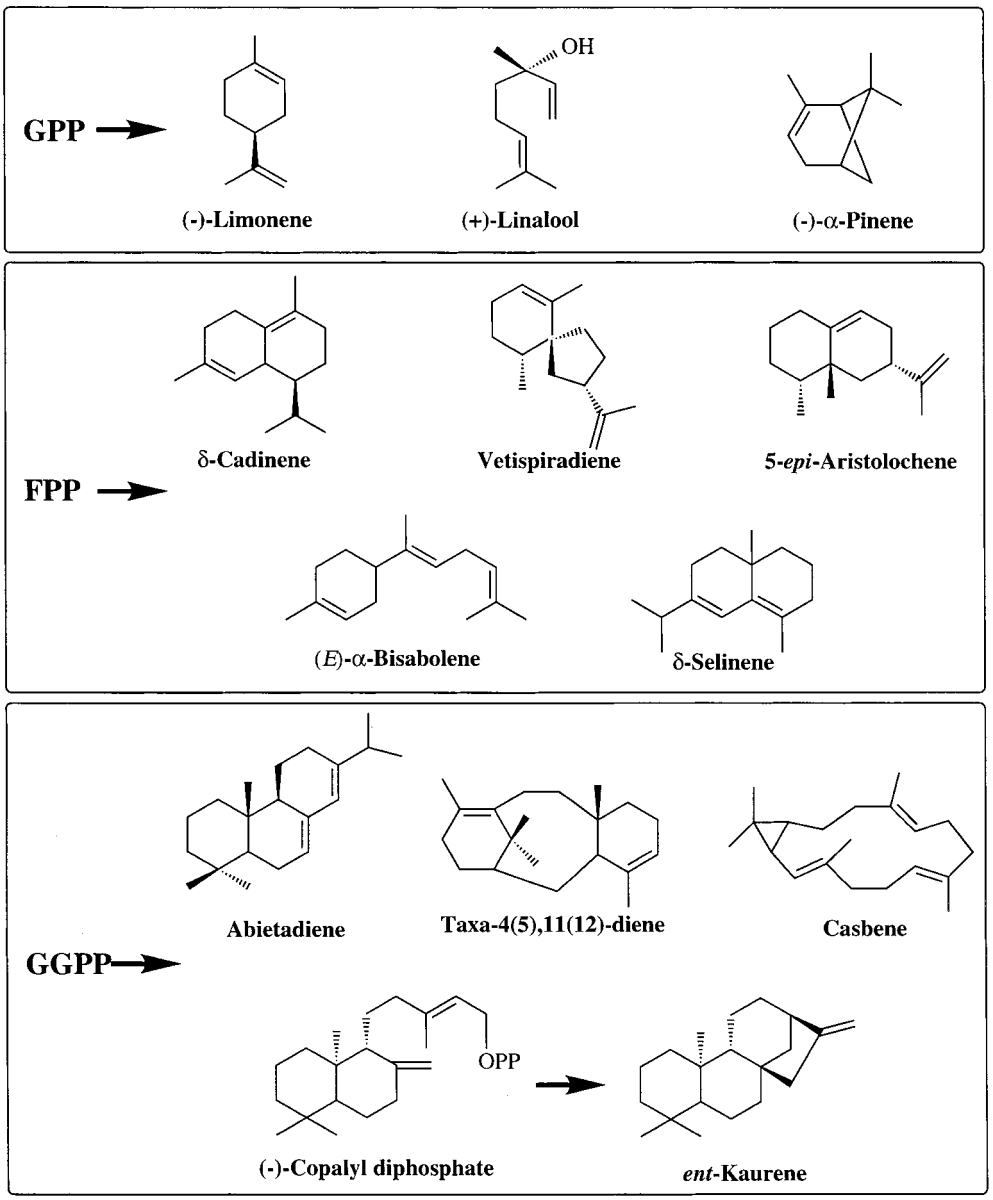


FIGURE 1.—Representative terpenoids biosynthesized by plants. Monoterpenes, sesquiterpenes, and diterpenes are derived from the prenyl diphosphate substrates geranyl diphosphate (GPP), farnesyl diphosphate (FPP), and geranylgeranyl diphosphate (GGPP), respectively, and are produced in both angiosperms and gymnosperms. (−)-Copalyl diphosphate and *ent*-kaurene are sequential intermediates in the biosynthesis of gibberellin plant growth hormones. Taxa-4(5), 11(12)-diene is the first dedicated intermediate in the biosynthesis of Taxol.

(C_{20}) classes, to which they are converted by a very large group of enzymes called the terpene (terpenoid) synthases. These enzymes are often referred to as terpene cyclases, since the products of the reactions are most often cyclic.

A large number of terpenoid synthases of the monoterpene (CROTEAU 1987; WISE and CROTEAU 1999), sesquiterpene (CANE 1990, 1999b), and diterpene (WEST 1981; MACMILLAN and BEALE 1999) series have been isolated from both plant and microbial sources, and these catalysts have been described in some detail. All terpenoid synthases are very similar in physical and chemical properties, for example, in requiring a divalent metal ion as the only cofactor for catalysis, and all operate by unusual electrophilic reaction mechanisms. In this regard, the terpenoid synthases resemble the prenyltransferases; however, it is the tremendous range of possible variations in the carbocationic reactions (cyclizations, hydride shifts, rearrangements, and termination steps) catalyzed by the terpenoid synthases that sets them apart as a unique enzyme class. Indeed, it is these variations on a common mechanistic theme that permit the production of essentially all chemically feasible skeletal types, isomers, and derivatives that form the foundation for the great diversity of terpenoid structures.

Several groups have suggested that plant terpene synthases share a common evolutionary origin based upon their similar reaction mechanism and conserved structural and sequence characteristics, including amino acid sequence homology, conserved sequence motifs, intron number, and exon size (MAU and WEST 1994; BACK and CHAPPELL 1995; BOHLMANN *et al.* 1998b; CSEKE *et al.* 1998). Sequence comparison between the first three plant terpenoid synthase genes isolated [a monoterpene

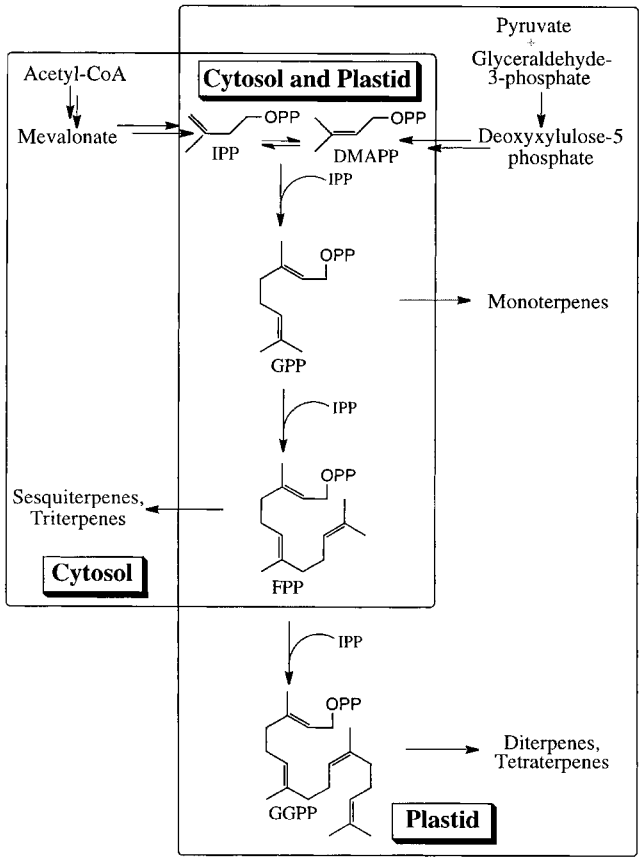


FIGURE 2.—Overview of terpenoid biosynthesis in plants. The intracellular compartmentalization of the mevalonate and mevalonate-independent pathways for the production of isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP), and of the derived terpenoids, is illustrated. The cytosolic pool of IPP, which serves as a precursor of farnesyl diphosphate (FPP) and, ultimately, the sesquiterpenes and triterpenes, is derived from mevalonic acid (left). The plastidial pool of IPP is derived from the glycolytic intermediates pyruvate and glyceraldehyde-3-phosphate and provides the precursor of geranyl diphosphate (GPP) and geranylgeranyl diphosphate (GGPP) and, ultimately, the monoterpenes, diterpenes, and tetraterpenes (right). Reactions common to both pathways are enclosed by both boxes.

cyclase limonene synthase (COLBY *et al.* 1993), a sesquiterpene cyclase *epi*-aristolochene synthase (FACCHINI and CHAPPELL 1992), and a diterpene cyclase casbene synthase (MAU and WEST 1994)] gave clear indication that these genes, from phylogenetically distant plant species, were related, a conclusion supported by very limited genomic analysis of intron number and location (MAU and WEST 1994; BACK and CHAPPELL 1995; CHAPPELL 1995a,b). More recently, phylogenetic analysis of the deduced amino acid sequences of 33 terpenoid synthases from angiosperms and gymnosperms allowed recognition of six terpenoid synthase (Tps) gene subfamilies on the basis of clades (BOHLMANN *et al.* 1998b; see Figure 6B). The majority of terpene synthases analyzed produce secondary metabolites and are classified into three subfamilies, Tpsa (sesquiterpene and diterpene synthases from angiosperms), Tpsb (monoterpene synthases from angiosperms of the Lamiaceae), and Tpsd (11 gymnosperm monoterpene, sesquiterpene, and diterpene synthases). The other three subfamilies, Tpsc, Tpse, and Tpsf, are represented by the single angiosperm terpene synthase types copalyl diphosphate synthase, kaurene synthase, and linalool synthase, respectively. The first two are diterpene synthases involved in early steps of gibberellin biosynthesis (MACMILLAN and BEALE 1999). These two Tps subfamilies are grouped into a single clade and are involved in primary metabolism, which suggests that the bifurcation of terpenoid synthases of primary and secondary metabolism occurred before the separation of angiosperms and gymnosperms (BOHLMANN *et al.* 1998b). A detailed analysis of the latter monoterpene synthase, linalool synthase from Clarkia representing Tpsf, suggests that this is a composite gene of recent origin (CSEKE *et al.* 1998).

In this article, we address the evolutionary relationships among plant monoterpene, sesquiterpene, and diterpene synthase genes from gymnosperms and angiosperms by examination of gene architecture. The genomic sequencing and organization of six new terpenoid synthase genes from gymnosperms are described, and these sequences are compared to those of eight defined terpenoid synthases and six putative sequences from angiosperms in the databases. A model for the evolutionary history of plant terpene synthases from primary to secondary metabolism is presented based upon the evaluation of intron number, size, placement and phase, and exon size, and upon the assumption that descent was accompanied by intron loss. This model has allowed a more refined analysis of structure and classification of these genes, from which their evolutionary origin by divergence from a common ancestor, with a progression of sequence loss, can be inferred.

MATERIALS AND METHODS

Materials and general procedures: Pacific yew (*Taxus brevifolia*) saplings (2–4 yr old, from the Weyerhaeuser Research Center, Centralia, WA) were raised in a greenhouse as described and were previously verified to produce Taxol (KOEPP *et al.* 1995). A mature grand fir (*Abies grandis*) tree (~50 yr old, from the University of Idaho Arboretum, Moscow, ID) was analyzed by a standard protocol (LEWINSOHN *et al.* 1993) to ensure an oleoresin composition within the typical range (KATOH and CROTEAU 1998). Immature needles from both species were frozen in liquid N_2 after harvest and used immediately, or stored at −80° until use.

Restriction enzymes and T4 DNA ligase were obtained from Promega (Madison, WI). Plasmid miniprep isolations were performed by a modified alkaline lysis procedure (SAMBROOK *et al.* 1989) using the Miniprep Express Matrix (BIO 101, La Jolla, CA) or the QIAGEN Miniprep kit (QIAGEN, Valencia, CA). PCR amplifications employed PCR SUPERMIX high fidelity, *Taq* DNA polymerase, or ELONGASE SUPERMIX (Life Technologies, Gaithersburg, MD) according to the manufacturer's instructions. The QiaQuick gel purification kit (QIAGEN) or the Geneclean kit (BIO 101) was used for all steps

requiring purification of DNA fragments in excised agarose gel bands in accordance with the manufacturer's instructions. For plasmid clones generated by PCR, the TOPO-TA cloning kit, including the pCR2.1-TOPO vector and *Escherichia coli* One Shot TOP10F′ competent cells (Invitrogen, San Diego), were used according to the manufacturer's protocols. For the Tb*ggtax* gene plasmid clone, pBluescript II KS(−) and *E. coli* XL2-Blue cells (Stratagene, La Jolla, CA) were employed. All custom primers were synthesized by Life Technologies. The ABI DyeDeoxy Terminator cycle sequencing kit and FS*Taq* were used for DNA sequencing with either an ABI PRISM 373 or 377 system (PE Applied Biosystems, Foster City, CA).

DNA Isolation from *A. grandis* and *T. brevifolia*: Genomic DNA from *A. grandis* and *T. brevifolia* needles was isolated using a modification of the procedure of Patterson *et al.* (1993). Frozen tissue (1–2 g) was pulverized to a fine powder in a chilled mortar and pestle and transferred to a 50-ml conical tube on ice containing 10 ml/g tissue of buffer A [100 mM Tris-Cl (pH 8.0) containing 350 mM glucose, 5 mM EDTA, 2% (w/v) polyvinylpyrrolidone (PVP40), 0.1% (w/v) diethyldithiocarbamate, and (added immediately before use) 0.1% (w/v) ascorbic acid and 0.2% (v/v) mercaptoethanol]. The tubes were flushed with N_2, sealed, centrifuged for 20 min at 2700 × *g* at 4°, and the supernatant was discarded. Pelleted material was resuspended in 10 ml of buffer B [100 mM Tris-Cl (pH 8.0) containing 1.4 M NaCl, 20 mM EDTA, 2% (w/v) polyvinylpyrrolidone (PVP40), 0.1% (w/v) diethyldithiocarbamate, and (added immediately before use) 0.1% (w/v) ascorbic acid and 0.2% (v/v) mercaptoethanol], and 125 μg/ml of RNase A was added to the mixed sample, which was incubated for 2–5 hr at 50°.

To purify the DNA, 10 ml of chloroform:isoamyl alcohol (24:1, v/v) was thoroughly mixed into the sample (by inverting), which was centrifuged for 5 min at 2700 × *g* and the upper phase was transferred to a new tube. DNA was precipitated by gently mixing in 0.6 vol of isopropanol, and the DNA was collected by spooling with a glass hook. Final purification was achieved using a QIAGEN 500-Tip column by following the manufacturer's genomic tip protocol for plant DNA isolation (QIAGEN). Thus, the isolated DNA was suspended in 5 ml of 1 M NaCl by heating at 62° for 30 min. After cooling to room temperature, 1.7 ml of autoclaved deionized water and 3.3 ml of QIAGEN's QBT buffer were added, and the DNA suspension was loaded onto the column and eluted as described.

Genomic cloning of terpene synthases: Genomic clones (see Table 2) corresponding to previously described cDNAs ag1 (Bohlmann *et al.* 1998a), ag3, ag10 (Bohlmann *et al.* 1997), ag4 (Steele *et al.* 1998), ag22 (Stofer Vogel *et al.* 1996) from grand fir, and tb1 (Wildung and Croteau 1996) from Pacific yew were obtained by PCR amplification using primers designed to the sequence termini (see Table 3) with genomic DNA from the appropriate conifer species as target. The gel-purified amplicons were sequenced directly to verify correspondence of genomic sequences to the previously defined cDNAs. Amplicons corresponding to each grand fir terpene synthase gene were ligated into pCR2.1-TOPO and transformed into *E. coli* One Shot TOP10F′ competent cells using the TOPO-TA cloning kit. The taxadiene synthase gene corresponding to Tb*ggtax* cDNA from yew was digested with *Eco*RI and *Xba*I, the fragment was gel purified and ligated into pBluescript similarly digested, and the plasmid then transformed into *E. coli* XL2-Blue competent cells. Following plasmid preparation, pAGg1-1, pAGg1-11A, pAGg3-37, pAGg4-34, pAGg10-1, and pAGg22-4 were digested with *Eco*RI (pTBtaxg-1-1 was digested with *Eco*RI and *Xba*I) and subjected to gel electrophoresis to verify insert size (see Table 2).

Polymerase chain reactions: The amplification reactions contained either ELONGASE SUPERMIX or SUPERMIX high fidelity DNA polymerase (Life Technologies), 6.4 pmol of each primer and 10–100 ng of genomic DNA in a 50-μl volume. General reaction conditions were 30 sec initial denaturation at 94° (hotstart), followed by 35 cycles, each of 30 sec at 94°, 30 sec at 50–52°, and 4–5 min at 68°, with final extension for 7 min at 68°. Only Ag*fEαbis* could not be successfully amplified by this approach and so was divided into two segments, AG1-1 (the 3′-terminal half) and AG1-11A (5′-terminal half). AG1-1 was successfully amplified by the above conditions. A modified touchdown PCR procedure (Don *et al.* 1991) utilizing the ELONGASE polymerase mix was used to amplify the AG1-11A segment. The conditions for the first round of PCR were 30 sec initial denaturation at 94° (hotstart), followed by 7 cycles, each of 30 sec at 94°, 30 sec at 60°, and 3 min at 67°, followed by 30 additional cycles, each of 30 sec at 94°, 30 sec at 57°, and 3 min at 67°, with final extension for 7 min at 67°. The resulting gel-purified amplicon was used as a template for the second round of Touchdown PCR employing 30 sec initial denaturation at 94°, followed by 5 cycles, each of 30 sec at 94°, 30 sec at 60°, and 3 min at 68°, followed by 35 cycles, each of 30 sec at 94°, 30 sec at 57°, and 3 min at 68°, with final extension for 7 min at 68°.

Genomic sequencing: Genomic sequencing was carried out on both strands using the plasmid clones or the original PCR amplicons directly as template. In addition to the original sequencing primers, nondegenerate primers, 18–21 nucleotides (nt) in length, were designed to span distances of 250–300 nt, with overlaps as necessary to fill gaps and resolve uncertainties. The Lasergene programs EDITSEQ and SEQMAN (DNASTAR, Madison, WI) were utilized for basic editing and assembly of fragment sequences into a finished contig, respectively. The Lasergene MEGALIGN program was used for routine comparison of multiple amino acid sequences (Clustal method) and for pairwise comparisons (Lipman-Pearson method). MEGALIGN was also used for the final multiple protein alignment (see WebFigure 1 at http://ibc.wsu.edu/faculty/croteau.html). BLAST programs (Altschul *et al.* 1990, 1997) were utilized to search for other defined and putative plant terpene synthase genes in the database using the characterized gymnosperm terpene synthase amino acid sequences. The acquired sequences were downloaded to EDITSEQ for editing and transferred to MAPDRAW and MEGALIGN programs for analysis.

Intron identification and analysis: For previously characterized terpene synthase sequences, the general placement of introns was determined by MEGALIGN pairwise comparison of cDNA nucleotide sequence to genomic sequence. The genomic sequence of (−)-limonene synthase (Ml*g-lim*) from *Mentha longifolia* was kindly provided by T. Davis, University of New Hampshire, and the genomic sequence of casbene synthase from castor bean was obtained from C. West (University of California, Los Angeles). In the cases of vetispiradiene synthase and Ml*g-lim*, one or both intron borders had been identified previously; however, the gene had not been sequenced entirely (J. Chappell and T. Davis, personal communications). For δ-cadinene synthase, intron placement was determined by comparison of the δ-cadinene synthase genomic sequence (Ga*f*δ*cad1b*) from *Gossypium arboreum* (Chen *et al.* 1996) to the cDNA sequences of Ga*f*δ*cad1b* and Gh*f*δ*cad1* from *G. hirsutum* (Davis *et al.* 1998; see Table 1). For the putatively identified terpene synthases of *Arabidopsis thaliana*, the exon/intron borders specified by the genome project were reidentified and reverified by MEGALIGN pairwise analysis of the nucleotide sequence in the database to a corresponding known angiosperm terpene synthase of similar type (*i.e.*, mono-, sesqui-, or diterpene synthase); intron borders in some instances were corrected on the basis of intron phase and

exon size pattern established for other defined terpene synthases (see Figure 4), and new predictions were used for subsequent analyses. The 5′- and 3′-intron splice sites were determined by comparing sequence data in the intron regions to published consensus sequences (Hanley and Schuler 1988; Turner 1993). All sequences were entered into the Lasergene program MAPDRAW, which displays both the nucleotide and amino acid translation together, to define intron phase by determining at which nucleotide within a codon the coding sequence was interrupted. If the last coded amino acid of an exon was interrupted by an intron (phase 1 or phase 2), then this amino acid was defined as the first amino acid of the next exon for the purpose of describing exon size.

Gene organization comparison: The architecture of all terpene synthase genes was determined by manual analysis of the coding region, including evaluation of intron phase, intron placement and number (1–14 when extant), exon number (1–15 when extant) and exon size (amino acid), and conserved amino acids and/or motifs. The architecture of each terpene synthase gene (known and putative) was summarized in table format and a physical map (exons and introns) of each gene was created. Terpene synthase gene architectural maps were aligned by hand with the assistance of a computer drawing program (Adobe Illustrator 6.0) and then compared. The classification of the terpene synthase genes into class I, II, or III types is based upon grouping by physical similarities of gene architectures.

Phylogenetic analysis: The hypothesis for the phylogenetic relationship among the defined terpenoid synthase gene sequences was generated by evaluating observed architectural patterns of intron number, the presence or absence of the conifer diterpene internal sequence (CDIS) domain, exon number and size, and intron phase conservation. The evolutionary history of the terpene synthases was proposed after the most parsimonious explanation (the fewest steps to account for intron and CDIS domain loss) was schematically diagrammed. Although putative terpenoid synthases were utilized to evaluate initial patterns of exon size, intron number, placement, and loss, and thus gene classification, to affirm the observed patterns, only the genomic sequences of defined terpene synthases were utilized in the phylogenetic analyses. A distance tree based upon an algorithm utilizing a distance substitution (amino acid) method (within the MEGALIGN module) was used to produce the phylogenetic tree model in Figure 6A. The previously published phylogenetic tree (Bohlmann *et al.* 1998b), prepared by the neighbor-joining method, is presented for comparison in Figure 6B. Proposed losses of introns and the CDIS domain were charted upon the trees (Figure 6, A and B) by hand with the aid of a computer drawing program (Adobe Illustrator 6). The nucleotide coding sequences (excluding introns) and deduced protein sequences of the terpene synthases were subsequently analyzed by PAUP and MacClade programs (courtesy of Pamela Soltis, Washington State University) to further evaluate possible phylogenetic relationships in attempts to recreate a tree (similar to the distance tree) by rigorous analyses (data not shown).

RESULTS

A previous phylogenetic reconstruction based upon amino acid sequence comparison of 12 gymnosperm and 21 angiosperm terpene synthases, representing 18 different species (Bohlmann *et al.* 1998b), characterized the general structural features of these plant enzymes and indicated that conifer monoterpene, sesquiterpene, and diterpene synthases are more closely related to each other than they are to their respective, mechanism-based, counterparts of angiosperm origin (Bohlmann *et al.* 1998b). Because of this apparent lack of structure-function correlation, it is not yet possible to predict the catalytic capability of a terpene synthase solely on the basis of sequence relatedness. Thus, for example, monoterpene synthases sharing 70–90% identity at the amino acid level can catalyze biochemically distinct reactions, while synthases sharing <30% amino acid identity can catalyze the same cyclization reaction (Bohlmann *et al.* 1997).

To extend the previous analysis, genomic structures of several terpene synthases were determined in order to refine the phylogenetic relationships among and between these gymnosperm and angiosperm genes. Although gene sequences for several angiosperm terpene synthases were found in the public database (Table 1), no genomic sequences encoding terpene synthases from gymnosperms could be identified. Therefore, we determined the genomic (gDNA) sequences corresponding to 6 (Ag**gg***abi*, Ag*f*Eα*bis*, Ag*g-pin1*, Ag*f*δ*sel1*, Ag*g-lim*, Tb*ggtax*) of 12 previously reported conifer terpene synthase cDNAs (Table 1); 5 of these genes were isolated from grand fir (*A. grandis*) and the sixth, Tb*ggtax*, was isolated from Pacific yew (*T. brevifolia*). This selection of genes represents constitutive and inducible terpenoid synthases from each class (monoterpene, sesquiterpene, and diterpene). Sequence alignment of each cDNA with the corresponding gDNA, including putative terpene synthases from Arabidopsis, established exon and intron boundaries, exon and intron sizes, and intron placement; generic dicot plant 5′- and 3′-splice site consensus sequences (5′ NAG$^{\nabla}$GTAAGW WWW; 3′ YAG$^{\nabla}$) were used to define specific boundaries (Hanley and Schuler 1988; Turner 1993). These analyses revealed a distinct pattern of intron phase for each intron throughout the entire Tps gene family, as summarized below.

A wide range of nomenclatures has been applied to the terpenoid synthases, none of which is systematic. Here we use a unified and specific nomenclature system in which the Latin binomial (two letters), substrate, (one- to four-letter abbreviation), and product (three letters) are specified. Thus, ag22, the original cDNA designation for abietadiene synthase from *A. grandis* (a Tpsd subfamily member), becomes Ag**gg**ABI for the protein and Ag*ggabi* for the gene, with the remaining conifer synthases (and other selected genes) described accordingly (Table 1).

Terpene synthase genomic sequences from *A. grandis* and *T. brevifolia*: To isolate the genes encoding abietadiene, (*E*)-α-bisabolene, (−)-pinene, (−)-limonene, δ-selinene, and taxadiene synthases (Ag**gg***abi*, Ag*f*Eα*bis*, Ag*g-pin1*, Ag*f*δ*sel1*, *Agg-lim*, and Tb*ggtax*, respectively), PCR was performed with nondegenerate primers (Table 2) designed to the 5′ and 3′ termini of the coding region of the corresponding cDNAs (Table 3) using the

TABLE 1

Conifer and other selected terpene synthases

Products	Terpene synthase name				GenBank accession no.		Reference		Region on chromosome
	Species	Former gene	Enzyme[b]	cDNA/genomic[b]	cDNA	gDNA	cDNA	gDNA	
Abietadiene	*A. grandis*	ag22	AgggABI	Agg*gabi*	U50768	AF326516	Stofer Vogel *et al.* (1996)	Trapp and Croteau[tc]	—
(E)-α-Bisabolene	*A. grandis*	ag1	AgfEαBIS	Agf*Eαbis*	AF006195	AF326515	Bohlmann *et al.* (1998a)	Trapp and Croteau[tc]	—
(−)-Camphene	*A. grandis*	ag6	Agg-CAM	Agg-*cam*	U87910	—	Bohlmann *et al.* (1999)	—	—
γ-Humulene	*A. grandis*	ag5	AgfγHUM	Agf*γhum*	U92267	—	Steele *et al.* (1998)	—	—
(−)-Limonene	*A. grandis*	ag10	Agg-LIM1	Agg-*lim*	AF006193	AF326518	Bohlmann *et al.* (1997)	Trapp and Croteau[tc]	—
Myrcene	*A. grandis*	ag2	AggMYR	Agg*myr*	U87908	—	Bohlmann *et al.* (1997)	—	—
(−)-α/β-Pinene	*A. grandis*	ag3	Agg-PIN1	Agg-*pin1*	U87909	AF326517	Bohlmann *et al.* (1997)	Trapp and Croteau[tc]	—
(−)-α-Pinene/ (−)-limonene	*A. grandis*	ag11	Agg-PIN2	Agg-*pin2*	AF139207	—	Bohlmann *et al.* (1999)	—	—
(−)-β-Phellandrene	*A. grandis*	ag8	Agg-βPHE	Agg-β*phe*	AF139205	—	Bohlmann *et al.* (1999)	—	—
δ-Selinene	*A. grandis*	ag4	AgfδSEL1	Agfδ*sel1*	U92266	AF326513	Steele *et al.* (1998)	Trapp and Croteau[tc]	—
			AgfδSEL2	Agfδ*sel2*		AF326514			
Taxadiene	*T. brevifolia*	Tb1	TbggTAX	Tbgg*tax*	U48796	AF326519	Wildung and Croteau (1996)	Trapp and Croteau[tc]	—
Terpinolene	*A. grandis*	ag9	AggTEO	Agg*teo*	AF139206	—	Bohlmann *et al.* (1999)	—	—
5-*epi*-Aristolochene	*Nicotiana tabacum*	TEAS3	NtfeARI3	Nt*feari3*	L04680	L04680	Facchini and Chappell (1992)	Facchini and Chappell (1992)	—
		TEAS4	NtfeARI4	Nt*feari4*	L04680	L04680			
5-*epi*-Aristolochene[b]	*A. thaliana*	—	AteARI	At*eari*	—	AL022224	—	Bevan *et al.*[ds]	Chromosome 4 BAC F1C12 (ESSA) nt 44054–38820
δ-Cadinene	*G. arboreum*	CAD1-A	GafδCAD1A	Gaf*δcad1a*	X96429	Y18484	Chen *et al.* (1996)	Liang *et al.*[ds]	—
δ-Cadinene	*G. hirsutum*	CAD1-A	GhfδCAD1	Ghf*δcad1*	U88318	—	Davis *et al.* (1998)	—	—
δ-Cadinene	*G. arboreum*	gCAD1-B	GafδCAD1B	Gaf*δcad1b*		X95323	—	Chen *et al.*[ds]	—
Cadinene[p]	*A. thaliana*	—	AtCAD	At*cad*	—	AL022224	—	Bevan *et al.*[ds]	Chromosome 4 BAC F1C12 (ESSA) nt 44054–38820
Casbene	*Ricinus communis*	cas	RcggCAS	Rc*ggcas*	L32134	NA	Mau and West (1994)	West[pc]	—
(−)-Copalyl diphosphate[a]	*A. thaliana*	GA1	Atgg-COPP1	Atgg-*copp1*	U11034 —	NA AC004044[p]	Sun and Kamiya (1994)	Sun *et al.* (1992) — Bastide *et al.*[ds, c]	Chromosome 4 (Top) BAC T5J8 nt 34971–41856
ent-Kaurene[a]	*A. thaliana*	GA2	Atgg-KAU	Atgg-*kau*	AF034774	AC007202	Yamaguchi *et al.* (1998)	Vysotskaia *et al.*[ds, c]	Chromosome 1 BAC T8K14 nt 43552–47420
(−)-Limonene	*Perilla frutescens*	PFLC1	Pfg-LIM1	Pfg-*lim1*	D49368	AB005744	Yuba *et al.* (1996)	Tsubouchi[ds]	—
(−)-Limonene	*Mentha spicata*	LMS	Msg-LIM	Msg-*lim*	L13459	—	Colby *et al.* (1993)	—	—
(−)-Limonene	*M. longifolia*	LMS	Mlg-LIM	Mlg-*lim*	AF175323	—	Crock and Croteau[ds, c]	Jones and Davis[ps]	—
Limonene[p, i]	*A. thaliana*	—	AtLIMA1	At*lima1*	—	Z97341	—	Bevan *et al.*[ps]	Chromosome 4 CF6 (ESSA I) nt 164983–170505
			AtLIMA2	At*lima2*					

(*continued*)

TABLE 1

(Continued)

Products	Terpene synthase name				GenBank accession no.		Reference		Region on chromosome
	Species	Former gene	Enzyme[b]	cDNA/genomic[b]	cDNA	gDNA	cDNA	gDNA	
Limonene[p]	*A. thaliana*	—	AtLIMB	At*limb*	—	Z97341	—	Bevan *et al.*[ps]	Chromosome 4 CF6 (ESSA I) nt 172598–175344
(S)-Linalool	*Clarkia concinna*	LIS	Cc**g**LINOH	Cc*glinoh*	—	AF067602	Cseke *et al.* (1998)	Cseke *et al.* (1998)	—
Linalool[p]	*A. thaliana*	—	At**g**LINOH	At*glinoh*	—	AC02294	—	Federspiel[ds]	Chromosome 1 BAC FIIP17 nt 73996–78905
Vetispiradiene	*Hyoscyamus muticus*	Chimera	Hm**f**VET	Hm*fvet*	U20187	NA	Back and Chappell (1995)	Chappell[pc]	—
Vetispiradiene[p]	*A. thaliana*	—	AtVET	At*vet*	—	AL022224	—	Bevan *et al.*[ds]	Chromosome 4 BAC F12C12 (ESSA) nt 54692–56893

tc, genomic sequences by Trapp and Croteau (accession nos. pending); NA, sequences unavailable in the public databases but disclosed in journal reference; pc, sequences obtained by personal communications; ds, sequences in public database by direct submission but not published; p, sequences in database with putative function; c, confirmed gene by experimental determination stated in database; i, two possible isozymes reported for the same region referred to as A1 and A2; —, no former gene name or accession number. Species names are: *Abies grandis, Arabidopsis thaliana, Clarkia concinna, Gossypium arboreum, Hyoscyamus muticus, Mentha longifolia, Mentha spicata, Nicotiana tabacum, Ricinus communis, Perilla frutescens, Taxus brevifolia, Zea mays.*

[a] Former names, respectively, for (−)-copalyl diphosphate synthase and *ent*-kaurene synthase were *ent*-kaurene synthase A (KSA) and *ent*-kaurene synthase B (KSB), and mutant phenotypes were ga1 and ga2; these designations have been used loosely.

[b] Nomenclature architecture is specified as follows. The Latin binomial two-letter abbreviations are in spaces 1 and 2. The substrates (1- to 4-letter abbrev.) are in spaces 3–6, consisting of 1- or 2-letter abbrev. for substrate utilized in boldface (*e.g.*, **g**, geranyl diphosphate; **f**, farnesyl diphosphate; **gg**, geranylgeranyl diphosphate; **c**, copalyl diphosphate; **ch**, chrysanthemyl diphosphate; in lowercase) followed by stereochemistry and/or isomer definition (*e.g.*, α, β, δ, γ, etc. followed by epi (e), E, Z, -, +, etc.). The 3-letter product abbrev. indicates the major product is an olefin; otherwise the quenching nucleophile is indicated, (*e.g.*, ABI, abietadiene synthase; BORPP, bornyl diphosphate synthase; CEDOH, cedrol synthase); uppercase specifies protein and lowercase specifies cDNA or gDNA. All letters except species names are in italics for cDNA and gene. Distinction between cDNA and gDNA must be stated or a g is added before the abbreviation, *e.g.*, Tb***gg****tax* cDNA and gTb***gg****tax*, or Tb***gg****tax* gene (nomenclature system devised by S. Trapp, E. Davis, J. Crock, and R. Croteau).

TABLE 2

Primers used to isolate conifer genomic sequences from corresponding cDNAs

Gene name	Primer name	Primer sequence	Position[a]
Ag*f*Eα*bis*[b]	agc1.7F	5′ ACTTCAAAGATGCCAATGGG 3′	nt 1134–1114
	AG1/06R	5′ TGATTACAGTGGCAGCGGTTC 3′	nt 2439–2429
	AG1/05F	5′ GCTGGCGTTTCTGCTGTATC 3′	nt 3–21
	AG1/14R	5′ GCCAAGAAGTCTTCAGCGCG 3′	nt 1361–1341
	AG1/16R	5′ TGTTGTTCCAGTCACTGG 3′	nt 1314–1294
Ag*g-pin*	AG3/03F	5′ TTCTACCGCACCGTTGGC 3′	nt 12–29
	AG3/04R	5′ AACCCGACATAGCATAGG 3′	nt 1924–1907
Ag*f*δ*sel*	AG4/01F	5′ ATGGCTGAGATTTCTGAATC 3′	nt 1–20
	AG4/02R	5′ GACCATCACTATTCCTCC 3′	nt 1753–1737
Ag*g-lim*	AG10/01F	5′ GGCAGGAATCCATGGCTCTCCTT 3′	nt −1 to 24
	AG10/02R	5′ GAATAGTCTAGATTATAGACTTCCCAC 3′	nt 1985–1960
Ag*gg*abi	AG22/03F	5′ TGCTCATCATCTAACTGC 3′	nt 45–62
	AG22/04R	5′ ACACAATACCATGAGGGC 3′	nt 2645–2628
Tb*gg*tax	tax1/01F	5′ GAATTCCTTCCCCTGCCTCTCTGG 3′[c]	nt −21 to 5
	tax1/02R	5′ GCTCTAGAGCGCCAATACAATAATAAGTC 3′[c]	nt 2642–2624

[a] Number one and positive numbers represent ATG start site and downstream nucleotides (nt), respectively, designed from cDNA; minus numbers are upstream of the ATG start site.

[b] Sequencing the bisabolene synthase genomic sequence was accomplished by sequencing two overlapping fragments designated pAGg1-1 (1.7F and 06R) and pAGg1-11A (05F and 14R, 16R).

[c] Restriction enzyme (*Eco*RI or *Xba*I) sites were incorporated at the termini of these primers for cloning purposes.

appropriate genomic DNA as template. For Ag*f*Eα*bis*, Touchdown PCR amplification (Don *et al.* 1991) of the 5′ portion of two overlapping DNA templates was required (Table 2). Once the correspondence of cDNA to gDNA was confirmed, the latter was entirely sequenced. With the exception of Ag*g-pin1* and Ag*f*δ*sel*, all genomic sequences (after intron deletion) exhibited ≥98% identity to the corresponding cDNA. For the exceptions, two different products were observed in each case, corresponding to sizes of 3.3 and 2.8 kb for Ag*f*δ*sel*, and 3.2 and 2.8 kb for Ag*g-pin*. The Ag*f*δ*sel1* and Ag*f*δ*sel2* products were sequenced, and the deduced amino acid sequences were 92 and 87% identical to that of the published Ag*f*δ*sel* cDNA (formerly ag4; Steele *et al.*

TABLE 3

Terpene synthases sequenced in this study

Terpene synthase		cDNA			Genomic				
Product	Class	Former name[a]	Size[b]	aa[b]	Name	Size[b]	aa[b]	Clone name[c]	Introns[d]
Bisabolene	sesqui^{in}	ag1	2.541	817	Ag*f*Eα*bis*	4.647	817	pAGg1-1/1-11A[e]	11
(−)-Pinene	mono^{in}	ag3	1.884	627	Ag*g-pin1*	3.280	628	pAGg3-37	9
δ-Selinene	sesqui^{co}	ag4	1.743	581	Ag*f*δ*sel1*	3.346	579	pAGg4-31	9
					Ag*f*δ*sel2*	2.789	525	pAGg4-34	8
(−)-Limonene	mono$^{in, co}$	ag10	1.913	637	Ag*g-lim1*	1.913	637	pAGg10-1	9
Abietadiene	di^{in}	ag22	2.604	868	Ag*gg*abi	4.664	868	pAGg22-3/22-4[f]	14[c]
Taxadiene	di^{un}	tb1	2.586	862	Tb*gg*tax	3.999	862	pTBgtax1-1	12

The superscript notations in, co, and un refer to inducible, constitutive, or unknown-type enzyme expression, respectively.

[a] Former names for cDNAs are from Bohlmann *et al.* 1998b.

[b] Size of cDNA and genomic sequences are in kilobases; deduced amino acid (aa) sequences are translated from the cDNA, or from the genomic sequence without introns spliced.

[c] Plasmid clones were used for full-length genomic sequencing, with some exceptions (see Table 2).

[d] All introns are conserved in the same positions with reference to Ag*gg*abi, although the specific gene may not contain all 14 introns. Six introns are positionally conserved in all cases (Chappell 1995a).

[e] Bisabolene synthase sequence determined with the overlapping plasmid clones pAgc1-1 (3′-end) and pAgc1-11A (5′-end) containing inserts of 1.907 and 2.750 kb, respectively (see text).

[f] Both plasmid clones pAgc22-4 and pAgc22-3 were used to determine the full genomic sequence.

1998), indicating the presence of allelic variants, pseudogenes or, possibly, distinct but related genes. Only the 3.2-kb gDNA fragment of Ag*g-pin1* was completely sequenced and this version showed 99% identity to the published Ag*g-pin1* cDNA (formerly ag3; Bohlmann *et al.* 1997) at the deduced amino acid level.

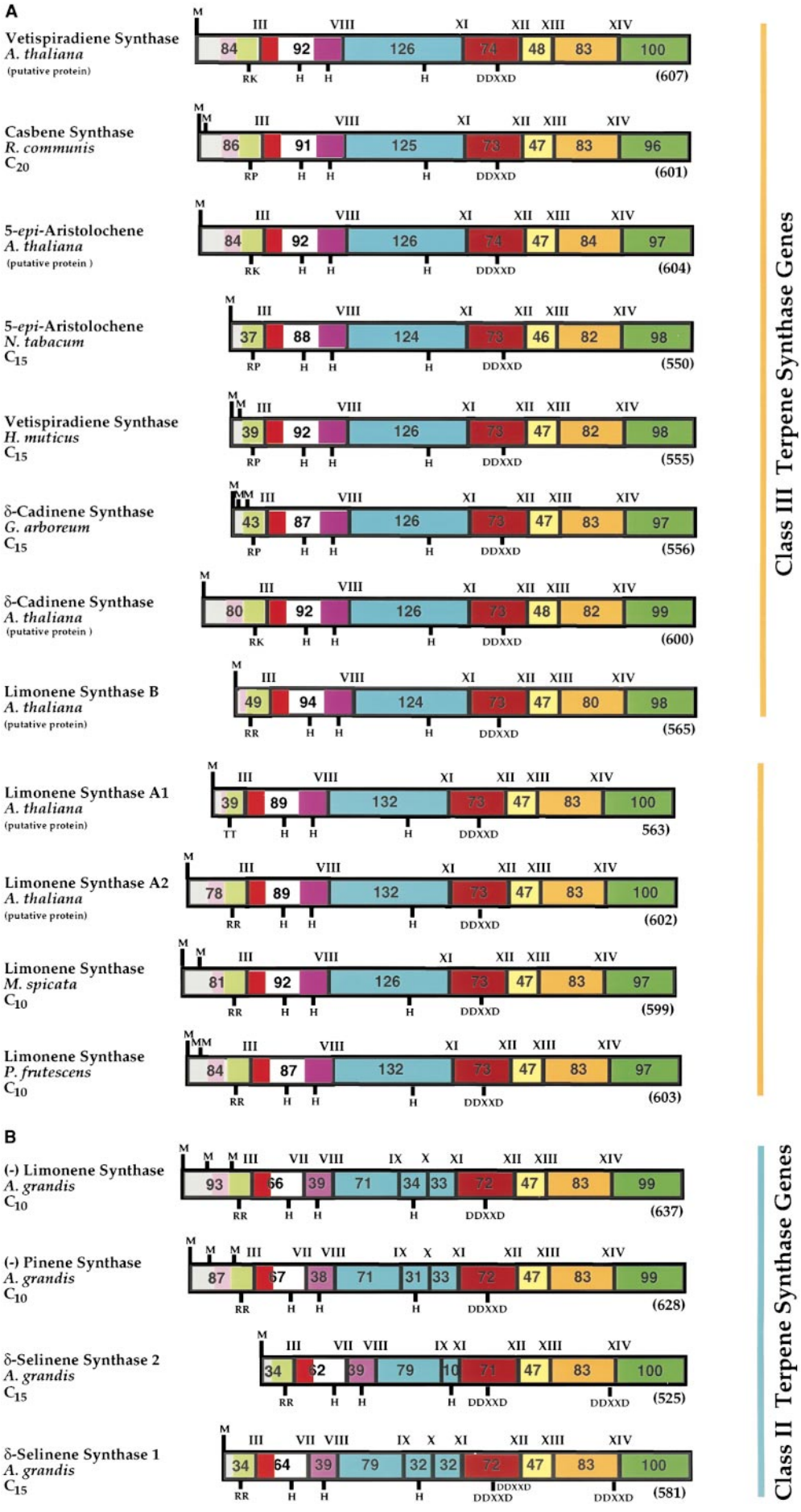
Intron/exon structure of terpene synthase genes: In addition to the new genomic sequences of the conifer terpene synthases (Table 3), genomic sequences were available for 13 angiosperm terpene synthases (Table 1). Seven of these are characterized terpene synthases (At*gg-copp1*, Cc*glinoh*, Ga*f*δ*cad*, Hm*fvet1*, Nt*feari4*, Rc*ggcas*, Pf*g-lim1*) and one is a chimera constructed from the published Ms*g-lim* cDNA sequence (Colby *et al.* 1993) and an unpublished Ml*g-lim* sequence. To complete the analysis, six putative terpene synthase sequences from Arabidopsis (pAt*eari4*, pAt*cad*, pAt*lim1a*, pAt*lim1b*, pAt*linoh*, pAt*vet*) were acquired by database searching (Table 1). It is of note that the public genomic databases poorly identify putative terpene synthase genes and are not very accurate in the prediction of intron splice sites and, thus, proper intron phase and exon definition; the algorithms could easily be improved. It is also worth noting that the putative terpene synthase genes of Arabidopsis cluster on a small portion of chromosome 4 (Table 1). Clustering of related pathways genes in plant genomes has not received much attention, and the terpenoid synthases and associated metabolic enzymes may be representative of this phenomenon.

Each of the 21 terpene synthase genomic sequences was analyzed for the number and size of exons and introns, as well as intron placement and position (phase). A distinct pattern of exon sizes emerged, and introns were observed at a total of 14 positions (Figure 3; Table 4). Introns were numbered according to placement starting with intron 1 closest to the 5′ terminus, and, in all of the observed introns, placement and phase are conserved (Figure 4). Intron phase is defined as the placement of the intron before the first, second, or third nucleotide position of the proximate codon and is referred to as phase 0, 1, or 2, respectively (Li 1997; *e.g.*, all Tps genes that contain intron 3 have a phase of 0). The one exception to intron phase conservation is intron 11 of the Hm*fvet* gene, which has a phase of 0 instead of 2, as do all other Tps genes from gymnosperms and angiosperms (Table 4). This discrepancy may represent an error in sequencing across the splice site or an example of intron sliding (Mathews and Trotman 1998).

An obvious pattern of intron size or sequence of the 14 introns was not detected, although a more rigorous comparison is required to determine this for certain. The introns of the conifer genes are relatively small (~150 nt on average) compared to those of the angiosperm genes (196 nt), especially those of Arabidopsis (266 nt). Moreover, Pf*g-lim1* and the putative pAt*lim*A1 gene contain several exceptionally large introns, ranging from 698 to 2850 nt in length (Table 4). In addition, the introns of the gymnosperm synthases are AT rich, with repetitive sequences rich in T (3–10 mers). Splice sites of all of the terpene synthases in this study have been compiled (see WebTables 1 and 2 at http://ibc.wsu.edu/faculty/croteau.html). The 5′-splice site consensus sequence for the conifer terpene synthases is NNNG$^{\nabla}$GTNNNN; however, there is a clear preference for G$^{\nabla}$GTAWD. The 3′-splice site consensus sequence, consistent with that of dicots (Hanley and Schuler 1988), is YAG$^{\nabla}$ (a minority of the sites consist of AAG$^{\nabla}$). A chart of amino acid sequence pair distances and alignments (showing intron splice sites) for all of the terpene synthases in this study is also available (see WebFigures 1 and 2 at http://ibc.wsu.edu/faculty/croteau.html).

Classification of terpene synthase genes: Comparison of genomic structures (Figures 3 and 5) indicates that the plant terpene synthase genes consist of three classes based on intron/exon pattern; 12–14 introns (class I), 9 introns (class II), or 6 introns (class III; Figure 3). Using this classification, based upon distinctive exon/intron patterns, the seven conifer genes are assigned to class I or class II (Figure 3–C). Class I comprises conifer diterpene synthase genes Ag*ggabi* and Tb*ggtax* and sesquiterpene synthase Ag*f*α*bis* and angiosperm synthase genes specifically involved in primary metabolism (At*gg-copp1* and Cc*glinoh*). Terpene synthase class I genes contain 11–14 introns and 12–15 of exons of characteristic size (Figure 3C), including the CDIS domain comprising exons 4, 5, and 6, and the first ~20 amino acids of exon 7, and introns 4, 5, and 6 (this unusual sequence element corresponds to a 215-amino-acid region [Pro^{137}-Leu^{351}] of the Ag*ggabi* sequence). Class II Tps genes comprise only conifer monoterpene and sesquiterpene synthases, and these contain 9 introns and 10 exons; introns 1 and 2 and the entire CDIS element have been lost, including introns 4, 5, and 6. Class III Tps genes comprise only angiosperm monoterpene, sesquiterpene, and diterpene synthases involved in secondary metabolism, and they contain 6 introns and 7 exons. Introns 1, 2, 7, 9, and 10 and CDIS domain have been lost in the class III type. The introns of class III Tps genes (introns 3, 8, 11–14) are conserved among all plant terpene synthase genes and were described as introns 1–6, respectively, in previous analyses (Mau and West 1994; Back and Chappell 1995; Chappell 1995b).

Of the class I Tps genes, introns 1 and 2 are observed only in Ag*ggabi* and At*gg-copp1* (Figure 3C), in which phase is also conserved. However, the placement of intron 1 is not conserved between the two genes; the slight discrepancy in placement of intron 1 may reflect poor alignment in this portion of the terpene synthase preproteins that defines the plastidial targeting sequence (Figures 3C and 4; Bohlmann *et al.* 1998b). (For more detail, see amino acid alignments at http://ibc.wsu.edu/faculty/croteau.html.) It is also notable that Ag*f*α*bis* (a


A
Vetispiradiene Synthase
A. thaliana
(putative protein)
84 92 126 74 48 83 100
(607)
Casbene Synthase
R. communis
C20
86 91 125 73 47 83 96
(601)
5-epi-Aristolochene
A. thaliana
(putative protein)
84 92 126 74 47 84 97
(604)
5-epi-Aristolochene
N. tabacum
C15
37 88 124 73 46 82 98
(550)
Vetispiradiene Synthase
H. muticus
C15
39 92 126 73 47 82 98
(555)
δ-Cadinene Synthase
G. arboreum
C15
43 87 126 73 47 83 97
(556)
δ-Cadinene Synthase
A. thaliana
(putative protein)
80 92 126 73 48 82 99
(600)
Limonene Synthase B
A. thaliana
(putative protein)
49 94 124 73 47 80 98
(565)
Limonene Synthase A1
A. thaliana
(putative protein)
39 89 132 73 47 83 100
563)
Limonene Synthase A2
A. thaliana
(putative protein)
78 89 132 73 47 83 100
(602)
Limonene Synthase
M. spicata
C10
81 92 126 73 47 83 97
(599)
Limonene Synthase
P. frutescens
C10
84 87 132 73 47 83 97
(603)
B
(-) Limonene Synthase
A. grandis
C10
93 66 39 71 34 33 72 47 83 99
(637)
(-) Pinene Synthase
A. grandis
C10
87 67 38 71 31 33 72 47 83 99
(628)
δ-Selinene Synthase 2
A. grandis
C15
34 62 39 79 10 71 47 83 100
(525)
δ-Selinene Synthase 1
A. grandis
C15
34 64 39 79 32 32 72 47 83 100
(581)
Class III Terpene Synthase Genes
Class II Terpene Synthase Genes

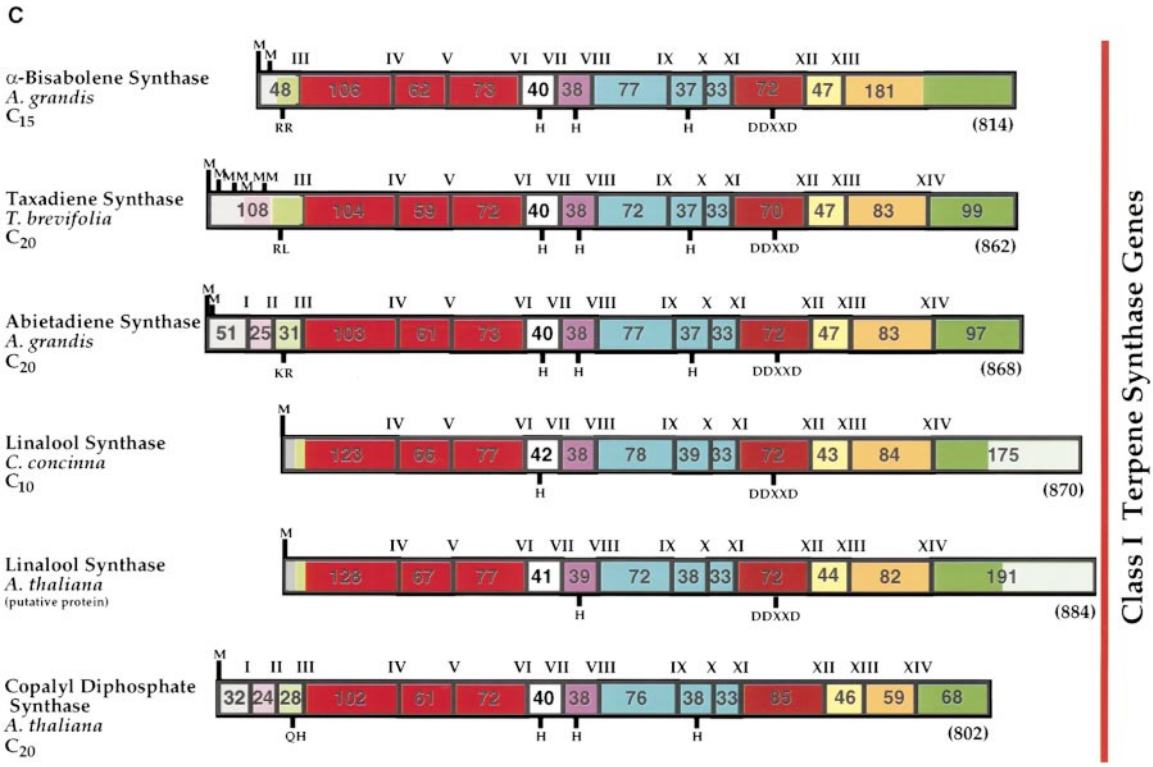


FIGURE 3.—Genomic organization of plant terpenoid synthase genes. Black vertical bars represent introns 1–14 (Roman numerals in figure) and are separated by colored blocks with specified lengths representing exons 1–15. The terpenoid synthase genes are divided into three classes (class I, class II, and class III), which appear to have evolved sequentially from class I to class III by intron loss and loss of the conifer diterpene internal sequence domain (CDIS; see Figure 4). (C) Class I Tps genes comprise 12–14 introns and 13–15 exons and consist primarily of diterpene synthases found in gymnosperms (secondary metabolism) and angiosperms (primary metabolism). (B) Class II Tps genes comprise 9 introns and 10 exons and consist of only gymnosperm monoterpene and sesquiterpene synthases involved in secondary metabolism. (A) Class III Tps genes comprise 6 introns and 7 exons and consist of angiosperm monoterpene, sesquiterpene, and diterpene synthases involved in secondary metabolism. Exons that are identically colored illustrate sequential loss of introns and the CDIS domain, over evolutionary time, from class I through class III. The methionine at the translational start site of the coding region (and alternatives), highly conserved histidines, and single or double arginines indicating the minimum mature protein (WILLIAMS *et al.* 1998) are represented by M, H, RR, or RX (X representing other amino acids that are sometimes substituted), respectively. The enzymatic classification as a monoterpene, sesquiterpene, or diterpene synthase is represented by C_{10}, C_{15}, C_{20}, respectively. Conifer terpene synthases were isolated and sequenced to determine genomic structure; all other terpene synthase sequences were obtained from public databases or by personal communication (see Table 1). Putative terpene synthases are referred to as putative proteins and are illustrated based upon predicted homology. Two different predictions of the same putative protein (accession no. Z97341) are shown as limonene synthase A1 and A2; if A1 is correct, the genomic pattern suggests that At*lim* (accession no. Z97341) is a sesquiterpene synthase; if A2 is correct, then At*lim* (accession no. Z97341) is a monoterpene synthase. In the analysis of intron borders of the Ms*g-lim*/Ml*g-lim* chimera and Hm*fvet1* genes (see Table 1), only a single intron border (5′ or 3′) was sequenced to determine intron placement; size was not determined. The intron/exon borders predicted for a number of terpene synthases identified in the Arabidopsis database were determined to be incorrect; these data were reanalyzed and new predictions used. The number in parentheses represents the deduced size (in amino acid residues) of the corresponding protein or preprotein, as appropriate.

sesquiterpene synthase) and Cc*glinoh* (a monoterpene synthase) are class I genes, although all other class I Tps genes are diterpene synthases. Furthermore, Cc*glinoh* and Ag*fαbis* are the only monoterpene or sesquiterpene synthase genes that contain the CDIS domain. Moreover, Ag*fαbis* and Cc*glinoh* genes are exceptions, even within the class I group, in that they both are devoid of introns at the extremes of the coding region; Ag*fαbis* lacks intron 14, and Cc*glinoh* lacks intron 3 (as well as 1 and 2). Finally, the angiosperm terpene synthase genes that fall within class I all encode enzymes involved in primary metabolism, with the exception of Cc*glinoh*.

Evolutionary history of the Tps gene family by gene architectural comparison: A schematic flow chart for the evolution of terpene synthase genes (Figure 5) was proposed on the basis of the data presented in Figure 3 and Table 4. Figure 5 provides the simplest account of ancestry derived by charting the physical patterns of proposed intron and domain loss and the consideration of additional conserved patterns of gene architecture that are not explicitly shown. Other possible mecha-

TABLE 4

Comparison of terpene synthase introns: number, size, placement, and phase

Intron	Rc*ggcas* C_{20}	At*fvet1*[a] C_{15}	Hm*fvet1*[b,c] C_{15}	At*fear1*[a] C_{15}	Nt*fear4* C_{15}	Ga*fδcad* C_{15}	At*cad1*[a] C_{10}	Ml*g-lim*[c,d] C_{10}	Pf*g-lim* C_{10}	At*limb*[a]	At*lima*1[a,e]	At*lima*2[a,e]	Agg*-lim* C_{10}	Agg*-pin1* C_{10}	Ag*fδsel2* C_{15}	Ag*fδsel1* C_{15}	Ag*ggabi* C_{20}	Ag*fEαbis* C_{15}	Tb*ggtax* C_{20}	At*glinoh*[a] C_{10}	Cb*glinoh* C_{10}	At*gg-copp1* C_{20}
I	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	K51* (151)	—	—	—	—	S32* (822)
II	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	N76*	— (96)	—	—	—	Q56* (115)
III	S86 (111)	S84 (71)	Q39 (ND)	S84 (91)	Q37 (127)	K43 (100)	S80 (116)	91[U] (87)	K84 (698)	V49 (100)	L39 (2.850)	F78 (2.734)	G93 (83)	E87 (107)	G34 (122)	G34 (120)	K107 (96)	I48 (205)	E108 (88)	—	—	Q84 (283)
IV	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	K210* (188)	Q154* (164)	Q212* (197)	K128* (912)	K123[U] (440)	K186* (253)
V	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	K271** (106)	A216** (428)	T271** (88)	Q195** (92)	R189[U] (181)	T247** (198)
VI	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	H344* (121)	F289* (546)	C343* (98)	G272* (108)	G266[U] (246)	G319* (121)
VII	—	—	—	—	—	—	—	—	—	—	—	—	Y159** (278)	Y154** (85)	Y96** (84)	Y98** (102)	Y384** (87)	Y329** (156)	Y383** (83)	Y313** (102)	Q307[U] (181)	H359** (79)
VIII	S177* (133)	S176* (91)	P131* (ND)	Y176* (115)	P125* (87)	C130* (101)	S172* (99)	183[U] (ND)	Q171* (826)	Q143* (212)	Q129* (231)	Q167* (231)	S198* (83)	S192* (146)	S135* (147)	S137* (147)	S422* (88)	P367* (134)	S421* (197)	P352* (262)	P346[U] (129)	A397* (82)
IX	—	—	—	—	—	—	—	—	—	—	—	—	E269 (84)	E263 (92)	E214 (101)	E216 (98)	E499 (156)	E444 (108)	E493 (130)	K424 (55)	M424[U] (84)	E473 (212)
X	—	—	—	—	—	—	—	—	—	—	—	—		N294** (537)	—	L248** (84)	Y536** (510)	Y481** (96)	Y530** (110)	L462** (100)	L462[U] (96)	Y511** (357)
XI	F302** (428)	T302** (98)	S255 (ND)	T302** (90)	S251** (76)	F256** (322)	T298** (103)	309[U] (ND)	L303** (125)	S267** (437)	S260** (95)	S299** (95)	S336** (101)	Y327** (99)	R224** (55)	T279** (472)	R569** (90)	T514** (116)	T563** (106)	T495** (140)	E495[U] (73)	Q544** (263)
XII	A375** (93)	K376** (86)	Q328** (93)	E376** (92)	Q324** (131)	E329** (103)	E371** (122)	383[U] (ND)	R376** (326)	E340** (89)	K334** (174)	K372** (174)	K408** (108)	K399** (94)	R295** (142)	R351** (146)	K641** (190)	R586** (141)	K633** (99)	L567** (222)	R567[U] (119)	D629** (927)
XIII	A422 (79)	E424 (248)	R375 (ND)	E423 (193)	R370 (155)	A376 (165)	E419 (1844)	429[U] (ND)	S423 (91)	S387 (308)	V380 (73)	V419 (73)	A455 (102)	A446 (106)	C342 (361)	C398 (366)	V688 (82)	G633 (103)	P680 (118)	I611 (139)	L611[U] (82)	S675 (494)
XIV	V505 (81)	E507 (87)	E457 (ND)	E507 (84)	E452 (113)	K459 (286)	E501 (583)	512[U] (101)	L506 (119)	S467 (178)	T463 (402)	T502 (402)	K538 (127)	K529 (120)	E425 (76)	E481 (78)	Q771 (98)	—	Q763 (99)	Q693 (124)	L695[U] (84)	K734 (270)
No. of Introns	6	6	6	6	6	6	6	6	6	6	6	6	9	9	8	9	14	11	12	11	11	14
aa	601	607	555	604	550	556	600	599	603	516	563	602	637	628	525	582	868	814	862	884	870	802

No asterisk, *, **, indicates intron inserted, respectively, between codons (does not interrupt codon); between the first and second nucleotide; between second and third nucleotide. —, intron not present; ND, intron not determined; U, intron phase not determined; intron numbers in boldface represent highly conserved introns throughout.

[a] All *A. thaliana* genomic sequences reported here are putative with the exception of copalyl diphosphate synthase found in the database corresponding to a published cDNA formerly called GA1 (see Table 1).

[b] Information obtained from J. Chappell (personal communication).

[c] Sequences are not complete; the authors sequenced the regions across the 5′- and 3′-splice sites (either or both) to verify the intron placement.

[d] Ml*g-lim* was analyzed by preparing a chimeric genome sequence utilizing cDNA from *M. spicata* and incomplete gDNA sequence from *M. longifolia* (T. Davis, unpublished data).

[e] All three At*lim* A1, A2, and B sequences are putative limonene synthase homologs from the *A. thaliana* database that are adjacent to each other on chromosome 4 (see Table 1). At*lima1* and At*lima2* are predictions of two possible isozymes. In a recent report At*lima* and At*limb* are referred to as At*TPS02* and At*TPS03* with 52 and 56% identity, respectively, to a newly defined Arabidopsis monoterpene synthase Myrcene/(*E*)-β-Ocimene Synthase, AtTPS10cDNA (Bohlmann *et al.* 2000).

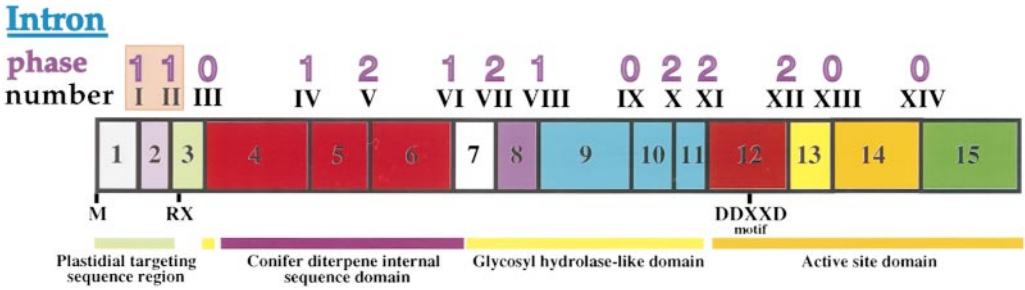


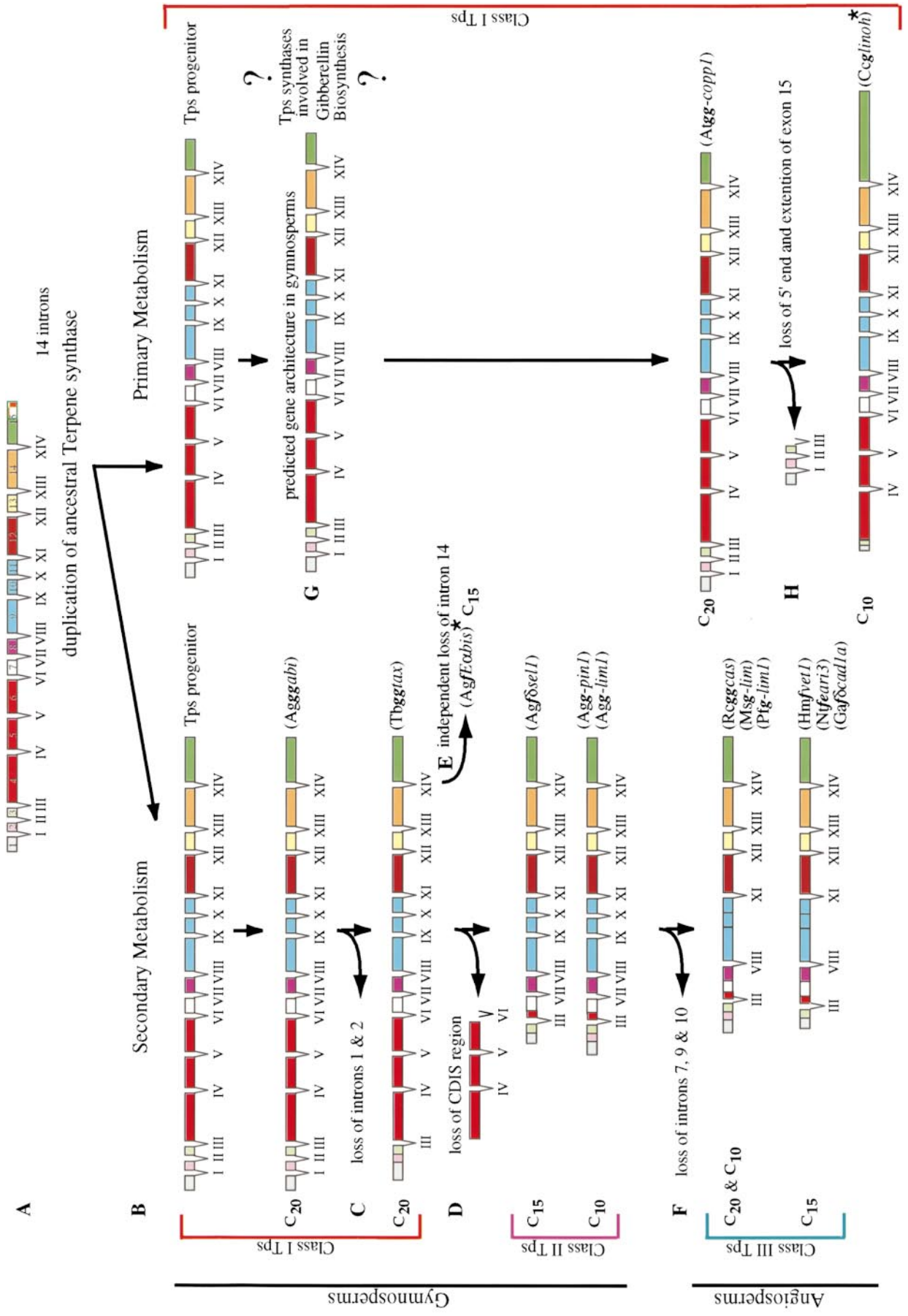
FIGURE 4.—General structural features of plant terpene synthase genes. A generic parent terpene synthase gene (class I type) illustrates the pattern of intron placement and phase conservation among all genes analyzed. Introns 1–14 (Roman numerals in figure) are represented by black vertical bars, and exons 1 through 15 are depicted by color boxes (with color coding as in Figure 3). Introns 1 and 2 are found in only two class I terpene synthase genes (Ag*ggabi* and gAt*ggcopp1*) and are boxed in peach to indicate that their presence is not a class I Tps gene requirement. Introns 3, 8, and 11–14 are conserved in all plant terpene synthases, with the exception of Ag*ffEαbis*, which has lost intron 14. The number above the intron Roman numeral represents the intron phase number and demonstrates conservation throughout this gene family. Introns are classified into three phase types according to LI (1997). General structural domains are labeled. The RX (representative of RR; see Figure 3) and DDXXD motifs are shown in boldface. The consecutive gray, pink, and light green boxes, representing exons 1–3 in Ag*ggabi* and At*ggcopp1*, comprise a single exon of variable size in all other terpene synthases; monoterpene and diterpene synthases comprise an exon size of 80–107 aa, whereas sesquiterpene synthases comprise an exon size of 30–50 aa due to the absence of a plastidial targeting sequence (depicted by the green bar). The conifer diterpene internal sequence domain (CDIS, red bar), identified by BOHLMANN *et al.* (1998a,b), is present only in class I type terpene synthase genes. The glycosyl hydrolase-like domain (yellow bar), the active site domain (orange bar), and the intradomain region (white space) are predicted (BOHLMANN *et al.* 1998b) on the basis of the crystal structure of Nt*fe*ARI4 (STARKS *et al.* 1997).

nisms of derivation are less suitable on the basis of the assumption that the gymnosperms predate angiosperms, that intron gain with conservation in placement and phase in both classes is less likely, and that terpene synthases of secondary metabolism almost certainly derived from those of primary metabolism. The schematic "model" (Figure 5) thus represents the most parsimonious, and biochemically consistent, explanation for Tps evolution when intron and CDIS alterations are taken into account.

This model of evolution was tested by computer-simulated analysis utilizing a simple algorithm, with which a satisfactory phylogentic tree was obtained (Figure 6A). By this analysis, class III type Tps genes are placed in a clade branching directly from class II types. However, when the model was tested more rigorously utilizing standard tree-searching computer-generated methods, we obtained a tree that is very similar to the previously published tree based on protein phylogeny (BOHLMANN *et al.* 1998b; Figure 6B). These protein sequence-based trees place the class III Tps genes as branching separately from both gymnosperm class I and class II Tps genes involved in secondary metabolism (Figures 5 and 6B; the group of class III Tps genes are highlighted for clarity in Figure 6, A and B). The latter derivation is not the most parsimonious when gene architecture is considered. It is possible, nevertheless, that the nucleotide sequence comparison of terpene synthases could generate a phylogenetic tree that differs from the protein sequence-based reconstruction. Preliminary analysis of the coding sequences of the defined terpene synthases to investigate this possibility was inconclusive and will require further investigation. Furthermore, the number of introns and variability in sequence made the alignment and phylogenetic analysis of genomic nucleotide sequences difficult. The lack of consonance between the gene architecture-based and protein homology-based models, in terms of placement of the class III Tps types, will likely be resolved with the acquisition of a larger, more diverse data set, including better outgroups to aid in inferring branching order. Although the protein-based phylogeny cannot be ignored, the gene architectural data would appear to be the more compelling at present.

DISCUSSION

Historically, the evolution and complexity of natural products has puzzled scientists across disciplines, from ecology to chemistry, who have been fascinated by the question as to why these chemicals, which are not essential for viability, are biosynthesized (MAPLESTON *et al.* 1992; CHRISTOPHERSEN 1996; JARVIS and MILLER 1996; JARVIS 2000). It seems most reasonable to assume that the precursors and pathways for the generation of natural products arose from mutations of enzymes involved in the synthesis of primary metabolites. Under the pressure of natural selection, natural products evolved as messages, broadly defined (JARVIS and MILLER 1996; JARVIS 2000), which increased the survival fitness of the producing organism (WILLIAMS *et al.* 1989; MAPLESTON *et al.* 1992; JARVIS and MILLER 1996). Addressing the evolutionary origins of natural products and their complex biosynthetic pathways is an arduous undertaking, because the genetic and functional bases for such a diverse repertoire of compounds are difficult to analyze, as many of the relevant biosynthetic enzymes are unknown; the task is often further complicated by the inability to distinguish between compounds involved in primary and secondary metabolism based on structure alone (JARVIS and MILLER 1996; JARVIS 2000). Few studies have focused on the evolution of natural products,


A
14 introns
duplication of ancestral Terpene synthase
Primary Metabolism
Secondary Metabolism
B
Tps progenitor
Class I Tps
G
predicted gene architecture in gymnosperms
Tps synthases involved in Gibberellin Biosynthesis
?
?
C
loss of introns 1 & 2
(Aggabi)
C20
D
loss of CDIS region
(Tbggtax)
E
independent loss of intron 14
(AgfEabis)
C15
(Agfδsel1)
(Agg-pin1)
(Agg-lim1)
C15
C10
Class II Tps
F
loss of introns 7, 9 & 10
C20 & C10
(Rcggcas)
(Msg-lim)
(Pfg-lim1)
C15
(Hmfvet1)
(Nyfeari3)
(Gafαcad1a)
Class III Tps
H
loss of 5' end and extention of exon 15
C20
(Atgg-copp1)
C10
(Ccglinoh)
Gymnosperms
Angiosperms

although a number of papers have discussed observations within specific classes of natural products from a variety of kingdoms (*e.g.*, fungi, plants, bacteria; Mapleston *et al.* 1992; Jarvis 2000). A notable example is the superfamily of chalcone synthase-related enzymes, which catalyze the first committed step in the biosynthesis of a number of biologically important plant products, *e.g.*, flavonoids (flower color) and phytoalexins (antimicrobials; Schrsder 1999), but which appear to be involved in the biosynthesis of a much larger range of substances than was previously realized (Eckermann *et al.* 1998). Recent analysis has revealed related sequences in bacteria, suggesting that this protein family is much older than was previously assumed (Schrsder 1997).

The majority of plant molecular evolutionary studies have focused on the chloroplast genome, and few have examined the molecular evolution of plant nuclear genes (Clegg *et al.* 1997). Most often, the gene families selected for such studies encode enzymes that play a well-defined role in a limited set of primary metabolic pathways, such as the small subunit of ribulose-bisphosphate carboxylase/oxygenase and alcohol dehydrogenase (Clegg *et al.* 1997). Very few studies have examined the origins of nuclear multigene families, *e.g.*, phytochrome, heat shock, *knox*, and catalase genes (Mathews *et al.* 1995; Frugoli *et al.* 1998; Bharathan *et al.* 1999; Waters and Vierling 1999). Even fewer studies have analyzed gene superfamilies such as that encoding chalcone synthase, which participates in a range of functionally diverse pathways. The terpene synthase genes, which encode multiple classes of mechanistically related enzymes, are arguably among the most functionally diverse group of genes and furthermore, as a group, are involved in both primary and secondary metabolism in many phyla.

The evolution of the terpene synthase gene superfamily is an instructive model to address the complex questions surrounding the origins of natural products. Terpenoids are the largest class of natural products, they are present and often abundant in all phyla, and they serve a multitude of functions in their internal environment (primary metabolism) and external environment (ecological interactions). The biosynthetic requirements for terpene production are the same for all organisms (a source of isopentenyl diphosphate, isopentyl diphosphate isomerase or other source of dimethylallyl diphosphate, prenyltransferases, and terpene synthases). The terpene synthases (regardless of phylogenetic origin) provide a unique focus since they are mechanistically very closely related yet are capable of producing a diverse array of structural types and derivatives. The conservation of genomic organization throughout the large multigene superfamily encoding plant monoterpene, sesquiterpene, and diterpene synthases, especially intron architecture, provides a compelling argument (Figures 3 and 4; Table 4) for reconstructing the evolutionary history of the terpene synthases from primary to secondary metabolism on the basis of the proposed pattern of gene sequence loss (introns and CDIS domain; Figure 5).

Current model for the evolution of terpene synthases: The three classes of terpene synthase genes exhibit clear intron phase conservation coupled to a distinct, and seemingly sequential, pattern of intron (and CDIS) loss, thereby suggesting the derivation of this gene family from an ancestral class I type terpene synthase of primary metabolism common to both gymnosperms and angiosperms. This proposed ancestral terpene synthase contained 12–14 introns, 13–15 exons, and the CDIS domain, as do all class I type Tps genes. The most obvious modern candidate that resembles this ancestral gene is postulated to be a contemporary Tps gene that contains the largest number of introns. This candidate comprises a genomic architecture most similar to either the conifer Ag**gg***abi* gene or the angiosperm At**gg**-*copp1* gene, both of which contain 14 introns and 15 exons as illustrated at the A1 and A2 branchpoints of Figure 7, with At**gg**-*copp1* most likely because it is involved in primary metabolism. Branchpoint B (Figure 7) indicates the loss of introns 1 and 2 within the class I type Tps genes to yield those containing 12 introns, 13 exons,

Figure 5.—Schematic proposal for evolution of the terpene synthase gene family in plants. This theory is based upon the most parsimonious account for the loss of introns and the conifer diterpene internal sequence domain (CDIS) over evolutionary time, sequentially from gymnosperms to angiosperms, and it takes into account the conserved pattern of exon domain size and intron phase data (not shown in schematic). Step A represents the predicted terpene synthase gene architecture for the ancestral gene of terpene synthases of both primary and secondary metabolism. Throughout, each colored block represents an exon (numbered 1–15) that is charted for defined terpene synthase genes. The downward facing open arrowhead (V) between each exon represents the positional placement of each intron for introns 1–14 (Roman numerals in figure). Steps B–H are hypothetical branchpoints in the proposed evolution of the Tps family, in which duplication of the depicted gene occurred. In steps C–H, the duplicated gene is not shown; a side arrow is used to represent the difference between the two duplicated genes illustrating how the second gene diverged. In steps B–F, divergence (illustrated by loss of introns or CDIS) of duplicated gene leads to novel terpene synthases. Step G is hypothetical, illustrating that the ancestral gene remained conserved in gene architecture within primary metabolism of gymnosperms, plausibly in gibberellin biosynthesis, and has remained conserved in gene architecture and function in angiosperms. The genomic sequence of gymnosperm gibberellin biosynthetic pathway genes is unknown and is depicted by the question marks. Steps E and H indicate exceptions to the otherwise conserved architecture and asterisks have been placed after the corresponding gene name. Bisabolene synthase has independently lost intron 14 (E), and a simple explanation for gene architecture and the origin of linalool synthase (Cc**g***linoh*) is presented (H). Each class of Tps gene is identified as type I, II, or III.

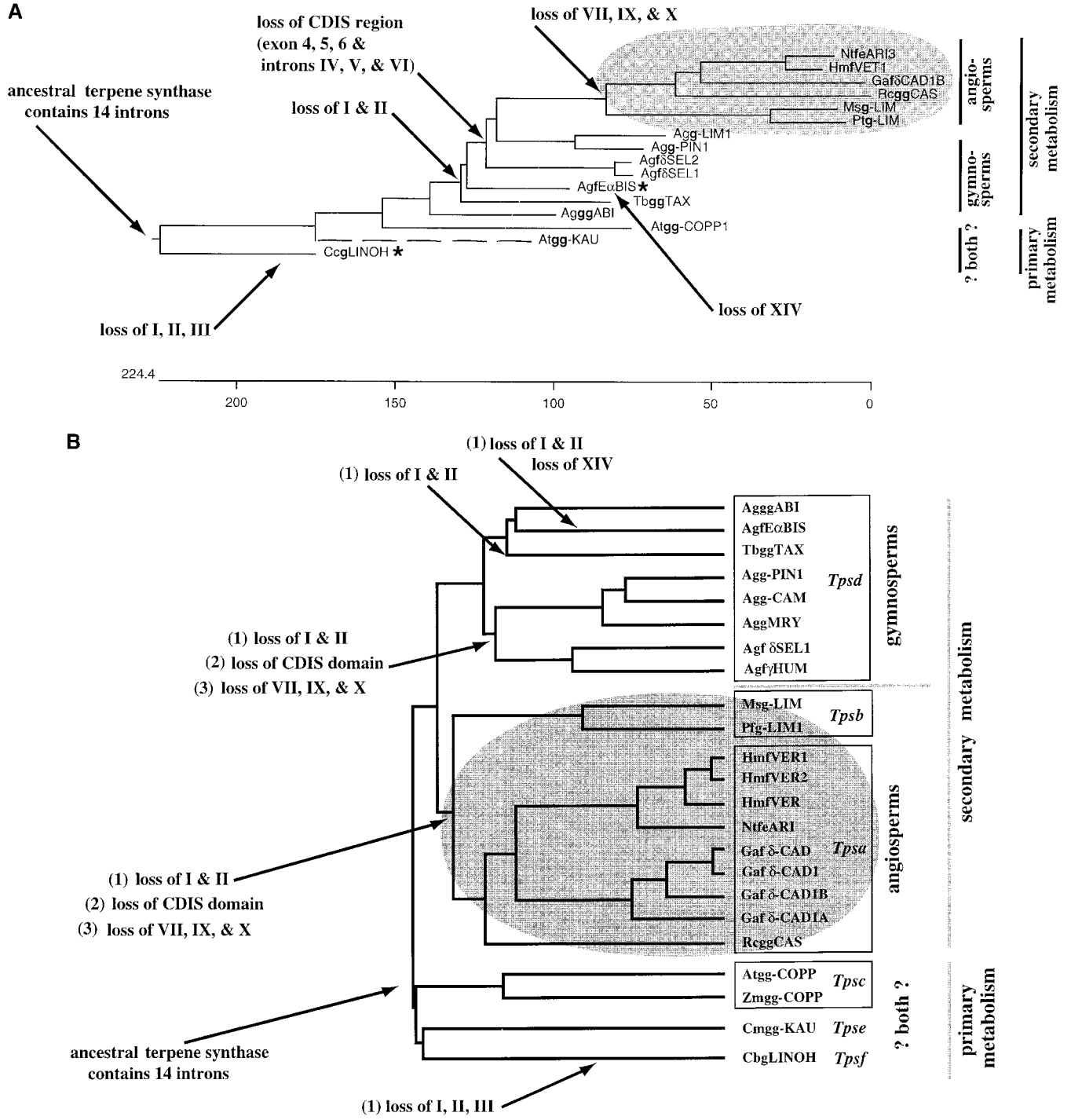


FIGURE 6.—Comparison of computer-generated phylogenetic trees for plant terpene synthases. Qualitative characters (intron and CDIS domain loss) are physically mapped on the trees. Amino acid sequences of defined terpene synthases were used to generate both trees by using a distance substitution method (A) or Dayhoff method previously published (BOHLMANN *et al.* 1998b) (B). Class III terpene synthases, consisting of all angiosperm terpene synthases of secondary metabolism, are highlighted in A and B by shaded ovals. The number scale on the bottom of the tree (A) represents amino acid substitution events.

and the CDIS domain (*e.g.*, Tb***gg****tax* gene). It is also plausible that the ancestral candidate gene resembles Tb***gg****tax*, and that introns 1 and 2 (as found in Ag***ggg****abi* and At***gg****-copp1*) resulted from recent intron acquisition. This rationale could explain why only Ag***ggg****abi* and At***gg****-copp1* contain intron 2 and the positionally nonconserved intron 1 compared to all other class I terpene synthase genes. The latter interpretation, however, seems less likely because At***gg****-copp1* encodes an enzyme that is essential for plant hormone production, which can be assumed to predate genes encoding enzymes of secondary metabolism (*i.e.*, Ag***ggg****abi*, Tb***gg****tax*, and Ag***f****Eα**bis*).

In our model for the molecular evolution of plant

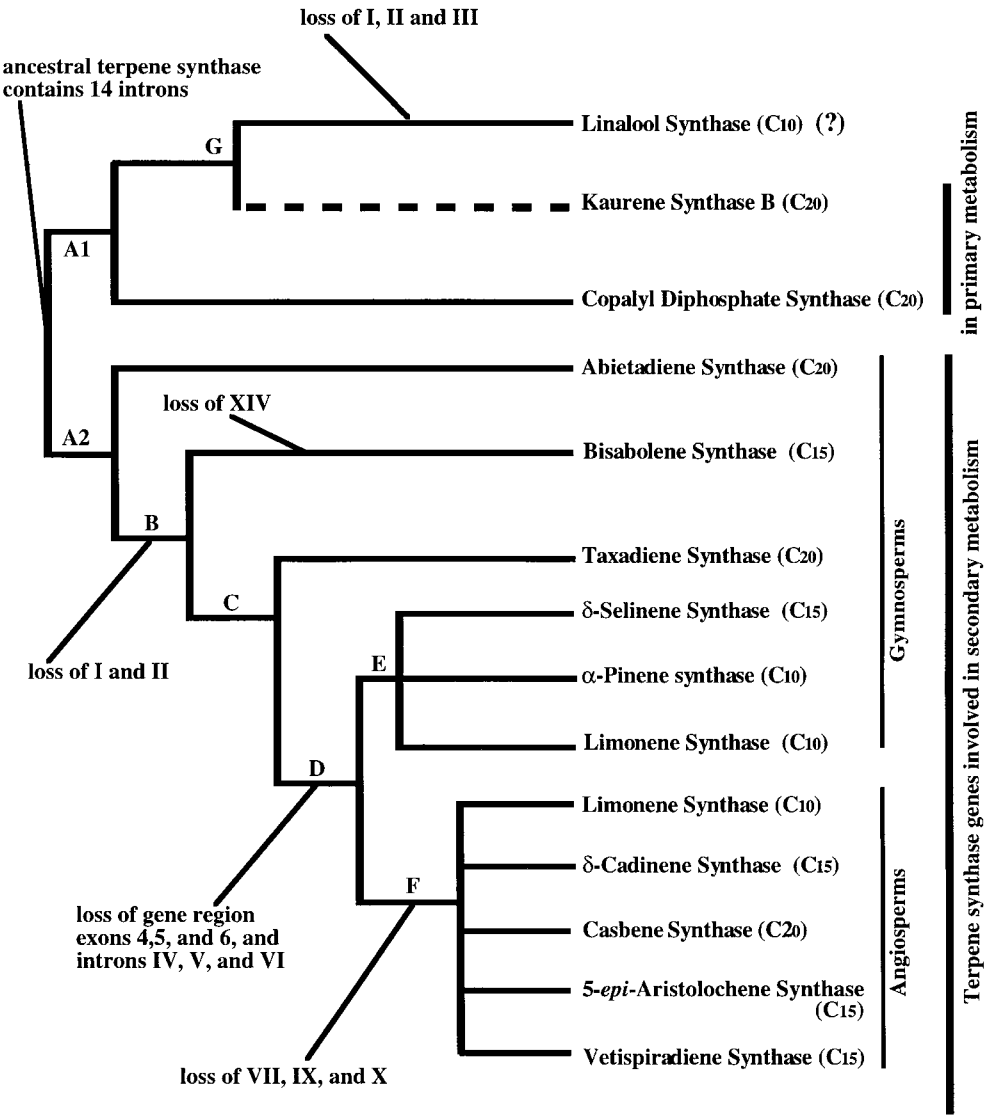


FIGURE 7.—A theoretical model for the evolution of plant terpene synthase genes with observed changes in gene architecture noted (from data summarized in Figure 5). Letters A–G of the dendrogram indicate inferred branchpoints, with A1 and A2 representing the progeny of the duplicated ancestral gene predicted to be similar to present-day gymnosperm diterpene synthase genes with 14 introns and 15 exons (class I Tps gene type). The duplication events to create present-day gymnosperm diterpene synthases (*e.g.*, At*gg-copp1* and Ag*ggabi* each containing 14 introns) occurred prior to the separation of gymnosperms and angiosperms and preceded the specialization of terpenoid synthases involved in primary or secondary metabolism. Subsequently, A1 progeny are predicted to have remained conserved in genomic structure and function, whereas A2 progeny have continuously specialized by gene duplication and divergence (involving sequential loss of introns and CDIS domain) to produce a superfamily of terpene synthases involved in secondary metabolic processes. The mechanism of concerted intron loss can account for the first (B) and third (F) loss events. The second loss event (D), involving an entire region spanning exons 4, 5, and 6, and introns 4–6, gave rise to the class II terpene synthase types and could have occurred by a number of recombination mechanisms. The third loss event gave rise to class III terpene synthase genes. The loss of single intron 14 in Ag*fαbis* (C) and introns 1–3 in Cc*glinoh* (G) appears to have occurred independently of the global evolutionary pattern by concerted loss or other mechanism and may account for the unusual classification of these genes. The question mark (?) indicates the uncertain placement of the linalool synthase gene within this scheme (G). The dotted line indicates the predicted placement of an angiosperm or gymnosperm kaurene synthase for which a genomic structure has not yet been described. The symbols C_{10}, C_{15}, and C_{20} represent the enzyme class corresponding to monoterpene synthases, sesquiterpene synthases, and diterpene synthases, respectively.

terpene synthase genes, based on both theoretical (Figure 7) and computer-generated phylogenetic trees (Figure 6A), class I type terpene synthase genes gave rise at branchpoint D to class II type terpene synthase genes (Ag*g-lim*, Ag*g-pin1*, and Ag*fδsel*) comprising 9 introns and 10 exons. Characteristically, class II type terpene synthase genes encode conifer monoterpene and sesquiterpene synthases involved in secondary metabolism, and they have lost the entire CDIS domain spanning exon regions 4, 5, 6, and a small portion of exon 7 that includes introns 4, 5, and 6. Class III type terpene synthase genes (Ga*fδcad*, Hm*fvet1*, Nt*feari4*, Rc*ggcas*, and Pf*g-lim*) derive from class II types by a further loss of intron 7 and sequential loss of introns 9 and 10, as depicted at branchpoint F (Figure 7). All class III type Tps genes contain the 6 conserved introns that are found in all terpene synthase genes (3, 8, 11–14), [these were previously described as introns 1–6 (MAU and WEST 1994; BACK and CHAPPELL 1995; CHAPPELL 1995a)], and class III types comprise all angiosperm monoterpene, sesquiterpene, and diterpene synthases involved in natural products biosynthesis.

The current evolutionary hypothesis and the previous phylogenetic tree (BOHLMANN *et al.* 1998b) both affirm that plant terpene synthases involved in secondary metabolism have a common ancestral origin from a terpene synthase involved in primary metabolism. However, these models differ in the placement of the class III

terpene synthases (Figures 6, A and B, and 7). The inferred phylogenetic tree in Figure 6A suggests that the class III types (angiosperm terpene synthases) involved in secondary metabolism are descendants (evolving by divergent evolution) of the class II types (gymnosperm monoterpene and/or sesquiterpene synthases) involved in secondary metabolism, whereas Figure 6B indicates that class III types are not direct descendants of class II types, yet share a common ancestor. The latter (including the gene architecture information as valid) infers that gymnosperm and angiosperm terpene synthases involved in secondary metabolism diverged prior to the three loss events illustrated in Figure 6A and that these subsequent loss events occurred in parallel after the split between the two groups. This further implies that terpene synthases underwent gene duplication and divergence several times to create enzymes of both primary and secondary metabolism (divergent evolution), which subsequently evolved by convergent (or concerted) evolution. Thus, if the latter holds true (as in 6B), the same set of introns (7, 9, 10), noting that intron 7 is nonadjacent to 9 and 10, and a complete domain (including 3 exons and 3 introns) have been lost at least twice. This possibility cannot be eliminated; yet it seems unlikely that at least three separate events of sequence loss (event 1: loss of introns 1 and 2; event 2: loss of introns 9, 10, and 7; event 3: domain loss consisting of 3 exons and 3 introns) occurred at least twice (if not more) in the two groups. Thus, Figure 6A appears closer to the true tree based upon maximum parsimony for the patterns observed in gene architecture of the terpene synthases. Nevertheless, to resolve the conflict in topology will require improvements in the data set, including the inclusion of better outgroups (such as ferns, mosses, and cycads) and additional phyla (other gymnosperms and monocots) to improve the inferred phylogenetic tree for rooting as well as biases. Sequences for gymnosperm terpene synthases involved in primary metabolism, such as kaurene synthase and copalyl diphosphate synthase, could greatly aid in rooting the tree.

Mechanism of intron loss: The patterns of structural change observed in the terpene synthase genes, as the basis of the maximum parsimony model, is concordant with an experimentally demonstrated and presumably common (Frugoli *et al.* 1998) process of concerted intron loss (Derr *et al.* 1991). Plausible mechanisms for concerted intron loss and individual intron loss have been proposed (Baltimore 1985; Fink 1987) and have been demonstrated for other plant gene families (Huang *et al.* 1990; Kumar and Trick 1993; Häger *et al.* 1996; Frugoli *et al.* 1998). For the evolution of terpene synthase genes, such mechanisms provide an established rationale by which contiguous blocks of introns can be lost in a single event, such as the concurrent loss of introns 9 and 10, as well as the loss of nonadjacent intron 7 in class III Tps types (Figures 5 and 6). Presently, the data are insufficient to determine unequivocally the chronology of loss of introns 7, 9, and 10 or the interdependence of these loss events. Nevertheless, the precedented mechanisms of concerted intron loss provide a most reasonable explanation for the observed differences in terpenoid synthase genes and the derivation of terpenoid synthases of natural products biosynthesis from those of primary metabolism.

Mechanism of evolution of terpene synthases genes: The derivation of large gene families presumably involved repeated gene duplication of the ancestral gene and divergence by functional and structural specialization, an evolutionary process now viewed as quite common (Fryxell 1996; Clegg *et al.* 1997). Tandem duplication occurs spontaneously in bacteria at a frequency of 10^{-3}–10^{-5} per locus per generation and is documented to occur with similar frequency in insect and mammalian genomes. If spontaneous mutation of duplicated genes occurs frequently, the rate-limiting step in retention occurs at the level of natural selection (or genetic drift) since the altered gene will be lost rapidly unless it acquires a new (functional mutation) and useful function (divergence; Fryxell 1996). These considerations suggest that family trees of functionally related genes coevolved because functionally complementary gene duplications and divergence events tended to be retained by natural selection (Fryxell 1996). This hypothesis provides a plausible mechanism for the origin and evolutionary relationships of plant terpene synthase genes and their encoded enzymes.

On the basis of limited data, several groups have suggested that all plant terpene synthases share a common evolutionary origin (Colby *et al.* 1993; Mau and West 1994; Back and Chappell 1995), with the protein-based phylogenetic analysis of 33 plant terpene synthases providing the most compelling evidence thus far (Bohlmann *et al.* 1998b). The conservation of genomic organization that underlies the current evolutionary model presented here provides substantial further evidence that the terpene synthases from gymnosperms and angiosperms constitute a superfamily of genes derived from a shared ancestor. Prior to the divergence of gymnosperms and angiosperms, during the carboniferous period ~300 mya (Doyle 1998), the duplication of an ancestral terpene synthase gene, resembling most closely a contemporary diterpene synthase of primary metabolism, occurred. One copy of the duplicated ancestral gene remained highly conserved in structure and function, and this gene may have contemporary descendants in the terpene synthases involved in gibberellin biosynthesis. The second ancestral gene copy diverged in structure and function, by adaptive evolutionary processes, to yield the large superfamily of terpene synthases involved in secondary metabolic pathways. Although speculative, it is plausible that the early terpene synthase ancestors were functionally less specialized than modern forms and perhaps able to utilize several prenyl diphosphate substrates for the production of

multiple terpene types, the specialization into different synthase classes (for monoterpenes, sesquiterpenes, and diterpenes) having evolved much later.

The genetic changes (including distinct patterns of conservation in exon size, intron placement and intron phase, and the apparent sequential loss of introns and the CDIS element) observed among the plant terpene synthases suggest that gene organization may have been a greater driving force in the evolution of these enzymes than was previously thought. Gene organization may have played an important role in diversifying terpene structures and the ecological interactions that they mediate. Although the evolutionary connections are unclear, the absence of the CDIS domain in the angiosperm and gymnosperm monoterpene and sesquiterpene synthase could be significant for terpene structural diversification, and this apparent loss confirms the previous view (Bohlmann *et al.* 1998b) that the CDIS domain is not required for protein function. Furthermore, there is an obvious correlation between the genetic changes and an important structural aspect of the terpene synthases. Thus, considerable genetic variation occurs in the 5′ portion (N-terminal region of uncertain function) of the terpene synthase genes (mutations in the form of intron and CDIS loss), whereas the 3′ portion (which encodes the C-terminal active site including exons 12–15; see Figure 4) remains highly conserved in organization and catalytic function (no intron losses or change in exon size over evolutionary time).

Variations and exceptions: It was recently suggested that linalool synthase from several Clarkia species is a composite gene resulting from a discrete recombination event (*e.g.*, domain swapping) between the 5′ half of a copalyl diphosphate synthase type gene (class I type) and the 3′ half of a limonene synthase type gene (class III type; Cseke *et al.* 1998). On the basis of this apparent chimeric structure, Cseke *et al.* (1998) postulated that other terpene synthases could have arisen by similar recombination between extant terpene synthase types. Although this mechanism provides a credible explanation for the derivation of linalool synthase, it does not account for the global evolutionary history of the terpenoid synthases, as does the present model, which implies that all Tps genes share a common ancestor and are sequentially derived from class I to class III, regardless of precise enzyme mechanism or the presence or placement of specific primary structural elements.

There are several terpene synthase genes of the class I type that vary in structure from the general classification. Cc*glinoh* and At**gg**-*copp1* differ in exon size at the 3′ termini; Cc*glinoh* has lost a significant portion of the distal 5′ region (exons 1–3 including the conserved intron 3), and Ag*fEα bis* apparently has independently lost conserved intron 14 (Figure 3). In linalool synthase, exons 4–14 have similar exon sizes as do all other Tps class I types (Figure 3); however, exon 15 of Cc*glinoh* and putative At*glinoh* contain an additional 78 and 91 amino acids (aa), respectively. These additions to exon 15 (an increase from an average size of 100 aa) might be explained by a number of mechanisms, including internal duplications, a mutational change that converts a stop codon into a sense codon, insertion of a DNA segment into the exon, or a mutation obliterating a splice site [also a plausible explanation (Li 1997) for intron 14 loss in Ag*fEα bis*]. These mechanisms, and other discrete recombination events such as gene conversion, unequal crossing, and mutations leading to loss of amino acids within an exon without functional loss (Li 1997; Cseke *et al.* 1998), could account for the exon variation observed in At**gg**-*copp1* (exons 12, 13, and 15; see Figure 3) and for the partial N-terminal domain loss observed in Cb*glinoh* (including consecutive introns 1, 2, and 3, assuming the ancestral gene was similar in architecture to At**gg**-*copp1* or Ag**gg***abi*). Some uncertainty exists in the evolutionary placement of Cc*glinoh* (branchpoint G; Figure 7) because it is the only monoterpene synthase of angiosperm origin that is included in class I (which contains diterpene synthases involved in gibberellin biosynthesis). This unusual placement suggests that Cc*glinoh* has remained more conserved in structure than any other angiosperm monoterpene synthase, or, as proposed by Cseke *et al.* (1998), that the gene contains a fragment of a copalyl diphosphate synthase (class I) that resembles the ancestral type.

The monoterpene synthase Cc*glinoh* and the sesquiterpene synthase Ag*fEα bis* most closely resemble conifer diterpene synthase genes, all of which contain the CDIS element. Intriguingly, both genes have also lost terminal introns 3 and 14 that are conserved among all other plant terpene synthases. Most likely, both of these genes are defunct diterpene synthases that have retained sufficient 3′ sequence to encode a functional carboxy-terminal active site domain (Starks *et al.* 1997) and have undergone nondeleterious mutations that have altered substrate preference.

Predictions, prospect, and significance: Given the substantial primary sequence differences between gene types, the evolutionary relationship of plant terpene synthase genes to microbial terpene synthase genes and to the mechanistically related prenyl transferases is unclear, although common ancestry has been suggested (Lesburg *et al.* 1997; Wendt *et al.* 1997; Bohlmann *et al.* 1998b; Cane 1999a; Hohn 1999). Predictions based upon genomic structural organization (data not shown) suggest that prenyl transferases and microbial terpene synthases do not share a recognizable, common ancestor with plant terpene synthases. Thus, two farnesyl diphosphate synthase genes from the Arabidopsis database (both contain 10 introns) were compared to several terpene synthases from the present study, and no significant alignment, or conservation of intron placement or phase, was noted. Aside from the conservation of the mechanistically relevant DDXXD sequence motif, prior comparisons of plant terpene synthase genes to micro-

bial terpene synthase genes have not demonstrated significant conservation in deduced amino acid sequence (Bohlmann *et al.* 1997). The genomic organization of microbial terpene synthase genes (Hohn and Beremand 1989; Trapp *et al.* 1998; Hohn 1999) also differs significantly from that of plants, in that the former are only half the length of the latter and contain only 1–2 introns. The clear implication from these preliminary comparisons is that the mechanistically related terpene synthases from these very distant phyla arose by convergent evolution.

This study represents the first attempt to trace the molecular evolutionary history of the large multigene family of plant terpene synthases by comparison of genomic architecture and has predicted the appearance of a plant terpene synthase ancestor that existed prior to the division of angiosperms and gymnosperms and to the separation between primary and secondary terpenoid metabolism. Most likely, this ancestral terpene synthase gene resembled an extant relative of a conifer diterpene synthase of primary metabolism, prior to duplication and differentiation in which conservation of gene organization was maintained for most descendant monoterpene, sesquiterpene, and diterpene synthases. To refine the ancestry and mechanism of evolutionary descent proposed here and to verify the generality of the predictions made will require evaluation of a larger sample size of terpene synthases from throughout the plant kingdom, including triterpene and tetraterpene synthases. The nonvascular plants are of particular interest in this regard, especially the liverworts as an ancient group of land plants that are a rich source of terpenoid natural products (Asakawa 1995) and for which, consistent with the intron-early hypothesis (Li 1997; Mathews and Trotman 1998), a Tps class I gene architecture would be predicted. Data from a wider range of plant species will be important also for more accurately designing domain swapping and site-directed mutagenesis experiments that could permit more detailed understanding of structure-function relationships in this enzyme class. Finally, this new approach to understanding the origins of natural product diversity may have broad implications for plant molecular phylogenetics in general, particularly in the provision of novel molecular markers and criteria for assessing relatedness.

We thank P. Soltis, J. Bohlmann, G. Turner, E. Davis, R. Peters, and M. Wise for helpful technical and editorial suggestions; T. Davis, J. Chappell, and C. A. West for providing access to unpublished sequence information; J. Crock, E. Stauber, J. Davis, and D. Pouchnik for technical assistance; and Joyce Tamura for assistance in manuscript preparation. This work was supported in part by a U.S. Department of Agriculture National Research Initiative grant and by grants from the National Institutes of Health and the U.S. Department of Energy.

LITERATURE CITED

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. **215:** 403–410.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

Asakawa, Y., 1995 *Chemical Constituents of the Bryophytes* (Progress in the Chemistry of Organic Natural Products, Vol. 65, edited by W. Hertz, G. W. Kirby, R. E. Moore, W. Steglich and C. Tamm). Springer, Vienna.

Back, K., and J. Chappell, 1995 Cloning and bacterial expression of a sesquiterpene cyclase from *Hyoscyamus muticus* and its molecular comparison to related terpene cyclases. J. Biol. Chem. **270:** 7375–7381.

Baltimore, D., 1985 Retroviruses and retrotransposons: the role of reverse transcriptase in shaping the eukaryotic genome. Cell **40:** 481–482.

Bharathan, G., B.-J. Janssen, E. A. Kellogg and N. Sinha, 1999 Phylogenetic relationships and evolution of the KNOTTED class of plant homeodomain proteins. Mol. Biol. Evol. **16:** 553–563.

Bohlmann, J., C. L. Steele and R. Croteau, 1997 Monoterpene synthases from grand fir (*Abies grandis*): cDNA isolation, characterization and functional expression of myrcene synthase, (−)-4*S*-limonene synthase and (−)-(1*S*,5*S*)-pinene synthase. J. Biol. Chem. **272:** 21784–21792.

Bohlmann, J., J. Crock, R. Jetter and R. Croteau, 1998a Terpenoid-based defenses in conifers: cDNA cloning, characterization and functional expression of wound-inducible (*E*)-α-bisabolene synthase from grand fir (*Abies grandis*). Proc. Natl. Acad. Sci. USA **95:** 6756–6761.

Bohlmann, J., G. Meyer-Gauen and R. Croteau, 1998b Plant terpenoid synthases: molecular biology and phylogenetic analysis. Proc. Natl. Acad. Sci. USA **95:** 4126–4133.

Bohlmann, J., M. Phillips, V. Ramachandiran, S. Katoh and R. Croteau, 1999 cDNA cloning, characterization, and functional expression of four new monoterpene synthase members of the *Tpsd* gene family from grand fir (*Abies grandis*). Arch. Biochem. Biophys. **368:** 232–243.

Bohlmann, J., D. Martin, N. J. Oldham and J. Gershenzon, 2000 Terpenoid secondary metabolism in *Arabidopsis thaliana*: cDNA cloning, characterization, and functional expressoin of a myrcene/*E*-β-ocimene synthase. Arch. Biochem. Biophys. **375:** 261–269.

Buckingham, J., 1998 *Dictionary of Natural Products on CD-ROM, Version 6.1.* Chapman & Hall, London.

Cane, D. E., 1990 Enzymatic formation of sesquiterpenes. Chem. Rev. **90:** 1089–1103.

Cane, D. E., 1999a Isoprenoid biosynthesis: overview, pp. 1–13 in *Comprehensive Natural Products Chemistry: Isoprenoids Including Steroids and Carotenoids*, Vol. 2, edited by D. E. Cane. Pergamon, Oxford.

Cane, D. E., 1999b Sesquiterpene biosynthesis: cyclization mechanisms, pp. 155–200 in *Comprehensive Natural Products Chemistry Isoprenoids: Including Steroids and Carotenoids*, Vol. 2, edited by D. E. Cane. Pergamon, Oxford.

Chappell, J., 1995a The biochemistry and molecular biology of isoprenoid metabolism. Plant Physiol. **107:** 1–6.

Chappell, J., 1995b Biochemistry and molecular biology of the isoprenoid biosynthetic pathway in plants. Annu. Rev. Plant Physiol. Plant Mol. Biol. **46:** 521–547.

Chen, X. Y., M. Wang, Y. Chen, V. J. Davisson and P. Heinstein, 1996 Cloning and heterologous expression of a second (+)-delta-cadinene synthase from *Gossypium arboreum.* J. Nat. Prod. **59:** 944–951.

Christophersen, C., 1996 Theory of the origin, function, and evolution of secondary metabolites, pp. 677–733 in *Studies in Natural Products*, Vol. 18, edited by Atta-ur-Rahman. Elsevier, Amsterdam.

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

Colby, S. M., W. R. Alonso, E. J. Katahira, D. J. McGarvey and R. Croteau, 1993 4*S*-Limonene synthase from the oil glands of spearmint (*Mentha spicata*): cDNA isolation, characterization and bacterial expression of the catalytically active monoterpene cyclase. J. Biol. Chem. **268:** 23016–23024.

Croteau, R., 1987 Biosynthesis and catabolism of monoterpenoids. Chem. Rev. **87:** 929–954.

Crowell, P. L., and M. N. Gould, 1994 Chemoprevention and

therapy of cancer by *d*-limonene. CRC Crit. Rev. Oncogenesis **5:** 1–22.

Cseke, L., N. Dudareva and E. Pichersky, 1998 Structure and evolution of linalool synthase. Mol. Biol. Evol. **15:** 1491–1498.

Davis, E. M., Y.-S. Chen, M. Essenberg and M. L. Pierce, 1998 cDNA sequence of a (+)-delta-cadinene synthase gene induced in *Gossypium hirsutum* L. by bacterial infection. Plant Physiol. **116:** 1192.

Dawson, F. A., 1994 The Amazing Terpenes. Naval Stores Rev. **March/April:** 6–12.

Derr, L. K., J. N. Strathern and D. J. Garfinkel, 1991 RNA-mediated recombination in *S. cerevisiae*. Cell **67:** 355–364.

Don, R. H., P. T. Cox, B. J. Wainwright, K. Baker and J. S. Mattick, 1991 Touchdown PCR to circumvent spurious priming during gene amplification. Nucleic Acids Res. **19:** 4008.

Doyle, J. A., 1998 Phylogeny of vascular plants. Annu. Rev. Ecol. Syst. **29:** 567–599.

Eckermann, S., G. Schrsder, J. Schmidt, D. Strack, R. A. Edrada *et al.*, 1998 New pathway to polyketides in plants. Nature **396:** 387–390.

Eisenreich, W., M. Schwarz, A. Cartayrade, D. Arigoni, M. H. Zenk *et al.*, 1998 The deoxyxylulose phosphate pathway of terpenoid biosynthesis in plants and microorganisms. Chem. Biol. **5:** R221–R233.

Facchini, P. J., and J. Chappell, 1992 Gene family for an elicitor-induced sesquiterpene cyclase in tobacco. Proc. Natl. Acad. Sci. USA **89:** 11088–11092.

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

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

Fryxell, K. J., 1996 The coevolution of gene family trees. Trends Genet. **12:** 356–369.

Häger, K.-P., B. Müller, C. Wind, S. Erbach and H. Fischer, 1996 Evolution of legumin genes: loss of an ancestral intron at the beginning of angiosperm diversification. FEBS Lett. **387:** 94–98.

Hanley, B. A., and M. A. Schuler, 1988 Plant intron sequences: evidence for distinct groups of introns. Nucleic Acids Res. **16:** 7159–7176.

Harborne, J. B., 1991 Recent advances in the ecological chemistry of plant terpenoids, pp. 396–426 in *Ecological Chemistry and Biochemistry of Plant Terpenoids*, edited by J. B. Harborne and F. A. Tomas-Barberan. Clarendon Press, Oxford.

Hohn, T. M., 1999 Cloning and expression of terpene synthase genes, pp. 201–215 in *Comprehensive Natural Products Chemistry: Isoprenoids Including Steroids and Carotenoids*, Vol. 2, edited by D. E. Cane. Pergamon, Oxford.

Hohn, T. M., and P. D. Beremand, 1989 Isolation and nucleotide sequence of a sesquiterpene cyclase gene from the trichothecene-producing organism *Fusarium sporotrichioides*. Gene **79:** 131–138.

Holmes, F. A., A. P. Kudelka, J. J. Kavanagh, M. H. Huber, J. A. Ajani *et al.*, 1995 Current status of clinical trials with paclitaxel and docetaxel, pp. 31–57 in *Taxane Anticancer Agents: Basic Science and Current Status*, edited by G. I. Georg, T. T. Chen, I. Ojima and D. M. Vyas. American Chemical Society Symposium Series 583, Washington, DC.

Huang, N., T. D. Sutliff, J. C. Litts and R. L. Rodriguez, 1990 Classification and characterization of the rice α-amylase multigene family. Plant Mol. Biol. **14:** 655–668.

Jarvis, B. B., 2000 The role of natural products in evolution, pp. 1–24 in *Recent Advances in Phytochemistry-Evolution of Metabolic Pathways*, Vol. 34, edited by J. T. Romeo, R. K. Ibrahim, L. Varin and V. DeLuca. Plenum Press, NY.

Jarvis, B. B., and J. M. Miller, 1996 Natural products, complexity, and evolution, pp. 265–293 in *Phytochemical Diversity and Redundancy in Ecological Interactions*, edited by J. T. Romeo. Plenum Press, NY.

Katoh, S., and R. Croteau, 1998 Individual variation in constitutive and induced monoterpene biosynthesis in grand fir (*Abies grandis*). Phytochemistry **47:** 577–582.

Koepp, A. E., M. Hezari, J. Zajicek, B. Stofer Vogel, R. E. LaFever *et al.*, 1995 Cyclization of geranylgeranyl diphosphate to taxa-4(5),11(12)-diene is the committed step of Taxol biosynthesis in Pacific yew. J. Biol. Chem. **270:** 8686–8690.

Koyama, T., and K. Ogura, 1999 Isopentenyl diphosphate isomerase and prenyltransferases, pp. 69–96 in *Comprehensive Natural Products Chemistry: Isoprenoids Including Steroids and Cartenoids*, Vol. 2, edited by D. E. Cane. Pergamon, Oxford.

Kumar, V., and M. Trick, 1993 Sequence of the *S* receptor kinase gene family in *Brassica*. Mol. Gen. Genet. **241:** 440–446.

Lesburg, C. A., G. Zhai, D. E. Cane and D. W. Christianson, 1997 Crystal structure of pentalenene synthase: mechanistic insights on terpenoid cyclization reactions in biology. Science **277:** 1820–1824.

Lewinsohn, E., T. Savage, M. Gijzen and R. Croteau, 1993 Simultaneous analysis of monoterpenes and diterpenoids of conifer oleoresin. Phytochem. Anal. **4:** 220–225.

Li, W., 1997 *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Lichtenthaler, H. K., 1999 The 1-deoxy-D-xylulose-5-phosphate pathway of isoprenoid biosynthesis in plants. Annu. Rev. Plant Physiol. Plant Mol. Biol. **50:** 47–66.

MacMillan, J., and M. Beale, 1999 Diterpene biosynthesis, pp. 217–243 in *Comprehensive Natural Products Chemistry: Isoprenoids Including Steroids and Carotenoids*, Vol. 2, edited by D. E. Cane. Pergamon, Oxford.

Mapleston, R. A., M. J. Stone and D. H. Williams, 1992 The evolutionary role of secondary metabolites—a review. Gene **115:** 151–157.

Mathews, C. M., and C. A. N. Trotman, 1998 Ancient and recent intron stability in *Artemia* hemoglobin gene. J. Mol. Evol. **47:** 763–771.

Mathews, S., M. Lavin and R. A. Sharrock, 1995 Evolution of the phytochrome gene family and its utility for phylogenetic analysis of angiosperms. Ann. Mo. Bot. Gard. **82:** 296–321.

Mau, C. J. D., and C. A. West, 1994 Cloning of casbene synthase cDNA: evidence for conserved structural features among terpenoid cyclases in plants. Proc. Natl. Acad. Sci. USA **91:** 8479–8501.

Newman, J. D., and J. Chappell, 1999 Isoprenoid biosynthesis in plants: carbon partitioning within the cytoplasmic pathway. Crit. Rev. Biochem. Mol. Biol. **34:** 95–106.

Ogura, K., and T. Koyama, 1998 Enzymatic aspects of isoprenoid chain elongation. Chem. Rev. **98:** 1263–1276.

Patterson, A. H., C. L. Brubaker and J. F. Wendel, 1993 A rapid method for extraction of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis. Plant Mol. Biol. Rep. **11:** 122–125.

Qureshi, N., and J. W. Porter, 1981 Conversion of acetyl-Coenzyme A to isopentenyl pyrophosphate, pp. 47–94 in *Biosynthesis of Isoprenoid Compounds*, Vol. 1, edited by J. W. Porter and S. L. Spurgeon. John Wiley & Sons, New York.

Ramos-Valdivia, A. C., R. van der Heiden and R. Verpoorte, 1997 Isopentenyl diphosphate isomerase: a core enzyme in isoprenoid biosynthesis. A review of its biochemistry and function. Nat. Prod. Rep. **14:** 591–603.

Sambrook, J., E. F. Fritsch and T. Maniatis, 1989 *Molecular Cloning: A Laboratory Manual*, Ed. 2. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schrsder, J., 1997 A family of plant-specific polyketide synthases: facts and predictions. Trends Plant Sci. **2:** 373–378.

Schrsder, J., 1999 The chalcone/stilbene synthase-type family of condensing enzymes, pp. 749–771 in *Comprehensive Natural Products Chemistry: Polyketides and Other Secondary Metabolites Including Fatty Acids and Their Derivatives*, Vol. 1, edited by U. Sankawa. Elsevier Science, Oxford.

Starks, C. M., K. Back, J. Chappell and J. P. Noel, 1997 Structural basis for cyclic terpene biosynthesis by tobacco 5-epi-aristolochene synthase. Science **277:** 1815–1820.

Steele, C. L., J. Crock, J. Bohlmann and R. Croteau, 1998 Sesquiterpene synthases from grand fir (*Abies grandis*). Comparison of constitutive and wound-inducible activities, and cDNA isolation, characterization, and bacterial expression of δ-selinene synthase and γ-humulene synthase. J. Biol. Chem. **273:** 2078–2089.

Stofer Vogel, B., M. Wildung, G. Vogel and R. Croteau, 1996 Abietadiene synthase from grand fir (*Abies grandis*): cDNA isolation, characterization and bacterial expression of a bifunctional diterpene cyclase involved in resin acid biosynthesis. J. Biol. Chem. **271:** 23262–23268.

Sun, T., and Y. Kamiya, 1994 The Arabidopsis GA1 locus encodes the cyclase *ent*-kaurene synthetase A of gibberellin biosynthesis. Plant Cell **6:** 1509–1518.

Sun, T. P., H. M. Goodman and F. M. Ausubel, 1992 Cloning the

Arabidopsis GA1 locus by genomic subtraction. Plant Cell **4:** 119–128.

Trapp, S., T. M. Hohn, S. McCormick and B. B. Jarvis, 1998 Characterization of the gene cluster for biosynthesis of macrocyclic trichothecenes in *Myrothecium roridum*. Mol. Gen. Genet. **257:** 421–432.

Turner, G., 1993 Gene organization in filamentous fungi, pp. 107–125 in *The Euykaryotic Genome: Organization and Regulation*, edited by P. M. A. Borda, S. Oliver and P. F. G. Sims. Cambridge University Press, New York.

Van Geldre, E., A. Vergauwe and E. Van den Eeckhout, 1997 State of the art of the production of the antimalarial compound artemisinin in plants. Plant Mol. Biol. **33:** 199–209.

Waters, E. R., and E. Vierling, 1999 The diversification of plant cytosolic small heat shock proteins preceded the divergence of mosses. Mol. Biol. Evol. **16:** 127–139.

Wendt, K. U., K. Poralla and G. E. Szhulz, 1997 Structure and function of a squalene cyclase. Science **277:** 1811–1815.

West, C. A., 1981 Biosynthesis of diterpenes, pp. 375–411 in *Biosynthesis of Isoprenoid Compounds*, Vol. 1, edited by J. W. Porter and S. L. Spurgeon. John Wiley & Sons, New York.

Wildung, M. R., and R. Croteau, 1996 A cDNA clone for taxadiene synthase, the diterpene cyclase that catalyzes the committed step of Taxol biosynthesis. J. Biol. Chem. **271:** 9201–9204.

Williams, D. C., D. J. McGarvey, E. J. Katahira and R. Croteau, 1998 Truncation of limonene synthase preprotein provides a fully active ''pseudomature'' form of this monoterpene cyclase and reveals the function of the amino-terminal arginine pair. Biochemistry **37:** 12213–12220.

Williams, D. H., M. J. Stone, P. R. Hauck and S. K. Rahman, 1989 Why are secondary metabolites (natural products) synthesized? J. Nat. Prod. **52:** 1189–1208.

Wise, M. L., and R. Croteau, 1999 Monoterpene biosynthesis, pp. 97–153 in *Comprehensive Natural Products Chemistry: Isoprenoids Including Steroids and Carotenoids*, Vol. 2, edited by D. E. Cane. Pergamon, Oxford.

Yamaguchi, S., T. P. Sun, H. Kawaide and Y. Kamiya, 1998 The GA2 locus of *Arabidopsis thaliana* encodes *ent*-kaurene synthase of gibberellin biosynthesis. Plant Physiol. **116:** 1271–1278.

Yuba, A., K. Yazaki, M. Tabata, G. Honda and R. Croteau, 1996 cDNA cloning, characterization, and functional expression of 4*S*-(−)-limonene synthase from *Perilla frutescens*. Arch. Biochem. Biophys. **332:** 280–287.

Zinkel, D. F., and J. Russell, 1989 *Naval Stores: Production, Chemistry, Utilization*. Pulp Chemicals Association, New York, 1060 pp.

Communicating editor: V. L. Chandler