Full Paper

DOI: 10.1002/minf.201100087

A Mechanism for Evolving Novel Plant Sesquiterpene Synthase Function

Troy Wymore,*[a] Brian Y. Chen,[a] Hugh B. Nicholas Jr.,[a] Alexander J. Ropelewski,[a] and Charles L. Brooks III[b]

Abstract: Plant sesquiterpene synthases, a subset of the terpene synthase superfamily, are a mechanistically diverse family of enzymes capable of synthesizing hundreds of complex compounds with high regio- and stereospecificity and are of biological importance due to their role in plant defense mechanisms. In the current report we describe a large-scale, high-resolution phylogenetic analysis of ~200 plant sesquiterpene synthases integrated with structural and experimental data that address these issues. We observe that all sequences that cluster together on the phylogenetic tree into well-defined groups share at least the first reaction in the catalytic mechanism subsequent to the initial ionization step and many share steps beyond this down to proton transfers between the enzyme and substrate. Most significant is the previously unreported high conservation of an Asp-Tyr-Asp triad. Due to its high conservation, patterns in the phylogenetic tree as well as experimental and modeling results, we suggest that this Asp-Tyr-Asp triad is an important functional element responsible for many proton transfers to and from the substrate and intermediates along the plant sesquiterpene synthase catalytic cycle and whose position can be tuned by residues outside the active site that can lead to the evolution of novel enzyme function.

Keywords: Sesquiterpene synthases · Molecular evolution · Phylogenetics · Enzymes · Structure-function relationships

1 Introduction

John Maynard Smith described protein evolution with great insight as a mutational search through a continuous network of functional intermediates.[1] Several classic studies of molecular evolution demonstrate that in many cases enzymes that carry out a specific function evolved from promiscuous ancestors;[2–4] i.e. able to act on different substrates or synthesize different products even if at very low efficacy. Often this evolution of a novel function is initiated with a gene duplication event.[5] A gauge of a protein's ability to traverse this functional landscape to develop a novel function is often referred to as its evolvability or plasticity. The plant sesquiterpene synthases (sTSs), a subset of the Terpene Synthase (TS) superfamily, are a highly evolvable family, since enabling adaptations is crucial for synthesizing compounds that protect against fast-evolving microbial pathogens and mediating other plant-environment interactions.[6] Plant sTSs span a large range of specifity from the *Nicotiana tabacum* (tobacco) 5-*epi*-aristolocholene synthase (5EAS), generating primarily one product,[7] to the very promiscuous *Abies grandis* (grand fir) sesquiterpene synthase that produces 52 different compounds.[8] These characteristics, among others, make the study of plant sTSs interesting targets for explorations of fundamental questions about enzyme evolution and evolvability.

Terpenes represent the largest and most diverse compilation of plant natural products with over 30 000 known compounds[9] and have uses as precursors to pharmaceutical agents, insecticides and fragrances besides their biological function in plants.[10] They are generated through catalysis by mono-, sesqui- and di- terpene synthases (sometimes called cyclases) that act on the acyclic isoprenoid substrates geranyl diphosphate (10 carbons), farnesyl diphosphate (15 carbons) or geranylgeranyl diphosphate (20 carbons) respectively. The focus of this report is on plant sesquiterpene synthases (sTSs) though TSs also exist within bacterial and fungal species.[11]

Plant sTSs have two domains, the N-terminal domain which has some structural similarity to glycosylhydrolases[12] and the C-terminal domain containing the active site cavity that also binds three divalent Mg^{2+} ions.[11] The function of the N-terminal domain is not precisely known though some results suggest it plays a role in folding.[13] This domain is not present in bacterial and fungi sTSs. Only three plant sTSs sequences have structural models from x-ray crystallography deposited in the PDB;[14] the *Nicotiana*

[a] *T. Wymore, B. Y. Chen, H. B. Nicholas Jr., A. J. Ropelewski*
National Resource for Biomedical Supercomputing, Pittsburgh Supercomputing Center
300 South Craig Street, Pittsburgh, PA 15213, USA
phone: 412-268-4960; fax: 412-268-8200
**e-mail: wymore@psc.edu*

[b] *C. L. Brooks III*
Department of Chemistry and Biophysics, University of Michigan
930 North University Avenue, Ann Arbor, MI 48109, USA

Supporting Information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201100087.

© 2011 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

tabacum 5EAS,[12] a four substitution mutant of this enzyme[15] and the *Gossypium arboreum* (+)-δ-cadinene synthase (DCS).[16] Of particular interest is the difference in the binding mode of farnesyl diphosphate substrate analogues. In the 5EAS structure, the hydrocarbon moiety is folded and appears poised for catalysis while in the DCS structure the hydrocarbon moiety binds unfolded in a secondary binding site. The biosynthesis of sesquiterpenes is initiated by ionization of the C1-OPP bond generating a reactive carbocation (see Figure 1) and the first major important branching of the mechanistic network. At this point, the diphosphate moiety can then bind to C3, isomerize about the C2-C3 bond, and then ionize the C3-OPP bond to form the reactive nerolidyl carbocation. At any point along the mechanistic decision network, carbocations can be quenched by proton transfer from the intermediate to the enzyme or by addition of water molecules. Otherwise, the chemistry can proceed through an intermolecular electrophilic attack on one of the two double bonds of the substrate to form a cyclic species. Possible subsequent reactions include proton transfers, hydride shifts, further intermolecular electrophilic additions, and methyl or methylene shifts. All of these reactions often occur with regio- and stereospecific control giving rise to a diversity of plant sTS products (see Figure 1). Much is still unknown about the atomic details surrounding these mechanisms, though the intrinsic reactivity of the substrate has in many cases been elucidated through quantum chemical calculations.[17,18] What is known has led researchers to classify TSs as a template and chaperone for the various cyclization reactions[19] and as a dancer partnering with the substrate traversing a mechanistic landscape.[20] The enzymes' contribution to catalysis may be relatively subtle based on residues appearing to act in concert and ones outside the active site affecting product distributions.[21,22] The role of conformational fluctuations in response to various intermediates along the catalytic cycle may also be important.[23]

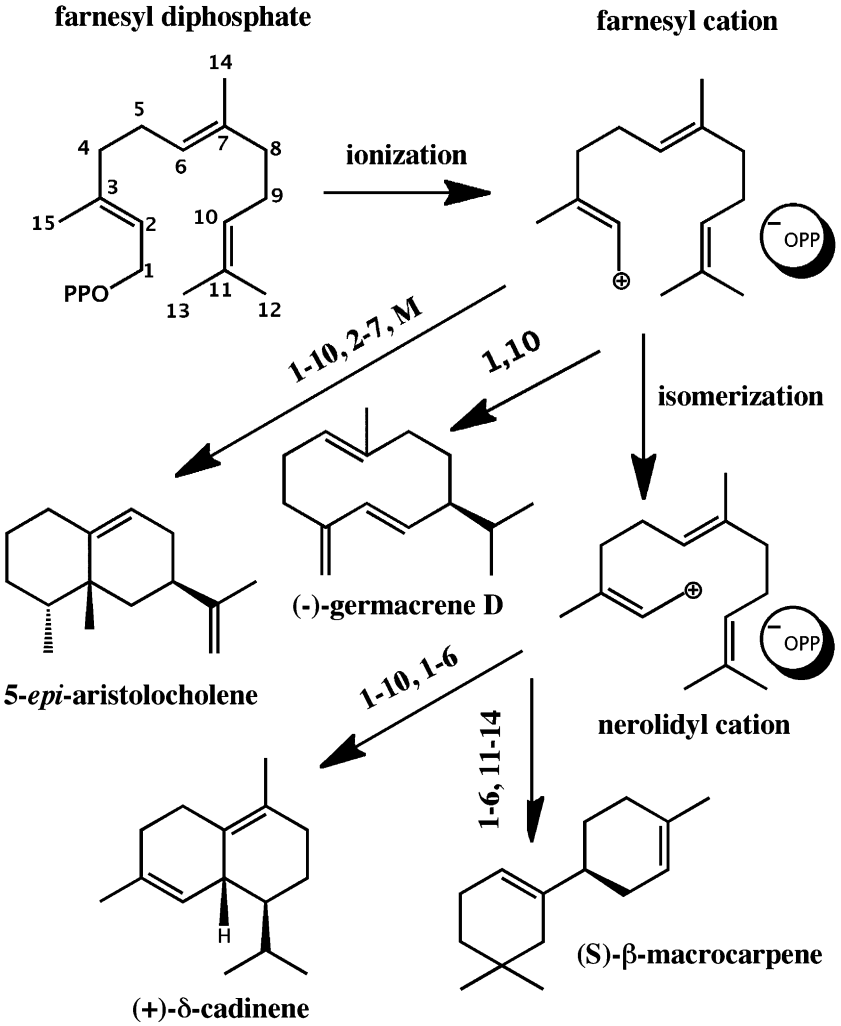

Figure 1. Reaction mechanism of sesquiterpene synthases begins with the ionization of farnesyl diphosphate which in some isozymes isomerizes to the nerolidyl cation. From the farnesyl and nerolidyl cation, cyclization reactions, proton and hydride transfers and Wagner-Meerwein rearrangements (M) may occur leading to a sampling of the products shown.[30,32]

© 2011 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

The number and complexity of sesquiterpene products has inhibited the use of phylogenetic analysis for constructing sequence-function relationships because of uncertainties in the statistical relevance of the resulting inferences. Since all plant TSs have a common evolutionary origin,[24] phylogenetic analysis has typically included mono-, sesqui- and diterpene synthase sequences. In an early review by Bohlmann et al.,[25] 28 TSs were placed into six subfamilies designated TpsA through TpsF, each sharing 40% identity between members. The TpsA and TpsB subfamilies consisted of angiosperm sesquiterpene and monoterpene synthases respectively. The TpsD subfamily consisted of gymnosperm mono-, sesqui- and diterpene synthases revealing that gymnosperm TSs are more related to each other than to their angiosperm counterparts. Phylogenetic analyses have also been performed as part of some new TS sequence characterizations.[26–28] While these studies have tentatively provided an evolutionary history between the different TSs, they have not resulted in relating sequence features to steps in the cyclization mechanisms beyond the initial ionization step. With such a low number of characterized sequences and a diversity of products, such early attempts may not have even been fruitful. A classification scheme was also developed by comparison of intron/exon patterns resulting in three classes,[29] though modifications to this scheme have been subsequently proposed.[27] Given the enormous complexity of this enzyme superfamily, it would seemingly make sense to take a reductionist approach and attempt to understand more fully a subset of the superfamily as exhibited in an excellent recent review by Degenhardt, Köllner and Gershenzon[30] on plant monoterpene synthases and sTSs. In this review, they also constructed a phylogenetic tree of plant sTSs using sequences with characterized product distributions and restricting the analysis to 45 residues in the active site. From this analysis, they concluded that no overall feature of plant sTSs exists that distinguishes multiple from single product isozymes, enzymes producing acyclic products had multiple evolutionary origins and there was not an absolute split between gymnosperms and angiosperms. Other excellent reviews that cover various aspects of TSs are also available.[31–33]

Though it has been claimed and often repeated that constructing sequence-function relationships in plant sTSs based on phylogenetic analysis is extremely difficult,[34] the validity of this claim is most certainly dependent on the phylogenetic analysis procedure and to what extent the analysis can be useful. For example, do closely related sequences according to the phylogenetic analysis share *any* common mechanistic steps? In this report, we begin to elucidate the evolutionary relationship between plant sTSs through a large-scale high-resolution phylogenetic analysis and the construction of enzyme structural models. Our findings are critical for placing the results of site-directed mutagenesis experiments in context of the entire plant sTS family and for making decisions on the type and location of residues to mutate that would yield the most insight. Finally, our results further demonstrate how plant sTSs evolve novel functions through substitutions to residues outside the active site.

2 Computational Methods

2.1 Sequence-Based Bioinformatics

Sequences were gathered from the UniProt database[35] through an inclusive text search for terpene synthase and through a recent review article on plant sTSs.[30] Sequences of lengths between 535–615 residues were retained, a range typical of plant sTSs. Monoterpene Synthases that were gathered by this criteria and have a relatively close relationship to sTSs were deleted in order to focus our structure-function analyses on plant sTS. A subset of 40 representative sequences was extracted using CD-HIT[36] and a multiple sequence alignment (MSA) was constructed using ClustalW.[37] The Meme program[38] was employed to search for the 20 most conserved patterns or motifs amongst the sequences. The alignment was then manually adjusted using the motifs and structural information as a guide within the GeneDoc editor.[39] Sequences were then added to the MSA (in sets of 20–30) using ClustalW to perform a profile alignment to the previous edited MSA and MEME rerun over the new expanded set of sequences. The new MSA was manually adjusted as before and the process was repeated until all sequences were included. Sequences missing sections of amino acids (larger than five residues) in conserved motifs were then deleted resulting in a final MSA of 189 sequences. These sequences and others with functional annotation not supported by experimental data are noted in the Supplemental Information on Problematic Sequences along with the final MSA labeled by their GenBank identifiers, a MSA figure of a selected subset of sequences, and a table that provides extra information on all sequences used in the analysis.

The quality of a phylogenetic tree is highly dependent on the quality of the MSA and the regions included for tree construction.[40,41] The final MSA was manually trimmed to 494 characters by deleting sections where the alignment was equivocal, primarily at the N- and C-terminal. All active site residues were included. One variable region spanning the J-K loop was retained since this may be an important distinguishing characteristic among groups of sequences. The trimmed MSA file was used to create a distance-based phylogenetic tree using the PHYLIP suite software.[42] 2500 data sets were generated through bootstrapping analysis using SEQBOOT. A distance matrix was computed for each of these data sets with the program PROTDIST using the Dayhoff PAM matrix. The resulting distance matrix data was used with NEIGHBOR to generate a tree for each of the 2500 data sets, all of which were run through CONSENSE to produce a consensus tree. The phylogenetic tree was visualized and a figure created with FigTree version 1.2.2. The retrieval of functional characteristics for all sequences,

 © 2011 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

ordering of the MSA file based on the phylogenetic tree and other information gathering tasks was facilitated by our program HarvestSeq. Finally, given an alignment and a set of protein sequences grouped according to some definition of function; such as whether the enzyme mechanism goes through the nerolidyl cation or not, the "sub-profile" analysis method[43,44] identifies positions that are correlated with functional differences. This analysis was performed with the program GEnt.[44] Various groupings of protein sequences were used in this analysis depending on the extent of similarity in the mechanistic network.

2.2 Molecular Modeling of Germacrene A synthases

Structural models were constructed by extracting the sequence to be modeled and the template 5EAS sequence from the MSA. The sequence alignment and template structure (PDB entry 5eat) were used as input to the program MODELLER version 9.4.[45] The best of five models according to the DOPE score[46] was used to rebuild the sidechains using SCWRL 4.0[47] with the ligands farnesylhydroxyphosphonate (FHP) and Mg^{2+} ions as steric boundaries. Stereo-visualization and comparison of structural models was done with VMD 1.8.6.[48] VMD was used to create the figure of conserved residues. Reaction schemes were created with ChemDraw.

2.3 Molecular Dynamics Simulations of the Eudesmane Carbocation Intermediate Bound to the Solvated 5-epi-Aristolocholene Synthase

Models of the eudesmane carbocation intermediate in the active site of 5-epi-aristolocholene synthase were constructed using the PDB coordinates from 5EAT that contains a co-crystallized Farnesyl Hydroxyphosphonate (FHP) molecule and three Mg^{2+} ions. The bicyclic eudesmane carbocation intermediate was placed so as to overlap with the folded FHP as much as possible. This initial structure appears to have many of the features described in the structure publication[12] such as the proximity of the C8 protons to Trp273 and the C6 protons to Tyr520. All subsequent energy minimizations and simulations were performed with CHARMM version c34b2.[49] The bond, angle, dihedral, Urey-Bradley and Lennard-Jones terms for a molecular mechanics force field of eudesmane carbocation was constructed through analogy to cholesterol parameters contained in the CHARMM force field.[50] The charges were taken from a CHELPG calculation performed at RHF/6-31G(d). In addition, the pyrophosphate moiety that is produced from the initial ionization reaction was placed so as to overlap with the Hydroxyphosphanate moiety and the MM force field was constructed through analogy to other phosphates given in the CHARMM force field. The positions of the oxygen atoms from water molecules observed in the electron density maps were maintained and 11 Å of water molecules (TIP3P model)[51] surrounded the entire system resulting in a system size of ~78 K atoms. The system was subsequently minimized using the CHARMM all-27 protein force field with CMAP corrections[52] for 1000 Steepest Descent steps. Periodic Boundary Conditions, Particle Mesh Ewald (~1 grid point/1 $Å^3$), SHAKE on all hydrogen atoms bonded to heavy atoms, and a 5.0 kcal/(mol Å) restraint on all protein heavy atoms was employed. Next molecular dynamics simulations were initiated by heating the system to 300 K over 30 picoseconds (ps) by scaling of velocities every 1000 steps. Constant pressure simulations were performed for 40 ps to equilibrate the box size. Finally the simulations were run for 10 nanoseconds (ns), the analysis was performed over the last 9 ns. Full results from this simulation and others are the subject of a manuscript in preparation.

3 Results

Conserved Residues in plant sTSs from multiple sequence alignment (MSA). Throughout the rest of the text, residue numbering if not explicitly stated otherwise refers to the *Nicotiana tabacum* 5-*epi* aristolocholene synthase (5EAS, Genbank: L04680) for which a structure is known (PDB entry 5eat).[12]

Our MSA (provided as Supporting Information) of plant sTSs reveals 15 highly conserved (95% or above, see Figure 2 and 3) residues on the C-terminal catalytic domain. These include the Asp residues of the DDxxD motif ($D^{301}D^{302}xxD^{305}$) forming the signature Mg^{2+} metal binding site and two Arg residues (Arg-264 and Arg-441) responsible for binding the phosphate moiety of farnesyl diphosphate. One Arabidopsis and two Solidago sequences have substituted an Asn for Asp-301 while only one Arabidopsis sequence contains an Asn for Asp-302. One Capsicum annuum (red pepper) sequence reports a Glu for Asp-305.

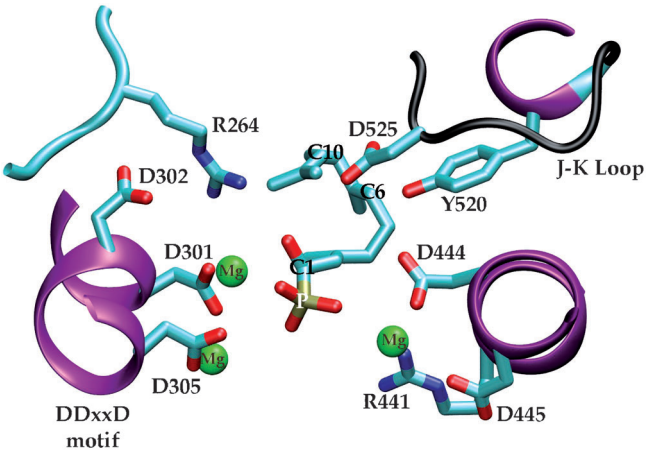

Figure 2. Highly conserved (>95%) residues highlighted on the 5-epi-aristolocholene synthase active site structure (PDB entry: 5eat). Also shown are the cocrystallized Mg^{2+} ions (green) and farnesylhydroxyphosphanate substrate mimic (middle). Helices are in purple.

© 2011 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

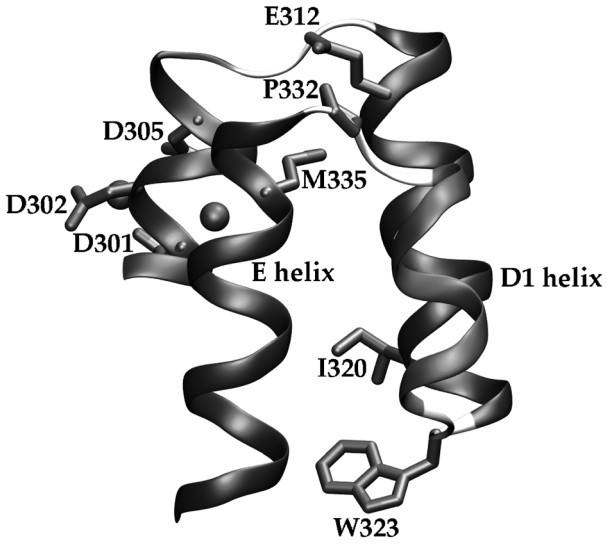

Figure 3. Highly conserved residues outside the active site and their relation to the DDxxD Mg^{2+} binding site.

Four sequences report a Lys in place of Arg-264 while Arg-441 is strictly conserved.

Other highly conserved residues include Asp/Asn-444, Asp-445, Tyr-520 and Asp-525. Asp444–445 are part of a (N,D)DxxX[448] second Mg^{2+} binding motif. In many plant terpenoid cyclases this motif occurs as DDxxTxxxE including some plant sTSs. The residue at position 448 is always one of five small or acidic residues: Gly, Ser, Thr, Asp, Glu. Plant sTSs that go through a (+)-germacrene A product/intermediate all have Thr, the (+)-δ-cadinene synthases have Glu while several conifers have Asp. The Gossypium (+)-δ-cadinene synthase has severely compromised activity when this residue is mutated to Ala.[16] Asp-444 is replaced by Asn in many monocots including all the (E)-β-caryophyllene synthases. Tyr-520 is positioned at the end of the J helix between Asp-444 and Asp-525 and has been shown to be an important for generating the (+)-germacrene A intermediate in 5EAS.[53] Yet, its high conservation among the plant sTSs has to our knowledge not been previously reported. Asp-525 is part of a loop region (the J-K loop) connecting the J and K helices. This residue is replaced by Glu in five Arabidopsis thaliana and two Medicago truncatula sequences. In all plant sTSs we have gathered, the loop region between these two highly conserved residues is between 3 and 4 residues.

Conserved residues outside the active site. Other highly conserved residues include many outside the active site that may be important for maintaining secondary structure and for stability of the protein (see Figure 3). These include Arg-266 and Glu-269 (which sit on the same side of a helix downstream from highly conserved Arg-264), Glu-312 and Trp-323 (which are all part of the D1 helix), Pro-332 (in a loop region connecting the D2 and E helix) and Trp-382 near the end of the F helix. Ile-320 (part of the D1 helix) and Met-335, which interacts with the D1 helix are also highly conserved if we expand our definition to include similar types of amino acids. The Glu-312/Ile-320 pair is replaced by Asp/Phe in several monocot (mostly maize) sequences, and by Gln/Val in conifers revealing possible co-evolution of this sequence pair. Met-335 is replaced by Ile in the asterid (+)-germacrence A synthases and Leu in monocot (E)-β-caryophyllene synthases.

Phylogenetic Analysis of plant sTSs. A consensus phylogenetic tree was constructed from the trimmed plant sTS sequence alignment (see Figure 4 and Table 1). The tree consists of four main branches; the Rosid/Asterid, Monocotyledoneae, Coniferae (mostly gymnosperms) and the Arabidopsis/Lamiids branch labeled A-D in both Figure 4 and Table 1. Due to the low bootstrap values, it is uncertain whether all the sequences within either the A or D branch should be grouped together. However, further along all the branches are sequences that have been grouped with high bootstrap support. There is a mixing of sequences that are grouped according to species with others that are grouped in higher taxonomic levels (see Table 1 and Supporting Information). Of particular overall significance, *all sequences that cluster together on the phylogenetic tree to form well-defined groups (high bootstrap support) share in common at least the first reaction in the catalytic mechanism beyond the initial ionization step* (for example, 1–6, 1–10, 1–11 or no cyclization, see Table 1). Similar to the phylogenetic study by Degenhardt, Köllner and Gershenzon,[30] we have not refined this analysis to include where and when proton/hydride transfers take place since correlations can be seen without going to this level of detail, even though these reactions are obviously an important aspect of the sesquiterpene synthase chemistry; one which could possibly be addressed as more plant sTS sequences and their product distributions become known. An additional striking result is the clear grouping of and distinction between the Solanoideae vetispiradiene synthases (VDSs, D1), the Solanceae 5EASs (D2), and the Asteraceae (+)-germacrene A synthases (D6) with high bootstrap support. All pass through the neutral (+)-germacrene A species but the 5EASs and VDSs continue on in their biosynthetic pathway by protonating the (+)-germacrene A intermediate, diverging only in the last two steps of the catalytic cycle. The sequence features responsible for interconverting between 5EASs and VDSs include many outside the active site.[21,22] This phylogenetic analysis strongly suggests that sequence features can be discovered to interconvert plant sTSs that are even more distantly related such as between 5EASs and (+)-germacrene A synthases and determine what features allow the (+)-germacrene A to be released instead of reprotonated. The Asteraceae (+)-germacrene A synthases are also noteworthy in that they are composed of six asterid species suggesting they are orthologous. This grouping was also observed in a phylogenetic tree constructed only from active site residues.[30] Any two members of this group share between 61 and 95% sequence identity. Finally, it is interesting that all (+)-germacrene A synthases (except

 © 2011 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

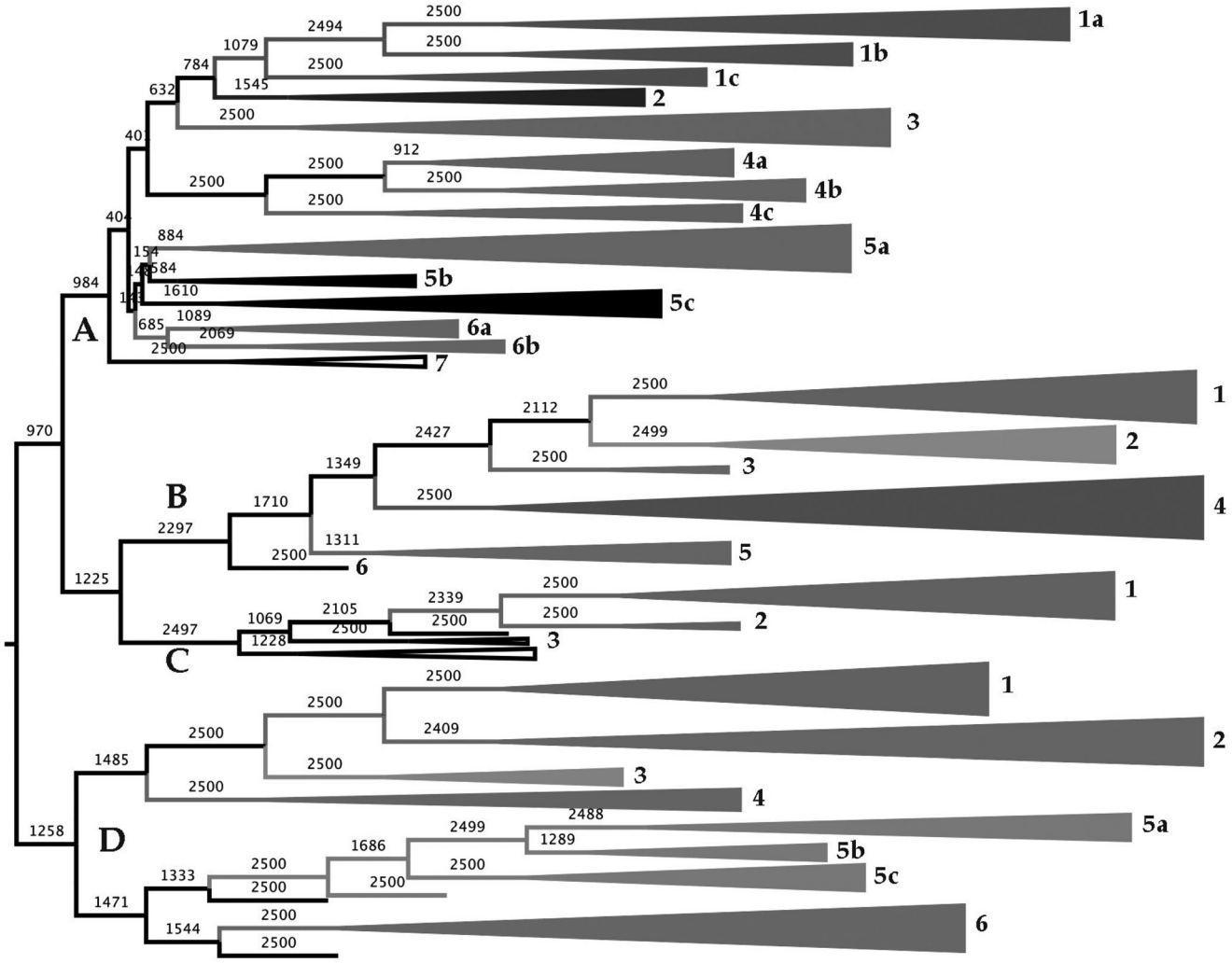

Figure 4. Phylogenetic tree of 187 plant sTS sequences labeled by groups (see Table 1). *Major branches are labeled A-D. The uppermost group located in the A branch is thus labeled A1a in Table 1.* Bootstrap values are given out of a maximum 2500. Further information about all sequences in the order of their location on the tree (top to bottom) including their GenBank ids are given in the supplemental information.

one) reside in the Lamiids (D) branch while all other germacrene synthases reside in the Rosid/Asterid (A) and Zingiber (B5) branches. The different germacrenes are distinguishable primarily by where the proton is removed from the germacradienyl cation.

In addition, groups within both the Monocotyledoneae (B) and Coniferae (C, gymnosperm) branches that do primarily initial 1–6 or 1–11 cyclization can be distinguished from each other. The gymnosperms (C branch) are clustered together separate from the rest of the angiosperms though six angiosperm sequences (C3, five eudicots and one monocot) are clearly closely related (high bootstrap support). Interestingly, these angiosperms do either initial 1–6 or 1–11 cyclization (they actually form different subbranches within the C3 group) while the majority of angiosperm sTSs classified so far do 1–10 cyclization. The Zingiber (B5) sequences that branch off first in the Monocot (B) branch also do 1–10 initial cyclization. The structure of the phylogenetic tree and the fact gymnosperms are more ancient than monocots which are themselves more ancient than eudicots, suggest that plant sTS may have originally performed 1–11 and 1–6 initial cyclization reactions. Plant sTSs that performed 1–10 initial cyclizations may have evolved in the monocot branch and were adopted more heavily by eudicot species.

Finally, because of the correlation of groups of sequences with biochemical function, the sub-profile analysis method[43,44] was employed. Over many protein families, this analysis often identifies residues most responsible for substrate selectivity. Yet, since sTSs all bind the same farnesyl diphosphate substrate, the sub-profile analysis may instead identify residues most responsible for product specificity. We focused on the analysis of groups D1, D2, and D6 (Vetispiradiene, 5-epi-aristolocholene and (+)-germacrene A

© 2011 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1. Groups of plant sTS sequences based on phylogenetic tree in Figure 4. (*)-low bootstrap support (> 50 %) for these sequences to be classified as a group.

Group	Group size	Classification	Representative Product(s)	Cyclization(s)
A1a	7	*Vitis vinifera*	(+)-Valencene/unclassified	1–10, 2–7
A1b	5	*Vitis vinifera*	Unclassified	–
A1c	4	*Vitis vinifera*	Unclassified	–
A2	4	*Pogostemon cablin*	(−)-Germacrene D	1–10
A3	8	*Gossypium*	(+)-δ-Cadinene	1–10, 1–6
A4a*	6	*Artemisia annua*	Amorpha-4,11-diene	1–6, 1–10
A4b	5	*Solidago canadensis*	Germacrenes	1–10
A4c	4	*Artemisia annua*	(E)-β-Farenesene	None
A5a*	10	Lamiids	Germacrenes, δ-elemene	1–10
A5b*	3	Rosids/Asterids	Diverse	1–10/1–11
A5c	6	Rosids/Asterids	(−)-Germacrene D	1–10
A6a*	4	Fabids	α-Farenesene	None
A6b	3	*Medicago truncatula*	(E)-β-Caryophyllene	1–11, 2–10
A7	3	*Citrus*	Valencene, (E)-β-farnesene	1–10/None
B1	11	Poaceae	(E)-β-Caryophyllene	1–11, 2–10
B2	8	Poaceae	(−)-β-Macrocarpene	1–6
B3	2	Liliopsida	Unclassified	–
B4	13	Poaceae	(E,E)-Farnesol/unclassified	None
B5	5	Liliopsida	Diverse	1–10
B6	1	*Magnolia grandiflora*	β-Cubebene	1–10, 1–6, 6–2
C1	10	Pinaceae	Longifolene/δ-selinene	1–11/1–10
C2	2	Pinaceae	(E)-α-Bisabolene	1–6/None
C3*	6	Angiosperms		1–11/1–6 /None
D1	11	Solanoideae	Vetispiradiene	1–10, 2–7
D2	10	Solanaceae	5-*epi*-Aristolochene	1–10, 2–7
D3	4	Lamiaceae	(+)-Germacrene A	1–10
D4	5	Lamiaceae	Diverse	1–10/None
D5a	6	*Arabidopsis thaliana*	Unclassified	–
D5b	4	*Arabidopsis thaliana*	(Z)-γ-Bisabolene	1–6
D5c	6	*Arabisopsis thaliana*	Unclassified	–
D6	10	Asteraceae	(+)-Germacrene A	1–10

synthases, respectively) since they contained many sequences within each group resulting in more robust inferences. Using the 5EAS structure[12] as a template for the location of residues specific to group D1, D2 and D6 shows that none of these residues would make contact with the farnesyl diphosphate substrate. Yet, several are located just outside the active site and could certainly contribute to tailoring the substrate-binding pocket (see Supplemental Data).

Comparative Modeling of Germacrene Synthases. Many of the hypotheses discussed in this report are based not just on the MSA and phylogenetic tree but on visual comparisons carried out in our stereo-visualization lab between the 5EAS structure and the three dimensional models we constructed through comparative modeling methods. We chose to focus on various germacrene synthases for two reasons. First, their mechanisms are relatively straightforward; a 1–10 initial cyclization reaction that concludes with a proton transfer from various sites on the germacadienyl intermediate to produce different germacrenes. For example, proton abstraction from C12 or C15 of the germacradienyl carbocation leads to either germacrene A or D respectively. And finally, examining different germacrene synthases is a reasonable first step towards understanding which substitutions are most responsible for new functions. Some of the questions we hoped to answer through comparative protein modeling of germacrene synthases were unsuccessful such as why the asterid (+)-germacrene A synthases do not reprotonate the product and instead release it, and therefore will require more advanced methods to answer.

By comparing the sequences and comparative protein models of (+)/(−)-germacrene D synthases (gDSs) versus the asterid (+)-gASs and 5EASs, a mechanism emerges on how plant sTSs switch from abstracting a proton from C12 and instead remove it from C15. In the Vitis vinifera (−)-gDS (grape, GenBank: AY561842) containing the Asp-444-Tyr-520-Asp-525 triad (5EAS numbering) residues near Tyr-520, including hydrophobic Ile-443 and Ala-447, have been substituted by hydrophobic Val/Met. In fact, this substitution is observed in most gDSs. These subtle substitutions involving hydrophobic residues are likely to function in a similar manner as those outside the active site of 5EAS/VDS

 © 2011 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

to steer intermediates towards different product outcomes.[21,22] As we concluded this study, a structure of a four-residue mutant of 5EAS that results in promiscuity showed a large change in the position of Tyr-520.[15] One substitution in particular, involving one hydrophobic residue for another Val516Ile, and positioned near Tyr-520 appears to be very important for shifting the location of the final proton transfer.[22]

In the (−)-gDSs from Pogostemon cablin (patchouli, Genbank: AY508727 and AY508729) that share 67% sequence identity, Asp-444 is replaced by Asn which again would result in changing Tyr-520's position as previously discussed. This substitution and other minor changes in the active site are enough to change the location of proton abstraction from the substrate.

Molecular Dynamics simulation of the eudesmane carbocation in active site of 5-epi-aristolocholene synthase. The 10-nanosecond simulations revealed a stable interaction between the phenolic oxygen atom of Tyr-520 and one of the C8 protons of the eudesmane carbocation. Trp-273 never presented its aromatic center to either of the C8 protons. This residue primarily interacts with the isopropyl group of eudesmane. The closest Trp-273 atom (labeled the CH2 atom in the CHARMM force field) to one of the C8 protons is shown in the time series graph and in a representative simulation snapshot of the active site (see Figure 5). These interactions were not present in the starting model but instead evolved rapidly during the equilibration steps.

4 Discussion

Properties of the Asn/Asp-Tyr-Asp triad. In the report on the 5EAS structure,[12] the Asp-444-Tyr-520-Asp-525 triad was proposed to remove the proton from C12 and effectively donate it to C6 forming the eudesmane carbocation (see Figure 6). Because of this and other studies, similar protonation mechanisms were proposed to take place in other plant sTSs.[26,54,55] The final proton removal from C6 or C8 forming vetispiradiene or 5-*epi*-aristolocholene respectively was postulated to be from Trp-273 based on its proximity to either proton in the 5EAS structure and because of its reduced nucleophilicity diminishing the likelihood of inadvertent enzyme alkylation by the highly reactive carbocation. Yet, the positive charge on the eudesmane carbocation is almost certainly delocalized reducing the likelihood of enzyme alkylation. In addition, our molecular dynamics simulations of the eudesmane carbocation in the active site of 5EAS reveal a definitive interaction between Tyr-520 and the C8 protons. The reorientation required for the substrate to form this interaction is minimal. Therefore, we postulate that the Asp-Tyr-Asp triad functions to abstract the C6 or C8 proton in vetispiradiene synthases (VDSs) and 5EASs. Whether all three residues function exactly as shown in Figure 6 is speculative. A study on the protonation of the neutral (S)-β-bisabolene intermediate in the maize (S)-β-macrocarpene synthase has Asn substituted in place of Asp-444 and yet it still performs three proton transfer events just like 5EAS.[56] The double mutant Tyr520/Phe//Asp525/Asn (5EAS numbering) blocked catalysis complete-

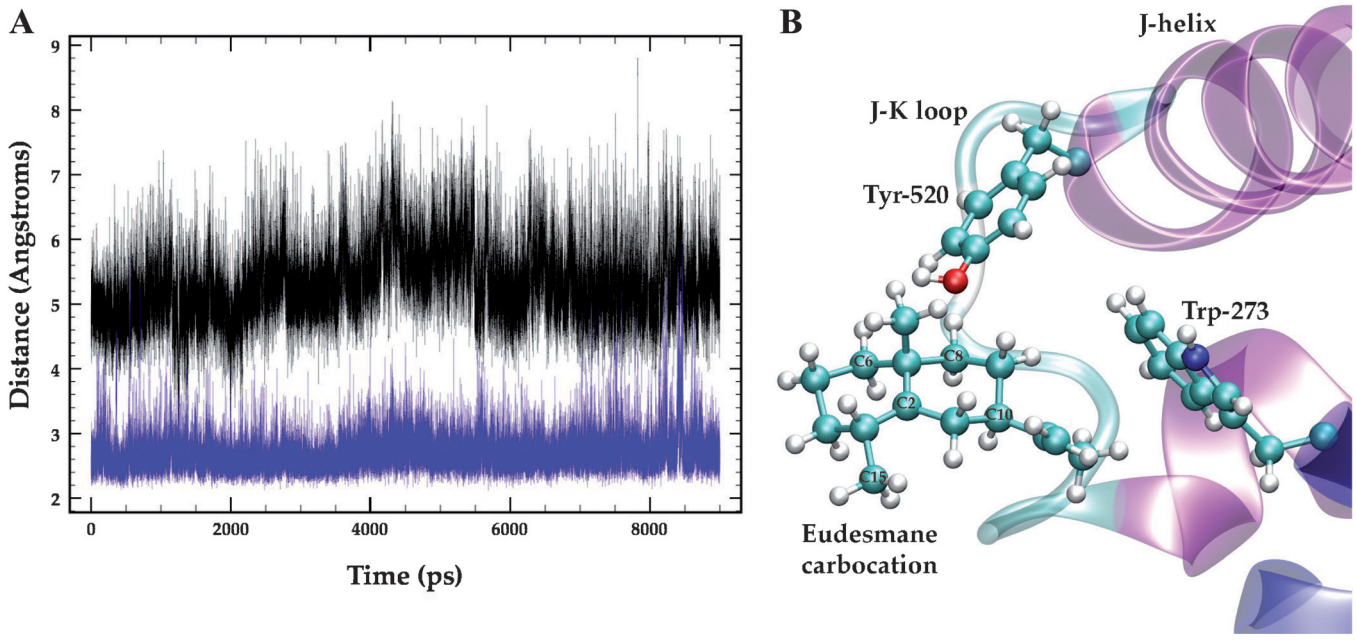

Figure 5. (a) Time series data over the last 9 ns of simulation showing the distance between the Trp-273 CH2 atom (black) and the closest proton on C8 of eudesmane. The distance between the phenolic oxygen atom of Tyr-520 and the same proton on C8 of eudesmane is shown in blue. (b) A representative snapshot from the MD simulation of eudesmane carbocation bound in the active site of 5-epi-aristolocholene synthase.

© 2011 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

ly.[56] Thus, the proton "flow" shown in Figure 6 might only be between Tyr-520 and Asp-525. This is not to argue that Asp-444, when present in a plant sTS sequence, plays no role in these proton transfers, it is just that its role is largely structural. Asp-444 interacts with one of the Mg^{2+} ions meaning the resonance structure shown in Figure 6d is likely an accurate depiction of its electronic distribution and therefore it is less likely to function as a proton acceptor. It is unknown how Asn-444 should be oriented in the active site since all known plant sTS *structures* contain an Asp at this position. If it orients its side chain carbonyl oxygen atom towards the phenolic oxygen atom of Tyr-520 it sacrifices a favorable interaction with a Mg^{2+} ion. If instead, its side chain amino group orients toward Tyr-520, then it likely alters the position of Tyr-520 and the substrate from what is seen in the 5EAS structure,[12] and thus changes where protons might be removed or donated to the substrate. This latter orientation seems more likely and would explain why Asp-444 does not seem to be absolutely necessary to remove and donate protons to the substrate.[56] Given the high conservation of the Asp444-Tyr-520-Asp-525 triad, its position relative to a folded substrate

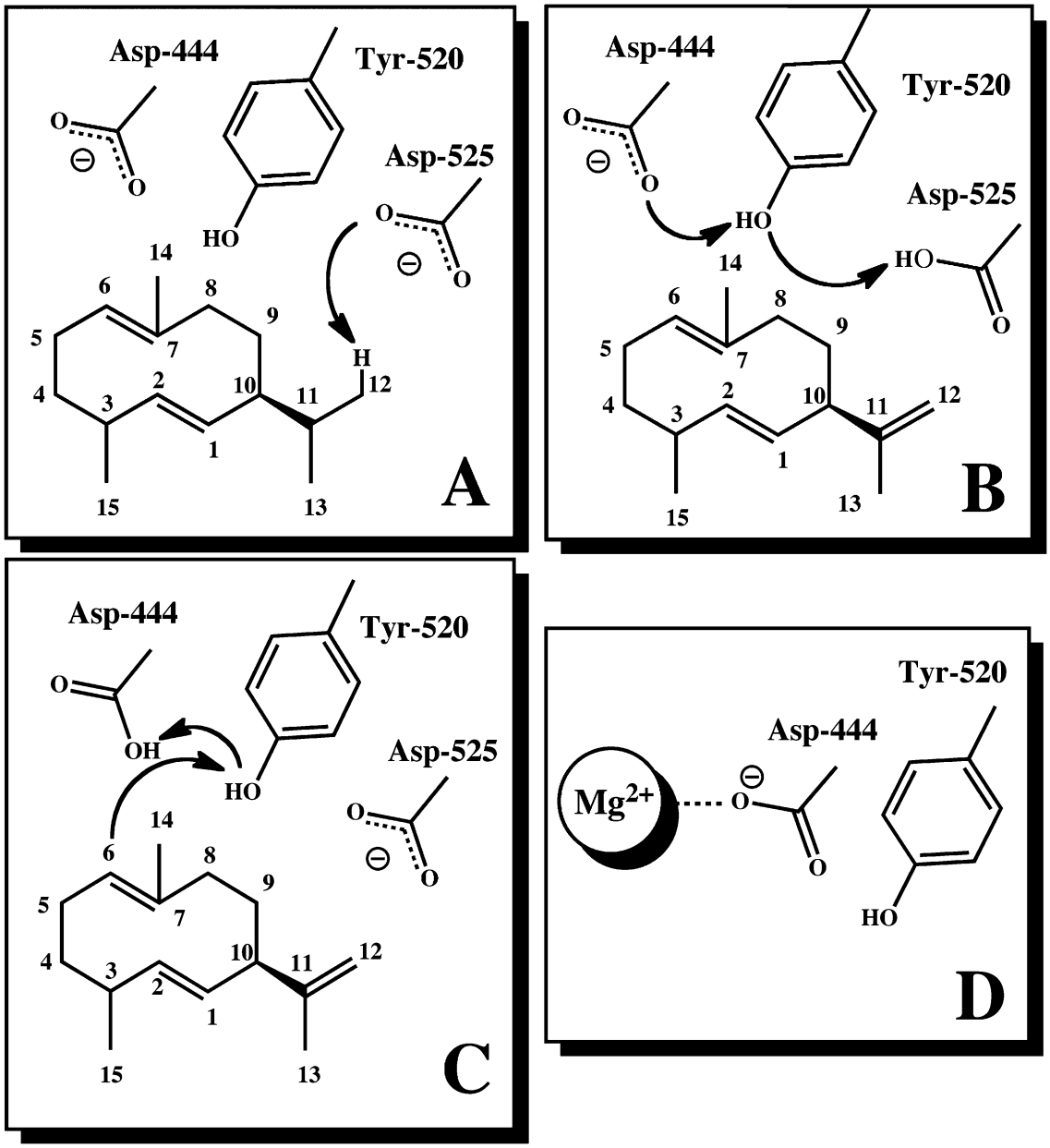

Figure 6. Proposed mechanisms for proton transfer reactions as part of the 5EAS catalytic cycle (detailed in Reference 10) (A) Asp-525 abstracts a proton from C12 to form the (+)-germacrene A intermediate (B) Asp-444 abstracts a proton from Tyr-520 as it abstracts the proton from Asp-525 (C) C6 of the intermediate abstracts a proton from Tyr-520 as it abstracts a proton from Asp-444 leading to the formation of the eudesmane carbocation (not shown). (D) the dominant resonance structure for Asp-444 given its interaction with Mg^{2+} ion.

 © 2011 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

analogue,[12] the demonstration of its importance to generating (+)-germacrene A either as an intermediate or product,[51] our own atomistic molecular dynamics simulations of the eudesmane carbocation in 5EAS, and finally the absence of likely proton donors/acceptors in other parts of many plant sTS active sites, *we propose that this triad is an important functional element* responsible for many proton transfers to and from the substrate and intermediates along the plant sesquiterpene synthase catalytic cycle. Though this triad is obviously not the key to understanding all of plant sTS enzymatic chemistry, we nevertheless propose that the triad can be tuned in a variety of ways to generate a diversity of products. These include 1) substituting residues surrounding the triad to shift their position relative to the substrate and/or intermediates and 2) substituting residues on the opposite side of the active site forcing the farnesyl diphosphate to fold in different ways so that through both mechanisms the triad will donate/abstract protons to and from different carbons.

5 Conclusions

Evolving a new enzyme function by mutation of active site residues can often have negative effects on protein stability[57] and thus compensating mutations are often necessary.[58] In addition, it has been shown that "nonfunctional" outside-the-active-site residue(s) must be mutated first before a "functional" active site residue can be mutated.[59] This context dependence, or epistasis, restricts the possible evolutionary pathways to novel functions.[60,61] Our phylogenetic analysis and molecular modeling results support the conclusion that plant sTS partially evolve new functions in a way that would have a minimal impact on protein stability by substituting residues outside the active site but still adjacent to the highly conserved Asp444-Tyr-520-Asp-525 triad which would shift the triads' position relative to the farnesyl substrate resulting in proton transfers from different locations on the substrate.

The construction of a functional landscape that interconverted 5EASs and VDSs clearly demonstrated that residues outside the active site can have an impact on the latter stages of the catalytic cycle while conserving the initial ones.[21,22] The different germacrene synthases we examined also conserve the initial 1–10 cyclization and appear to utilize residues not in contact with the substrate for shifting the location of the final proton transfer. The ability of plant sTSs to affect product distributions by substitutions to residues outside the active site also makes these enzymes an interesting target for the study of protein evolution and allosteric principles. Finally, as with most enzyme families this phylogeny reveals that closely related sequences share similar enzymatic mechanisms to various extents and can now be leveraged in future experiments to understand more distant relationships within the family and to guide homology-based functional annotation if it is deemed necessary (i.e., it should be avoided if the uncharacterized sequence falls into an unclassified or mechanistically diverse group). Our subprofile analysis identified several group-specific residues but none gave a clear indication on how they might impact the mechanistic network as these residues, unlike those identified in other protein families, were not located in the enzyme active site. Several of these residues were located on the periphery of the active site and likely modulate the function of the active site residues or possibly may act collectively to impact the product distribution.[22] Detailed structural studies will be crucial to understanding each residue's contribution to catalysis and how residues act collectively to steer the substrate or intermediates to various products. Some of the questions raised by this and other plant sTS studies are being investigated in our lab with computational methods.

Acknowledgement

This work was supported by a grant from the *National Institutes of Health* Grant RR06009.

References

[1] J. M. Smith, *Nature* **1970**, *225*, 563 – 564.
[2] J. T. Bridgham, S. M. Carroll, J. W. Thornton, *Science* **2006**, *312*, 97 – 101.
[3] A. Dean, J. W. Thornton, *Nat. Rev. Genet.* **2007**, *8*, 675 – 688.
[4] G. Conant, K. Wolfe, *Nat. Rev. Genet.* **2008**, *9*, 938 – 950.
[5] K. Dittmar, D. Liberles, *Evolution After Gene Duplication*, Wiley-Blackwell, New York **2010**.
[6] J. Gershenzon, N. Dudareva, *Nat. Chem. Biol.* **2007**, *3*, 408 – 414.
[7] J. Bohmann, et al. *Phytochemistry* **2002**, *60*, 109 – 116.
[8] C. L. Steele, J. Crock, J. Bohlmann, R. Croteau, *J. Biol. Chem.* **1998**, *273*, 2078 – 2089.
[9] D. E. Cane, "Isoprenoids, Including Cartenoids and Steroids" in: *Comprehensive Natural Products Chemistry* Vol. 2 (Eds: D. H. R. Barton, K. Nakanishi, O. Meth-Cohn), Elsevier, Oxford, **1999**.
[10] J.-M. Gao, W.-J. Wu, J.-W. Zhang, Y. Konishi, *Nat. Prod. Rep.* **2007**, *24*, 1153 – 1189.
[11] D. Christianson, *Chem. Rev.* **2006**, *106*, 3412 – 3442.
[12] C. M. Starks, K. Back, J. Chappell, J. P. Noel, *Science* **1997**, *277*, 1815 – 1820.
[13] T. G. Köllner, C. Schnee, J. Gershenzon, J. Degenhardt, *Plant Cell* **2004**, *16*, 1115 – 1131.
[14] H. M. Berman, et al. *Nuc Acids Res.* **2000**, *28*, 235 – 242.
[15] J. P. Noel, et al. *ACS Chem. Biol.* **2010**, *5*, 377 – 392.
[16] H. A. Gennadios, et al. *Biochemistry* **2009**, *48*, 6175 – 6183.
[17] Y. J. Hong, D. J. Tantillo, *J. Am. Chem. Soc.* **2009**, *131*, 7999 – 8015.
[18] M. D. Bojin, D. J. Tantillo, *J. Phys. Chem. A* **2006**, *110*, 4810 – 4816.
[19] D. Christianson, *Curr. Opin. Chem. Biol.* **2008**, *12*, 141 – 150.
[20] M. B. Austin, P. E. O'Maille, J. P. Noel, *Nat. Chem. Biol.* **2008**, *4*, 217 – 222.
[21] B. T. Greenhagen, P. E. O'Maille, J. P. Noel, J. Chappell, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 9826 – 9831.

© 2011 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

[22] P. E. O'Maille, et al. *Nat. Chem. Biol.* **2008**, *4*, 617 – 623.
[23] T. G. Köllner, et al. *Arch. Biochem. Biophys.* **2006**, *448*, 83 – 92.
[24] H. V. Thulasiram, H. K. Erickson, C. D. Poulter, *Science* **2007**, *316*, 73 – 76.
[25] J. Bohlmann, G. Meyer-Gauen, R. Croteau, *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 4126 – 4133.
[26] L. Sharon-Asa, et al. *Plant J.* **2003**, *36*, 664 – 674.
[27] S. Lee, J. Chappell, *Plant Physiol.* **2008**, *147*, 1017 – 1033.
[28] S. Aubourg, A. Lecharny, J. Bohlmann, *Mol. Genet. Genomics* **2002**, *267*, 730 – 745.
[29] S. C. Trapp, R. B. Croteau, *Genetics* **2001**, *158*, 811 – 832.
[30] J. Degenhardt, T. G. Köllner, J. Gershenzon, *Phytochemistry* **2009**, *70*, 1621 – 1637.
[31] D. Tholl, *Curr. Opin. Plant Biol.* **2006**, *9*, 297 – 304.
[32] D. E. Cane, *Chem. Rev.* **1990**, *90*, 1089 – 1103.
[33] D. J. Tantillo, *Chem. Soc. Rev.* **2010**, *39*, 2847 – 2854.
[34] Y. Yoshikuni, T. E. Ferrin, J. D. Keasling, *Nature* **2006**, *440*, 1078 – 1082.
[35] UniProt Consortium, *Nucleic Acids Res.* **2010**, *38*, D142 – D148.
[36] W. Li, A. Godzik, *Bioinformatics* **2006**, *22*, 1658 – 1659.
[37] M. A. Larkin, et al. *Bioinformatics* **2007**, *23*, 2947 – 2948.
[38] T. L. Bailey, C. Elkan, *Proc. Sec. Int.Conf. Intell. Syst. Mol. Biol.* **1994**, *2*, 28 – 36.
[39] K. B. Nicholas, H. B. Nicholas Jr., D. W. Deerfield II, *EMBNEW. NEWS* **1997**, *4*, 14.
[40] K. Sjölander, *Bioinformatics* **2004**, *20*, 170 – 179.
[41] K. Sjölander, *PLoS Comp. Biol.* **2010**, *6*, 1 – 3.
[42] J. Felsenstein, *Cladistics* **1989**, *5*, 164 – 166.
[43] S. Hannenhalli, R. B. Russell, *J. Mol. Biol.* **2000**, *303*, 61 – 76.
[44] J. Hempel, J. Perozich, T. Wymore, H. B. Nicholas Jr., *Chem. Biol. Int.* **2003**, *143–144*, 23 – 28.
[45] A. Fiser, A. Sali, *Methods Enzymol.* **2003**, *374*, 461 – 491.
[46] M. Y. Shen, A. Sali, *Protein Sci.* **2006**, *15*, 2507 – 2524.
[47] G. G. Krivov, M. M. Shapovalov, R. L. unbrack Jr., *Proteins* **2009**, *77*, 778 – 795
[48] W. Humphrey, A. Dalke, K. Schulten, *J. Mol. Graphics* **1996**, *14*, 33 – 38.
[49] B. R. Brooks BR, et al., *J. Comp. Chem.* **2009**, *30*, 1545 – 1614.
[50] A. D. MacKerell Jr., et al., *J. Phys. Chem. B* **1998**, *102*, 3586 – 3616.
[51] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926 – 935.
[52] A. D. MacKerell Jr., M. Feig, C. L. Brooks III, *J. Comp. Chem.* **2004**, *25*, 1400 – 1415
[53] K. A. Rising, C. M. Starks, J. P. Noel, J. Chappell, *J. Am. Chem. Soc.* **2000**, *122*, 1861 – 1866.
[54] K. Back, J. Chappell, *J. Biol. Chem.* **1995**, *270*, 7375 – 7381.
[55] Y. Iijima, et al. *Plant Physiol.* **2004**, *136*, 3724 – 3736.
[56] T. G. Köllner, et al. *J. Biol. Chem.* **2008**, *283*, 20779 – 20788.
[57] M. Depristo, D. Weinreich, D. Hartl, *Nat. Rev. Genet.* **2005**, *6*, 678 – 687.
[58] N. Tokuriki, F. Stricher, L. Serrano, D. Tawfik, *PLoS Comp. Biol.* **2008**, *4*, e1000002.
[59] P. Romero, F. H. Arnold, *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 866 – 876.
[60] F. Poelwijk, D. Kiviet, D. Weinreich, S. Tans, *Nature* **2007**, *445*, 383 – 386.
[61] D. M. Weinreich, N. F. Delaney, M. A. Depristo, D. L. Hartl, *Science* **2006**, *312*, 111 – 114.

Received: May 20, 2011
Accepted: September 11, 2011
Published online: October 10, 2011

 © 2011 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim