An analysis of characterized plant sesquiterpene synthases

Durairaj, J.; Di Girolamo, A.; Bouwmeester, H.J.; de Ridder, D.; Beekwilder, J.; van Dijk, A.D.J.

DOI
10.1016/j.phytochem.2018.10.020

Publication date
2019

Document Version
Final published version

Published in
Phytochemistry

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
An analysis of characterized plant sesquiterpene synthases

Janani Durairaja, Alice Di Girolamo\textsuperscript{b}, Harro J. Bouwmeester\textsuperscript{c}, Dick de Ridder\textsuperscript{a}, Jules Beekwilder\textsuperscript{b,d}, Aalt DJ. van Dijk\textsuperscript{a,e,}\textsuperscript{*}

\textsuperscript{a} Bioinformatics Group, Department of Plant Sciences, Wageningen University, Netherlands
\textsuperscript{b} Laboratory of Plant Physiology, Department of Plant Sciences, Wageningen University, Netherlands
\textsuperscript{c} Swammerdam Institute for Life Sciences, University of Amsterdam, Netherlands
\textsuperscript{d} Bioscience, Wageningen Plant Research, Wageningen University, Netherlands
\textsuperscript{e} Biometris, Department of Plant Sciences, Wageningen University, Netherlands

ARTICLE INFO

Keywords:
Database
Product specificity
Enzyme
Sesquiterpene
Sesquiterpene synthase
Terpene synthase

ABSTRACT

Plants exhibit a vast array of sesquiterpenes, C\textsubscript{15} hydrocarbons which often function as herbivore-repellents or pollinator-attractants. These in turn are produced by a diverse range of sesquiterpene synthases. A comprehensive analysis of these enzymes in terms of product specificity has been hampered by the lack of a centralized resource of sufficient functionally annotated sequence data. To address this, we have gathered 262 plant sesquiterpene synthase sequences with experimentally characterized products. The annotated enzyme sequences allowed for an analysis of terpene synthase motifs, leading to the extension of one motif and recognition of a variant of another. In addition, putative terpene synthase sequences were obtained from various resources and compared with the annotated sesquiterpene synthases. This analysis indicated regions of terpene synthase sequence space which so far are unexplored experimentally. Finally, we present a case describing mutational studies on residues altering product specificity, for which we analyzed conservation in our database. This demonstrates an application of our database in choosing likely-functional residues for mutagenesis studies aimed at understanding or changing sesquiterpene synthase product specificity.

1. Introduction

The terpenome represents a huge, ancient and diverse family of natural products. In addition to terpenes, it also encompasses steroids and carotenoids, comprising more than 60,000 members (Buckingham (1997)). These compounds all derive from the same 5-carbon precursor units, coupled together linearly and then cyclized, rearranged, and modified in various ways. Terpenes serve many roles in plants, for example as toxins against herbivores or pathogens, or as attractants for pollinators (Gershenzon and Dudareva (2007)). In turn, terpenes extracted from plants are used by mankind for a range of applications as pharmaceutical agents, insecticides, preservatives, fragrances, and flavors (Scheppe et al. (2017)).

Terpenes are built from 5-carbon isoprenoid units and they mainly exist as monoterpenes (C\textsubscript{10}), sesquiterpenes (C\textsubscript{15}) or diterpenes (C\textsubscript{20}), based on the number of such units used. In each case, a linear substrate loses a diphosphate group, usually cyclizes and then undergoes a variety of carbocation rearrangements. Though the exact number of sesquiterpenes found in nature is hard to determine, Tian et al. (2016) estimated computationally that the number of sesquiterpene intermediates far outnumber those of monoterpenes, due to the increase in chain length.

Interestingly, sesquiterpenes found in nature can be divided into seven groups based on their parent cation and the first cyclization step in their formation (Degenhardt et al. (2009)). Hence the extreme diversity of chemical compounds with desirable fragrances or medicinal properties is based on just seven initial carbocations. This makes the enzymes catalyzing their formation both interesting and difficult to characterize functionally.

Each plant species is capable of synthesizing a number of sesquiterpenes using a specialized class of enzymes called sesquiterpene synthases (STs). First, a farnesyl diphosphate synthase, produces the C\textsubscript{15} substrate for STs, farnesyl diphosphate (FPP), from the C\textsubscript{5}-unit isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP) (Ogura et al. (1997)). STs then create the myriad of sesquiterpenes found in nature by catalyzing carbocation formation from the linear FPP followed by a series of cyclizations and rearrangements (Fig. 1). Products are formed from intermediate carbocations after
deprotonation, phosphorylation, or hydration (Tian et al. (2016)). The STSs themselves represent a very diverse set of enzymes with a wide range of sequence similarities, despite having a common structural fold shared by plant, animal, fungal, and bacterial terpene synthases (TPSs) (Gao et al. (2012)). Hence, prediction of enzyme function from sequence is highly challenging in the case of STSs. Moreover, sequence diversity in STSs is not dependent on the products formed. This problem has been addressed so far by inspection of TPS structures (Gao et al. (2012)) and by mutational analyses that attempt to change the product of a synthase with the smallest number of residue changes (Li et al. (2013)). The former, though an attractive approach, is limited especially in plants due to the sparsity of experimentally determined structures, while the latter often leads to unnatural enzymes with lower catalytic activity than their wild-type parents. Characterization of multiple TPSs from the same species by the same study has allowed for some small-scale sequence comparison of those synthases (Martin et al. (2010); Martin et al. (2004)). However, no previous attempts have been made to compare all experimentally characterized plant STS sequences according to the products that they form. We have collated a curated database of plant STSs with experimentally characterized product data, shown in Table 1. The enzymes originate from a hundred different plant species and collectively account for the production of 117 different sesquiterpenes. Such a large number of possible products makes it difficult to find enough enzymes with the same product for a meaningful analysis of product specificity. To solve this, the sequences were divided into seven groups, making use of the sesquiterpene precursor carbocation scheme as specified by Degenhardt et al. (2009), described in Fig. 1. The reaction cascade of an STS is initiated by metal-mediated removal of the diphosphate anion in the FPP substrate, leading to the formation of a transoid (2Z,6E)-farnesyl cation (farnesyl cation) which can undergo cyclization either via 10-exo-trig or 11-endo-trig cyclizations on the C10–C11 double bond to form cyclic cations 1 or 2 respectively. However, the farnesyl cation can also isomerize to form a cisoid (2Z,6E)-farnesyl cation (nerolidyl cation). The nerolidyl cation, in finding residues involved in STS product specificity.

2. Results and discussion

2.1. Database of characterized STSs

To obtain a comprehensive set of annotated STSs, our starting point was the SwissProt database, a subset of UniProt (Boeckmann et al. (2003)) in which a curated and annotated set of proteins is available. This provided a set of 104 STSs. In addition, we manually reviewed literature linked to enzymes with the characteristic TPS domain in TremBL, the uncurated subset of UniProt. In this way, the number of curated plant STS sequences with experimentally characterized product data in the database was more than doubled.

We present a database of 262 manually curated characterized plant STSs, shown in Table 1. The enzymes originate from a hundred different plant species and collectively account for the production of 117 different sesquiterpenes. Such a large number of possible products makes it difficult to find enough enzymes with the same product for a meaningful analysis of product specificity. To solve this, the sequences were divided into seven groups, making use of the sesquiterpene precursor carbocation scheme as specified by Degenhardt et al. (2009), described in Fig. 1. The reaction cascade of an STS is initiated by metal-mediated removal of the diphosphate anion in the FPP substrate, leading to the formation of a transoid (2Z,6E)-farnesyl cation (farnesyl cation) which can undergo cyclization either via 10-exo-trig or 11-endo-trig cyclizations on the C10–C11 double bond to form the resulting cations 1 or 2 respectively. However, the farnesyl cation can also isomerize to form a cisoid (2Z,6E)-farnesyl cation (nerolidyl cation). The nerolidyl cation, in finding residues involved in STS product specificity.
in addition to a C1-attack (either via 10-exo-trig or 11-endo-trig) on the C10–C11 double bond to form cations 3 or 4, can also undergo cyclization at its C6–C7 double bond either via 6-exo-trig or 7-endo-trig, forming cations 5 or 6. These carbocations undergo multiple further skeletal rearrangements, cyclizations, hydride or methyl shifts, and other modifications to form the end products of the enzyme (Degenhardt et al. (2009)). Along with this myriad of cyclic products, acyclic sesquiterpenes can also be formed from either the farnesyl or the nerolidyl cation through proton loss or addition of water (Bairoch (2000); Christianson (2017); Degenhardt et al. (2009)). This schematic of carbocations derived from FPP can be used to divide sesquiterpenes produced by plants into seven groups both based on their parent cation (farnesyl or nerolidyl) and the first cyclization that occurs (by attack of the carbocation on the 10,1; 11,1; 6,1; or 7,1-double bond; or acyclic). For an STS enzyme, the carbocation of its major product is then used to determine its group in Table 1.

This division of STTs is in general straightforward even when multiple products are formed by one enzyme. Specifically, of the 98 sequences which also have minor products (Supplementary Table T1), only 17 have minor products whose precursor carbocation differs from the major product's. Nine of these produce acyclic products in addition to their major product. This could be the result of incomplete cyclization caused by premature termination of intermediates (Köllner et al. (2009)). Eight enzymes in the database either produce (−)-germacrene D or they produce germacrene D and the chirality was not determined during the enzyme’s characterization. (−)-germacrene D can be formed via either the farnesyl or the nerolidyl cation through proton loss or addition of water (Bairoch (2000); Christianson (2017); Degenhardt et al. (2009)). Along with this myriad of cyclic products, acyclic sesquiterpenes can also be formed from either the farnesyl or the nerolidyl cation through proton loss or addition of water (Bairoch (2000); Christianson (2017); Degenhardt et al. (2009)). This schematic of carbocations derived from FPP can be used to divide sesquiterpenes produced by plants into seven groups both based on their parent cation (farnesyl or nerolidyl) and the first cyclization that occurs (by attack of the carbocation on the 10,1; 11,1; 6,1; or 7,1-double bond; or acyclic). For an STS enzyme, the carbocation of its major product is then used to determine its group in Table 1.

This division of STTs is in general straightforward even when multiple products are formed by one enzyme. Specifically, of the 98 sequences which also have minor products (Supplementary Table T1), only 17 have minor products whose precursor carbocation differs from the major product’s. Nine of these produce acyclic products in addition to their major product. This could be the result of incomplete cyclization caused by premature termination of intermediates (Köllner et al. (2009)). Eight enzymes in the database either produce (−)-germacrene D or they produce germacrene D and the chirality was not determined during the enzyme’s characterization. (−)-germacrene D can be formed via either the farnesyl or a 10,1, or a 11,1-cyclization of the farnesyl cation (cation 1 or 2). Though each enzyme is likely to only follow one cyclization route to form its product, this route has so far not been determined, so these sequences are shown separately in and in the remainder of the text. The existence of other sesquiterpenes which can be formed via different cyclization routes cannot be ruled out, however in our analysis we stick to the cyclization routes involved in their production. To this end, chemical similarities between sesquiterpenes with the same parent cation can be seen that many pairs of sequences have less than 40% sequence identity. Similarly, Supplementary Fig. S2 shows the hierarchical clustering of concatenated N-terminal and C-terminal sub-sequences for spermatophyte enzymes. Both clusterings appear very comparable.

The phylogenetic tree of C-terminal sub-sequences of all 262 enzymes (Fig. 2) shows some grouping of spermatophyte enzymes based on their product precursor. In general, the neighbor of an enzyme is from the same or related species, and if there are enough examples from the same species then some product-based grouping is seen. For example, the clades containing mostly enzymes from Zea mays on the right are separated based on the product carbocation of the enzyme even while being grouped by the species. However, this is not a consistent trend enzymes from Vitis and Santalum at the top of the tree mainly by species and not by product type. In fact, the three Santalum synthase sequences marked in Fig. 2, making products derived from three different cyclic carbocations, have more than 90% in common. In any case, the product group of an enzyme from a species not present in the tree is nearly impossible to predict, while enzymes from species which are less represented in the tree can also be difficult to classify. In addition, clades forming predominantly one product carbocation are seen in many different parts of the tree, showing that strongly varying sequences can catalyze the same cyclization reaction and even produce the same product, such as the two marked β-carophyllene synthases from Arabidopsis lyrata and Zea perennis which have a sequence identity less than 30%. Hence phylogenetic analysis is biased and cannot be an accurate predictor of TPS product specificity. Supplementary Fig. S3, shows a similar tree considering both N-terminal and C-terminal sub-sequences concatenated together, for spermatophyte STS sequences only. N-terminal domain information again does not seem to effect the structure of the tree. Even though this does not rule out the possibility that residues in the N-terminal domain influence product specificity, it indicates that including the N-terminal domain as a large scale sequence analysis that we perform does not add information compared to using only the C-terminal domain. Since product and intermediate formation occur in the active site pocket, it may be easier to find sequence-function determinants in the C-terminal domain. Hence, from this point on we concentrate on the C-terminal sub-sequences of TPSs.

### 2.2. Chemical similarities between sesquiterpenes

Each of the seven possible sesquiterpene precursors (Fig. 1) usually undergoes a wide range of further rearrangements, cyclizations, and modifications, catalyzed by the STS enzyme, to finally result in a sesquiterpene product. To start exploring the enzyme grouping scheme, we initially investigated whether similarities between the final sesquiterpene chemical structures would reflect the parent carbocations involved in their production. To this end, chemical similarities between sesquiterpenes with the same parent cation were compared to similarities between those without. Chemical similarities were measured
using Dice similarity (Willett et al. (1998)) between extended connectivity fingerprints, as described by Rogers and Hahn (2010). Similarities between 165 sesquiterpenes are plotted using multi-dimensional scaling (MDS), in Fig. 3a with the color representative of the precursor cation. These 165 compounds collectively represent every enantiomer of the 117 sesquiterpenes produced by the enzymes in our database, since many of the experimental characterization studies used to build the database did not resolve the chirality of the STS’s product. MDS is a technique used to visualize the level of similarity of individual objects in a dataset using a distance matrix, such that the between-object distances are preserved as well as possible. Therefore, two objects appearing close to each other in the MDS plot represent sesquiterpenes which likely have a high chemical similarity, while those further away have lower similarity. Acyclic sesquiterpenes are clearly distinguishable in the plot, as they are linear in nature. Interestingly, many products derived from the 6,1-cyclized cation (cation 5) are also distinct from those derived from 10,1 or 11,1-cyclized cations despite further cyclizations and rearrangements after this first step. They cluster midway between the acyclic and other cyclic products, which makes sense given the presence of an acyclic tail portion in cation 5. The sesquiterpenes formed from the other cyclic cations seem less distinguishable.

2.3. Characterized sequence space

Though a manual literature search gave us access to more functionally characterized TPS sequences, there is a large and steadily
From our knowledge, uncharacterized. Fig. 3 has a few commonalities with the MDS plot of chemical similarities between sesquiterpenes, Fig. 3a. Many sequences catalyzing acyclic products as well as those derived from cation 5 cluster separately from the others. In fact, the enzymes making nerolidol, an acyclic sesquiterpene, cluster separately at the bottom right of the plot (light blue), leading us to hypothesize that perhaps many of the other uncharacterized STSs in this area also catalyze the formation of nerolidol. A second similarity is that enzymes forming products derived from 10,1 to 11,1-cyclized cations are difficult to distinguish. This again confirms, as was seen in the phylogenetic tree (Fig. 2), that overall sequence similarity by itself cannot be an accurate guide to product specificity.

Product specificity is even harder to identify in the case of gymnosperm synthases, as insufficient data is available to separate enzymes with different product cations. It has been noted before that gymnosperm TPSs resemble each other more than they do their angiosperm counterparts, regardless of catalytic activity (Chen et al. (2011); Trapp and Croteau (2001)). The enzymes from these species may be more informative if analyzed separately but this would require more gymnosperm sequences to be functionally annotated.

2.4. Comparing known TPS motifs across sequences

A database such as ours allows for a comparison of residues in previously studied structural elements across many STS sequences. A thorough study of TPS structures has led to the identification of several motifs important for catalytic activity (Gao et al. (2012)). In the case of STSs, the hydrophobic moiety of the STS substrate, FPP, is directed into the active site cavity, to undergo the cyclizations and rearrangements described in Fig. 1. Studies on STS structures have proposed that the diphosphate moiety is captured by the motif RxR and divalent metal ions like Mg$^{+2}$ or Mn$^{+2}$ which are themselves bound by motifs DDxxD and NSE/DTE, at the entrance of the active site (Starks et al. (1997)). Here, we compare these three motifs across the sequences in our database. Fig. 4a shows the motifs discussed below on a tobacco aristolochene synthase structure (Starks et al. (1997)). Fig. 4b shows each motif on a schematic representation of the alignment of all C-terminal sub-sequences in the database.

2.4.1. Aspartate-rich DDxxD motif conserved in plant STSs

The most conserved motif of TPSs is the metal binding aspartate-rich motif found both in plant and microbial TPSs as well as in isoprenyl diphosphate synthases.

Numerous studies performed on this motif, both site-directed mutagenesis and X-ray crystallography analysis, show that it is involved in binding the divalent metal ions in the active site entrance (Aaron and Christianson (2010)). The canonical form of the motif, DDxxD(E,D), where boldfaced residues indicate those proposed to bind Mg$^{+2}$ or Mn$^{+2}$, is found in 247 of the 249 spermatophyte enzymes. Of the remaining two, one is a (+)-germacrene-D synthase from Solidago canadensis with an Asn replacing the first Asp (Prosser et al. (2004)). The other is a bicyclogermacrene synthase from Matricaria chamomilla with an Asn replacing the second Asp (Son et al. (2014)). These examples indicate that either one of the first two Aspartates may be sufficient for maintaining catalytic activity.

2.4.2. Expanded NSE/DTE motif found in most sequences

The opposite site of the active site entry is also involved in metal-binding, due to the presence of a second, less-defined motif, termed the NSE/DTE motif (Christianson (2006)). An early form of this motif, as detailed by Christianson (2006) had a consensus of (L,V)(V,L,A)x(+)germacrene-D synthase. The residues in bold coordinate Mg$^{+2}$ ions. However, searching for a motif with this consensus only captured 38 of the 249 spermatophyte sequences in our database, indicating that it may be too restrictive given the current knowledge of sequences. When only the metal-binding portion of the motif is considered, the consensus sequence (N,D)Dxx(S,T)xxxE covers 219 spermatophyte sequences in the database. The possibility of Gly in the second metal-binding position is justified by Zhou and Peters (2009), with the proposal that Gly
may allow a water molecule to substitute for the hydroxyl group of Ser/Thr. Some TPSs however, are known to have a second, catalytically active, aspartate rich motif instead of the NSE/DTE motif (Gennadios et al. (2009); Little and Croteau (2002); Steele et al. (1998)) with the same consensus as the first, \( \text{D}^{Dxx}(\text{D,E}) \). This occurs in 20 sequences. Table 2 shows the distribution of the sequences over the different versions of the second motif.

A highly conserved Arg is found 3 residues upstream of all versions of the NSE/DTE motif or second aspartate-rich motif, in all of the spermatophyte sequences in the database. All 6278 uncharacterized spermatophyte TPS sequences also have an arginine in this position. Hence, an extended form of the motif may be more relevant for spermatophyte STSs, with the consensus \( \text{R}^{xx}(\text{N,D})\text{Dxx}(\text{S,T,G})\text{xxx}E \) or \( \text{RxxDDxx(D,E)} \).
A. annua source of many medicinal terpenes. Some of the STSs from Artepisia annua best characterized TPSs derive from 2.5. Comparing residues involved in product specificity across sequences acyclic sesquiterpene. This indicates that RxQ may be unable to capture of the remaining have an altered RxQ motif in the same region. Inter-
spermatophyte plant sequences have the canonical RxR motif while 18 
carbocationic intermediates (Starks et al. (1997)). 215 of the 249 
the substrate, thereby preventing nucleophilic attack on any of the 
carboxiontic intermediates (Starks et al. (1997)). The two Arg residues in the motif were proposed 
to be involved in the complexation of diphosphate after ionization of 
loop. This loop has been shown to become ordered upon FPP binding 
DDxxD motif, located on a flexible loop in the structure, termed the A-C 
2.4.3. RxR motif not conserved in nerolidol synthases 

The RxR motif is found about 35 amino acids upstream of the 
DDxxD motif, located on a flexible loop in the structure, termed the A-C loop. This loop has been shown to become ordered upon FPP binding 
(Sarks et al. (1997)). The two Arg residues in the motif were proposed to 
be involved in the complexation of diphosphate after ionization of the 
substrate, thereby preventing nucleophilic attack on any of the 
carboxiontic intermediates (Sarks et al. (1997)). 215 of the 249 
spermatophyte plant sequences have the canonical RxR motif while 18 
of the remaining have an altered RxQ motif in the same region. Interest-
ingly, these 18 enzymes all catalyze the formation of nerolidol, an 
acyclic sesquiterpene. This indicates that RxQ may be unable to capture diphosphate to the same extent as RxR, causing a premature quenching of an intermediate carbocation by water before cyclization has occurred (Degenhardt et al. (2009)).

2.5. Comparing residues involved in product specificity across sequences

Many studies have addressed the importance of specific residues 
located in the active site of TPSs via mutational analyses. Some of the 
best characterized TPSs derive from Artemisia annua, which is the 
source of many medicinal terpenes. Some of the STSs from A. annua have served as examples to identify residues involved in critical steps in the 
cyclization cascade. In this section three examples of A. annua STSs are described, for which residues involved in product specificity were 
experimentally investigated. We use these as a case-study to illustrate how the large set of characterized STSs that we make available can 
potentially be used to guide such experimental investigations. These examples are:

1. Salmon et al. (2015) tested a wide library of mutants for the (E)-β-
farnesene synthase (UniProt: Q9FXY7) from A. annua, an STS catal-
yzing the formation of an acyclic product. They discovered that a 
single substitution, Tyr402Leu, confers to the synthase a cyclase activity, resulting in zingerbene and β-bisabolene as the most 
abundant products. Both these sesquiterpenes derive from cation 5. In 
sequences catalyzing the formation of 10,1 and 11,1 cyclized 
products (cations 1, 2, 3 and 4), this position is highly conserved (> 95%) in cyclic STSs. Importantly, in 
acyclic STSs the most common amino acid is a Tyr, with a con-
servation of 38%. Acyclic STSs even have amino acids such as Gln, 
Gly and Ile in this position, never seen in any of the cyclic STSs in the 
database. The variability and low conservation score indicates that 
changing this position in cyclic STSs away from a Ser or Thr could 
result in the formation of acyclic products, as shown by Li et al. (2016).

In summary, analysis of these A. annua examples of residues in-
volved in the first cyclization step in STSs indicates that conservation patterns across all the annotated enzymes are consistent with the 
functional roles of these residues. This suggests it would be possible to 
obtain residues potentially involved in product specificity from this 
database. Such a data-driven approach is in contrast to how these 
mutational studies have traditionally been guided, i.e. by comparison of 
two or three sequences from the same or related species. Therefore, a 
potential application of our database is in guiding site-directed muta-
genesis studies in a way which avoids species bias and hence may reveal 
additional residues involved in product specificity. One such residue 
position obtained by studying conservation patterns has been discussed 
above in Section 2.4.3, namely the second arginine in the RxR motif. 
This position was found to be glutamine in most nerolidol synthases, 
something not seen in any of the cyclic synthases. Mutating this residue 
in cyclic synthases and monitoring for acyclic products, and vice versa, 
could confirm the residue’s role in the cyclization of sesquiterpene 
products.

3. Conclusions

We compiled a manually curated set of experimentally character-
ized plant STSs along with their major products. This database is the 
largest centralized resource of annotated plant STSs to date and allows 
for thorough sequence-based analysis of these diverse enzymes. The 
enzymes in the database are grouped according to the carbocational origin and cyclization of their major product. Such a division alleviates 
the task of functional analysis and comparison between the enzymes. 
Using the database we were able to extend and find variants of existing 
STS motifs. In addition, residues from previous mutational studies, 
when mapped onto the enzymes in the database, were found to have 
detectable conservation patterns that differed from group to group. 
Such properties of residues can be extrapolated and used to guide fur-
ther mutational studies. The database as a whole helps to understand 
the current state of STS sequence space characterization, and provides a 
starting point for future efforts to predict product specificity.

4. Experimental

4.1. Literature search for characterized STSs

To find potentially characterized STSs, an HMM search was performed using hmmer (version 3.1b2) (Eddy (1998)) on the UniProt

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Division of the different versions of the second metal-binding motif among characterized spermatophyte STS sequences. Sequences with motifs not covered by either motif consensus sequence (N,D)Dxx (S,T,G)xxxE or DDxx (D,E) are classified as “Other”.</td>
</tr>
<tr>
<td>Motif</td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td>DDxxTxxxxE</td>
</tr>
<tr>
<td>DDxxSxxxxE</td>
</tr>
<tr>
<td>NDxxSxxxxE</td>
</tr>
<tr>
<td>DDxxGxxxxE</td>
</tr>
<tr>
<td>NDxxTxxxxE</td>
</tr>
<tr>
<td>DDxxGxxxxE</td>
</tr>
<tr>
<td>NDxxxxxxE</td>
</tr>
<tr>
<td>DDxx(D,E)</td>
</tr>
<tr>
<td>Other</td>
</tr>
</tbody>
</table>

163
database (UniProt (2016)) using the HMM of the C-terminal domain of TPSs from Pfam (Bateman et al. (2004)) (Pfam ID: PF03936). Protein sequences with a hit having an E-value < 10^{-10} and a total protein length between 350 and 650 residues were selected. The UniProt IDs of these sequences were then linked to Pubmed IDs, either directly through programmatic access of UniProt if the Pubmed ID was present, or through a programmatic text search of the title and authors given in UniProt, using the Pubmed API (Wheeler et al. (2006)). The Pubmed articles thus obtained were searched manually for evidence of experimental characterization of sesquiterpenes through in-vivo or in-vitro GC-MS studies, and the corresponding UniProt IDs were collected.

For each UniProt ID found, the major product described in the corresponding paper was stored. Minor products with GC-MS peaks at least quarter the height of the major product peak were stored as well.

### 4.1.1. Measuring chemical similarities

The diagram of the sesquiterpene grouping scheme was made using ChemDoodle (version 9) (Todsen (2014)). The InChI strings for 165 sesquiterpenes were obtained from PubChem (Bolton et al. (2008)) using the Python wrapper for the PubChem REST API (Kim et al. (2015)), PubChemPy (version 1.0.4). To measure the similarity between different sesquiterpenes, rdkit (Release 2017.09.3) was used (Landrum et al. (2006)). A circular chemical fingerprint, called the Morgan fingerprint, with a radius of 2 Å, as explained by Rogers and Hahn (2010), was obtained for each sesquiterpene. The similarity between every pair of fingerprints was then calculated using Dice similarity (Willett et al. (1998)). The distance was given as 1 similarity. The distance matrix of all sesquiterpenes was then used to create a multi-dimensional scaling (MDS) plot using the Python scikit-learn library (version 0.19.1) (Pedregosa et al. (2011)), and then plotted using matplotlib (version 2.1.2) (Hunter (2007)).

### 4.1.2. Aligning sequences

For characterized spermatophyte plant STS sequences, the C-terminal catalytically active portion and the N-terminal portion of the enzyme were found with hmmer HMM searches (version 3.1.b2) (Eddy (1998)) using the TPS C-terminal Pfam domain (Pfam ID: PF003936) and the TPS N-terminal Pfam domain (Pfam ID: PF01397) respectively. These were then separately aligned using Clustal Omega (version 1.2.4) (Sievers et al. (2011)), with all heuristic features off and the respective Pfam domains as a guide for alignment. From these separate alignments, a concatenated N+C alignment was formed, covering both domains.

For some of the nonseed plant STS sequences however, a C-terminal Pfam domain search returned < 200 residues instead of the usual 250-270. Aligning the full nonseed sequences using the spermatophyte C-terminal sub-sequence alignment as a profile showed the position of the C-terminal portion for these sequences, so this was used to extract the required Cterminal sub-sequences for nonseed plants. An alignment consisting of both seed and nonseed characterized C-terminal sub-sequences was constructed using Clustal Omega with the same parameters as above.

### 4.1.3. Phylogenetic tree construction

A phylogenetic tree was built and visualized for the characterized spermatophyte and nonseed plant enzymes in the database using the ete toolkit (version 3.1.1) (Huerta-Cepas et al. (2016)). The previously explained alignment of all C-terminal sub-sequences was used, with columns having > 50% gaps removed using trimAL (Capella-Gutierrez et al. (2009)). The best protein model from JTT, WAG, VT, LG and MtREV was chosen using ProtTest (Abascal et al. (2009)), and finally a RaxML maximum likelihood tree was built (Stamatakis (2014)). Similarly, a phylogenetic tree for the spermatophyte sequences was built with the same approach using the concatenated N+C alignment.

### 4.1.4. Finding mono-, di-, and uncharacterized TPSs

Characterized plant monoand diterpene synthases were obtained from SwissProt (Boeckmann et al. (2003)) using a C-terminal TPS Pfam domain hmmr (version 3.1.b2) (Eddy (1998)) HMM search followed by collecting the sequences from plant species for which the catalytic activity was mentioned. These were not manually checked.

Uncharacterized TPS C-terminal sub-sequences were then obtained from plant species in TremBl (Boeckmann et al. (2003)), Ensembl Plants (release 38) (Kersey et al. (2017)), and the 1000 Plants Transcriptome Project (Matasci et al. (2014)) again using a Pfam domain search. Only those sequences where the search returned a sub-sequence having both DDxx (D,E) and (N,D)Dxx (S,T,G)xxxE or two DDxx (D,E) motifs within it, and whose subsequence length was within two standard deviations of the mean C-terminal sub-sequence length of characterized STS enzymes were retained. In both sets, sequences from nonseed plant species were discarded.

### 4.1.5. Measuring sequence similarities

A distance matrix of all spermatophyte TPS C-terminal sub-sequences: characterized mono-, di and sesquiterpene synthases as well as uncharacterized enzymes, was constructed using the pairwise sequence k-tuple measure described by Wilbur and Lipman (1983), implemented in Clustal Omega (version 1.2.4) (Sievers et al. (2011))). This distance matrix was then used to construct an MDS plot using scikit-learn (version 0.19.1) (Pedregosa et al. (2011)) and plotted using matplotlib (version 2.1.2) (Hunter (2007)). A cluster-map of sequence identities between characterized STS enzymes was made using the distance matrix of just these enzymes and complete hierarchical clustering using scipy (version 1.0.0) (Eric et al. (2001)) and seaborn (version 0.8.1) (Michael et al. (2017)).

### 4.1.6. Visualizing an STS structure

The SEAT tobacco 5-epi-aristolochene synthase structure from the Protein Data Bank (PDB) (Bernstein et al. (1977)) was used to visualize known TPS motifs, along with Mg^2+ ions and farnesyl hydroxyphosphonate (FHP) substrate analog. Visualization was done in Pymol 2.1 (DeLano (2002)).

### Acknowledgements

This work is part of the research programme Novel Enzymes for Flavour and Fragrance with project number TTW 15043 which is financed by the Netherlands Organisation for Scientific Research (NWO).

### Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jphytochem.2018.10.020.

### References


of small molecules and biological activities.


Little, D.B., Croteau, R.B., 2002. Alteration of product formation by directed mutagenesis and truncation of the multiple-product sesquiterpene synthases $\delta$-selinene synthase and $\delta$-bisabolol by a sesquiterpene synthase from cha-


