

# The ABCs (and Ds) of Class I Terpene Cyclase ( $\alpha$ C), *Trans*- Head-to-Tail ( $\alpha$ HT) and Head-to-Head ( $\alpha$ HH) Prenyltransferase Structures and Mechanisms of Action

Chun-Chi Chen<sup>1, #</sup>, Satish R. Malwal<sup>2, #</sup>, Xu Han<sup>3, #</sup>, Weidong Liu<sup>1,3</sup>, Lixin Ma<sup>1</sup>, Chao Zhai<sup>1</sup>, Longhai Dai<sup>1</sup>, Jian-Wen Huang<sup>1,3</sup>, Alli Shillo<sup>2</sup>, Janish Desai<sup>4</sup>, Xianqiang Ma<sup>5-7</sup>, Yonghui Zhang<sup>5-7</sup>, Rey-Ting Guo<sup>1,3</sup> \* and Eric Oldfield<sup>2, \*</sup>

<sup>1</sup>State Key Laboratory of Biocatalysis and Enzyme Engineering, Hubei Collaborative Innovation Center for Green Transformation of Bio resources, Hubei Key Laboratory of Industrial Biotechnology, School of Life Sciences, Hubei University, Wuhan 430062, China

<sup>2</sup>Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

<sup>3</sup>Industrial Enzymes National Engineering Laboratory, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 200208, China

<sup>4</sup>Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

<sup>5</sup>School of Pharmaceutical Sciences, MOE Key Laboratory of Bioorganic Phosphorus Chemistry and Chemical Biology, Tsinghua University, 100084 Beijing, China

<sup>6</sup>Joint Graduate Program of Peking-Tsinghua-NIBS, School of Life Sciences, Tsinghua University, 100084 Beijing, China

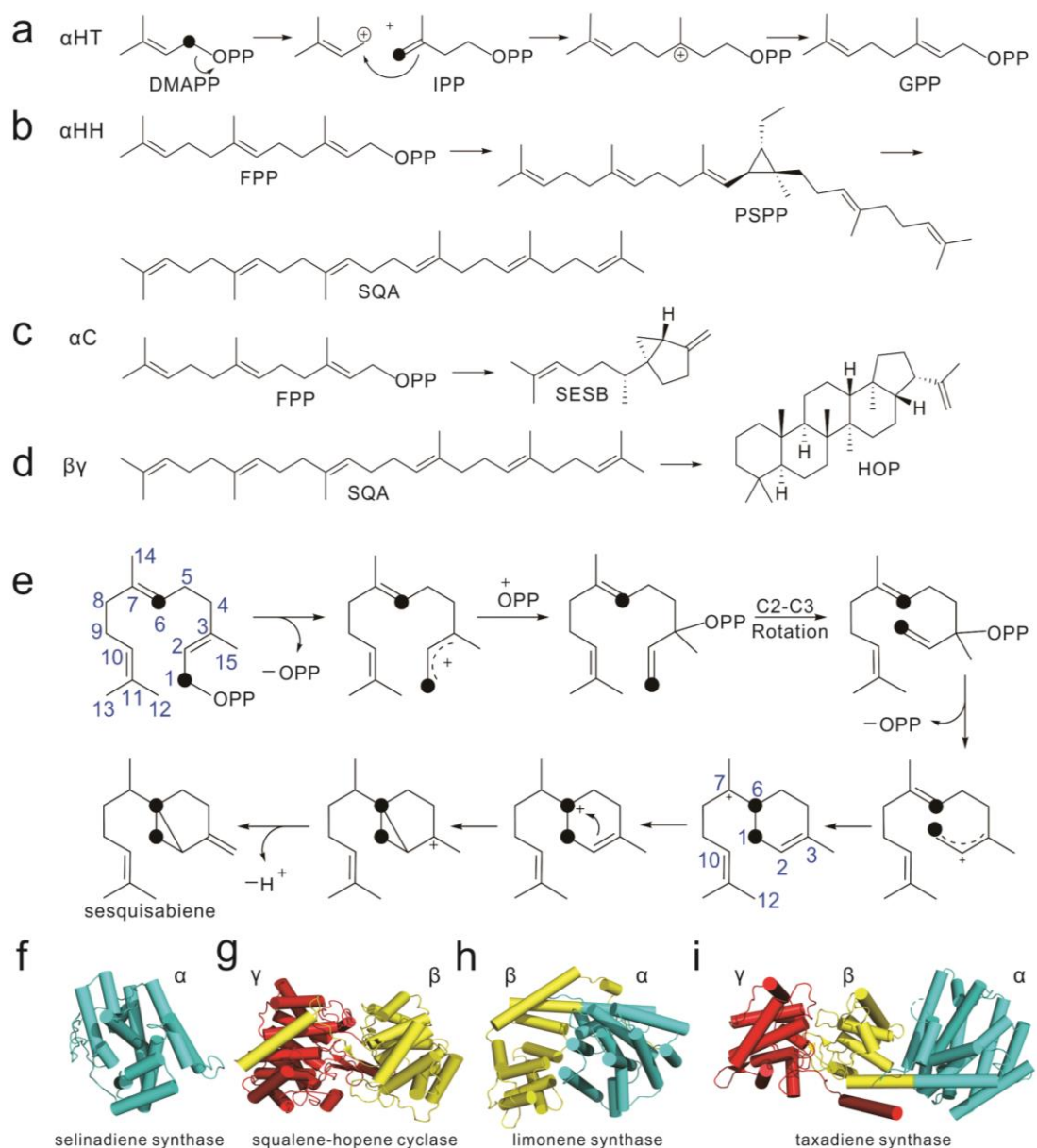
<sup>7</sup>Collaborative Innovation Center for Biotherapy, Sichuan University, Chengdu, 610041 Sichuan, China

**ABSTRACT:** We report the structures of the terpene cyclases *Santalum album* L. sesquisabinene synthases 1 and 2 and of santalene synthase, in apo forms, and with the sesquisabinene synthases, bound to either farnesyl diphosphate (FPP), farnesyl S-thiolodiphosphate, FPP containing a POP bridging O-to-CCl<sub>2</sub> substitution, or to sabinene, leading to a sequential mechanism for substrate binding and catalysis. We trapped early pre-catalytic inactive open forms that show how ligands initially bind to the apo-proteins, then when the pocket closes, catalysis can proceed. We also show that there are strong structural similarities between the most highly conserved residues in class I cyclases and those in head-to-tail ( $\alpha$ HT) trans-prenyl transferases—outside the well-known DDXXD-like and NSE/DTE-like domains. In the  $\alpha$ HT prenyltransferases there is a highly conserved Thr>Gln>Asp>Tyr motif and in the cyclases, a similar Thr>Arg>Asp>Tyr domain, these residues forming very similar, extended H-bond networks (rmsd ~1.4 Å) that are involved in catalysis, leading to the proposal that there are 3 key domains in both the cyclases and the  $\alpha$ HT prenyltransferases: The AC-domain that binds MgA and MgC; the B domain that binds MgB and leads to pocket closure, ionization, and condensation or cyclization; and the D-domain H-bond network, involved in H<sup>+</sup> elimination. In  $\alpha$ HH prenyltransferases the overall folds and MgABC motifs are similar to those found in the cyclase and  $\alpha$ HT proteins, but the full Thr>Arg/Gln>Asp>Tyr domain is absent and instead there are Tyr/Asp or Tyr/Glu residues that bind to MgC and are highly conserved. Overall, the results are of general interest since they show unexpected similarities between the enzymes that produce the most diverse molecules on Earth:  $\alpha$ HT and  $\alpha$ HH prenyltransferases, and terpenoid cyclases.

## INTRODUCTION.

There is great interest in the structures and mechanisms of action of isoprenoid and terpene/terpenoid synthases<sup>1-2</sup> because these enzymes produce molecules that have diverse and important functions and there are ~80,000 such compounds known.<sup>3</sup> Terpene or isoprenoid biosynthesis begins with the ionization of dimethylallyl diphosphate (DMAPP) to form the dimethylallyl cation which then reacts with the olefinic double bond in isopentenyl diphosphate (IPP) to form a geranyl cation which on H<sup>+</sup> elimination forms geranyl diphosphate (GPP), as shown in Figure 1a. The same type of ionization/condensation/elimination can then be repeated to form longer-chain species such as (C<sub>15</sub>) farnesyl diphosphate (FPP),

(C<sub>20</sub>) geranylgeranyl diphosphate (GGPP), (C<sub>35</sub>) heptaprenyl diphosphate (HepPP) or (C<sub>40</sub>) octaprenyl diphosphate (OPP) in reactions catalyzed by the respective prenyl diphosphate synthases (e.g. farnesyl diphosphate synthase, FPPS). These proteins exhibit an  $\alpha$  or more specifically an  $\alpha$ HT fold.<sup>4-5</sup> Two molecules of FPP can then react in a “head-to-head” manner to form (C<sub>30</sub>) presqualene diphosphate (PSPP), Figure 1b, which can then be converted to either dehydrosqualene or squalene in reactions catalyzed by dehydrosqualene synthase (a.k.a. CrTM) or squalene synthase (SQS), the latter reaction involving an additional, NADH/NADPH reductive step. A similar condensation reaction occurs with GGPP which is converted to the



**Figure 1:** Some reactions catalyzed by prenyltransferases and cyclases, and cartoon structures of some representative  $\alpha$ ,  $\alpha\beta$ ,  $\beta\gamma$  and  $\alpha\beta\gamma$ -fold proteins. (a)  $\alpha$ HT, GPP formation from IPP and DMAPP by head-to-tail transferases. (b)  $\alpha$ HH, PSPP and squalene formation from FPP by squalene synthase. (c)  $\alpha$ C, sesquisabinene B formation from FPP by sesquisabinene B synthase. (d) Hopene formation from squalene by squalene-hopene cyclase, a  $\beta\gamma$  fold protein. (e) Formation of sesquisabinene from FPP via the bisabolyl cation. The black dots highlight C1, C6. (f) Selinadiene synthase (PDB ID code 4OKZ). (g) Squalene-hopene cyclase (PDB ID code 1H3B). (h) Limonene synthase (PDB ID code 2ONG). (i) Taxadiene synthase (PDB ID code 3P5P).

$C_{40}$  by the enzyme phytoene synthase (PHYS). The structures of CrtM<sup>6-7</sup> and SQS<sup>8</sup> are known and both again exhibit a highly  $\alpha$ -helical fold,  $\alpha$ HH. Isoprenoid diphosphates such as FPP can also be cyclized to terpenes such as sesquisabinene,<sup>9</sup> Figure 1c, in reactions catalyzed by class I terpene cyclases, called here  $\alpha$ C cyclases.<sup>5</sup> In addition, the products of the  $\alpha$ HH prenyltransferases can undergo electrocyclic reactions, for example, the conversion of squalene to hopene, Figure 1d, catalyzed by squalene-hopene cyclase (SHC), a class II cyclase.<sup>10</sup> In this work, we were interested in the structure and mechanism of action of sesquisabinene synthase, a Class I cyclase, in which several bond rotations are involved, Figure

1e. Class I cyclases all contain the  $\alpha$ -fold originally discovered in FPPS<sup>11</sup> and a representative protein structure cartoon of one such cyclase, selinadiene synthase<sup>12</sup>, is shown in Figure 1f. Class II cyclases catalyze more extensive, electrocyclic transformations and contain the “ $\beta\gamma$ ” fold as found in squalene-hopene cyclase, Figure 1g. There are also many cyclases with “fused”  $\alpha\beta$  (Figure 1h, limonene synthase) structures as well as some  $\alpha\beta\gamma$  (Figure 1i, taxadiene synthase) structures and in previous work<sup>5</sup> we proposed that modern plant terpene cyclases might have originated by fusion of ancestral  $\alpha$ HT prenyltransferases with  $\beta\gamma$  cyclases to form  $\alpha\beta\gamma$  proteins that then evolved<sup>13</sup> via exon loss and recombination to the modern  $\alpha\beta$  and  $\alpha$  cyclases.

With  $\alpha\beta$  proteins, only the  $\alpha$  domain contains catalytic residues, but with  $\alpha\beta\gamma$  proteins, activity can be from  $\alpha$  only,<sup>14</sup>  $\beta\gamma$  only,<sup>15</sup> or  $\alpha+\beta\gamma$ .<sup>16</sup>

There have been many reports of the structures and mechanisms of action of these proteins<sup>1-2</sup> but there still remain questions as to the roles of some residues in ligand binding and catalysis, the time-line for ligand (e.g. FPP,  $\text{Mg}^{2+}$ ) binding, as well as the structural relationships between  $\alpha\text{C}$ ,  $\alpha\text{HT}$  and  $\alpha\text{HH}$  proteins. Here, we sought to answer the following questions. First, what are the structures and mechanisms of action of three class I sesquiterpene synthases? Second, what is the fingerprint of a cyclase? Third, how similar are the structures and mechanisms of action of the class I terpene cyclases  $\alpha\text{C}$ , and the  $\alpha\text{HT}$  and  $\alpha\text{HH}$  prenyltransferases? For example, are there highly conserved catalytic domains that have not previously been recognized? What are the main differences in ligand-binding locations? Are there previously unrecognized, conserved residues in the non-catalytic domains in the  $\alpha\beta$  proteins?

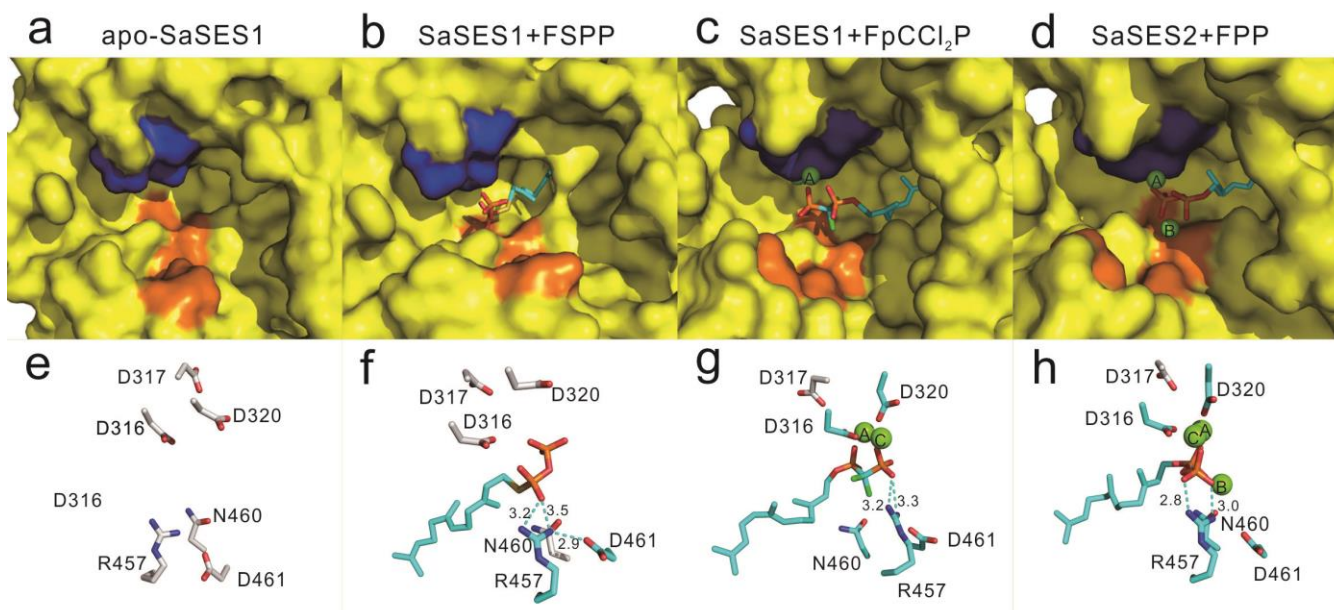
We first determined the structures of three sesquiterpene synthases, all from *Santalum album* L. (all  $\alpha\text{C}$ , class I cyclases): sesquisabinene B synthase 1 (SaSES1), sesquisabinene B synthase 2 (SaSES2), and santalene synthase (SaSS), in the presence or absence of various isoprenoid ligands. SaSES1 and SaSES2 both produce ~95% sesquisabinene B and ~5% sesquiphyllylandrene from farnesyl diphosphate,<sup>9</sup> while SaSS produces approximately equal amounts of  $\alpha$ -santalene,  $\beta$ -santalene, and *exo*- $\alpha$ -bergamotene, molecules of commercial interest as fragrances. The crystallographic results were of interest since all structures exhibited an “open”  $\alpha$ -fold—even with bound ligands—and this led us to a more general consideration of the catalytic mechanisms of action of the  $\alpha\text{C}$  cyclases, as well as of the  $\alpha\text{HT}$  prenyltransferases since while the role of many groups—involved in  $\text{Mg}^{2+}$  and diphosphate binding and ionization—are well known<sup>1-2</sup>, the roles of other residues are less well understood. Remarkably, we find that there are in fact a subset of residues in the class I cyclases and in the  $\alpha\text{HT}$  prenyltransferases (such as FPPS) that are actually the most highly conserved residues in these systems, and we propose that they have common roles in catalysis. We then compared these structures with those of the  $\alpha\text{HH}$  prenyltransferases CrtM and SQS, finding clear differences among the most highly conserved residues, but also interesting similarities in terms of ligand binding pockets and the presence of the MgABC clusters.

## RESULTS AND DISCUSSION

We first cloned, expressed, purified, crystallized and solved the X-ray crystallographic structures of apo-SaSES1, apo-SaSES2 and apo-SaSS using the methods described in the Supporting Information. Full crystallographic data acquisition and structure refinement details are given in Table S1. All three proteins adopt the  $\alpha\beta$  fold seen in many other class I terpene cyclases and the overall structures (stereo-views shown in Figures S1 and S2), are very similar to grey poplar (*Populus x canescens*) isoprene synthase<sup>17</sup> with a C $\alpha$  rmsd (alpha-carbon root mean square deviation) of ~1.45 Å, followed by *Salvia officinalis* bornyl diphosphate synthase<sup>18</sup> (a C $\alpha$  rmsd ~1.55 Å), as determined by use of the SSM server.<sup>19</sup> All three apo proteins crystallized in an “open” conformation in which the residues that close the hydrophobic pocket seen in many fully-liganded proteins<sup>1-2</sup> form a loop or are disor-

dered, facilitating ligand entry. In this work we used N-terminus truncated ( $\Delta 27$ -31 residue) species since they resulted in better diffraction (as low as 1.6 Å) but our structure of SaSES1 is very similar to that of the full length protein (at 1.9 Å resolution) reported<sup>20</sup> during preparation of this work (~0.3 Å C $\alpha$  rmsd, Table S2) in which the same residues we truncated were actually not observable, due to disorder PDB ID code 6O9Q.<sup>20</sup> Next, in order to learn more about the catalytic mechanisms of the sesquisabinene synthases, we obtained structures of SaSES1 and SaSES2 (which differ primarily in the  $\beta$ -domain outside the active site region, Figure S3) bound to either FPP, the *S-thiolo* analog of FPP (FSPP), FPP containing a POP bridging O $\rightarrow$ CCl<sub>2</sub> group (FpCCl<sub>2</sub>p), or sabinene (a homolog of sesquisabinene that lacks the additional isopentenyl fragment found in sesquisabinene B). Ligand electron density maps (2FoFc with 2 $\sigma$  and 1 $\sigma$  contours) and a ligand structure-superposition are shown in Figures S4a-f and omit maps are in Figures S4g-k. SaSS did not yield ligand-bound structures suitable for crystallography. Data acquisition and structure refinement details are in Table S1. We anticipated that FPP might not bind intact to these proteins (since it might react), so we also used the unreactive FSPP, as well as another analog, FpCCl<sub>2</sub>p. The latter species contains a CCl<sub>2</sub> fragment and we reasoned that it might bind to the proteins by displacing whichever Mg of the 3 (MgABC) typically found in fully liganded structures that was least strongly bound. We found that all four ligands—FPP, FSPP, FpCCl<sub>2</sub>p as well as sabinene—bound to the “open” or apo-form protein with no closure of the active site, Figure 2. This could be due to our use of a truncated protein (required for crystallization) and/or ligand soaking with apo crystals (since co-crystallization was not successful). In either case, the results obtained necessitated a revision of our initial plan—which was to determine closed form SaSES/SaSS structures—but led to a more general question: what are the “core” residues required for ligand binding and catalysis in these as well as other  $\alpha\text{C}$  cyclases (containing  $\alpha$ ,  $\alpha\beta$  or  $\alpha\beta\gamma$  folds)—as well as in the prenyltransferases. Based on initial inspection of our “open” structures we were interested in investigating a number of questions including the sequence of ligand and MgABC binding, the mechanism behind the “open $\leftrightarrow$ closed” conformational changes, the role of some very highly conserved residues (outside the well-known “DDXXD” and NSE/DTE domains)<sup>2</sup> as well as a comparison between the mechanisms of action of the  $\alpha\text{C}$  cyclases and the  $\alpha\text{HT}$  prenyltransferases, such as farnesyl diphosphate synthase (FPPS). We were also interested in a comparison between the structures of the  $\alpha\text{HH}$  prenyltransferases which are in a sense, cyclases (since they typically make or utilize presqualene diphosphate, which contains a cyclopropyl group). In the following sections, we consider first the mechanism of action of SaSES and other  $\alpha\text{C}$  cyclases. Then, we investigate the possible roles of the most highly conserved residues in the  $\alpha\text{C}$  cyclases and the  $\alpha\text{HT}$  prenyltransferases, followed by a structure/function comparison between the  $\alpha\text{C}$  cyclases and the  $\alpha\text{HT}$  and  $\alpha\text{HH}$  prenyltransferases.

**General Features for Class I Cyclase Catalysis.** Our structural results lead to the following proposals for the mechanism of action of the sesquisabinene synthases and of other class I terpene cyclases in which we use SaSES1, 2 and SaSS as model systems. Sequence alignments for these proteins are shown in Figure S5, representative structure superpositions are shown in Figures S6a,b and C $\alpha$  rmsd values between the



**Figure 2.** Ligand binding to sesquisabinene synthases. (a) Apo-SaSES1 active site surface view (PDB ID code 6A1I) showing open pocket. Blue = DDXXD; Orange = RNDTE motif. (b) Surface view of SaSES1 + FSPP (PDB ID code 6A1E). (c) Surface view of SaSES1 + FpCCl<sub>2</sub>p (PDB ID code 6A1D). (d) Surface view structure of FPP bound to SaSES2 (PDB ID code 6A2C). (e)-(h): Active site residues and Mg<sup>2+</sup> interactions. (e) Apo-SaSES1 (PDB ID code 6A1I). (f) SaSES1 + FSPP (PDB ID code 6A1E). (g) SaSES1 + FpCCl<sub>2</sub>p + MgAC (PDB ID code 6A1D). (h) SaSES2 + FPP + MgABC (PDB ID code 6A2C). Distances shown are in Å units.

different structures are shown in Table S2, including results for full length SaSES1 with and without the bisphosphonate drug, ibandronate (PDB ID code 6O9P).<sup>20</sup> All three apo proteins crystallized in an “open” conformation as shown for example, with SaSES1 in Figure 2a. Surface views of the three ligand-bound structures are shown in Figures 2b-d in which there is clearly a pocket into which the hydrophobic farnesyl side-chain can penetrate. For SaSES1 crystals, there is only one polypeptide chain in an asymmetric unit. For SaSES2 and SaSS crystals, four polypeptide chains are found in an asymmetric unit and they are essentially the same. Therefore, chain A in each SaSES2 and SaSS structure was chosen for examination.

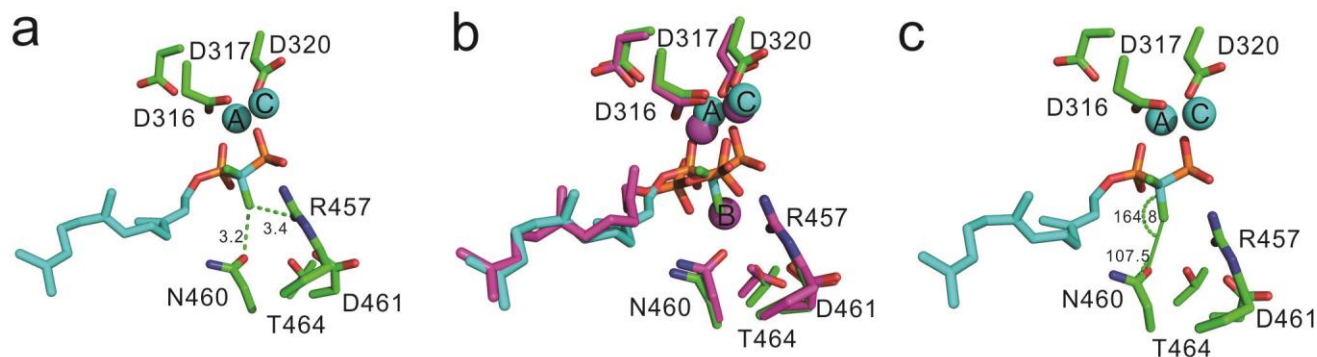
For clarity in describing the  $\alpha$ C cyclase,  $\alpha$ HT and  $\alpha$ HH prenyltransferase structures to be discussed below, we show Figure 2a-d, amino-acids that are part of the MgAC “DDXXD” motif, which we call the AC-domain, as blue surfaces. In e.g. FPPS, these residues comprise the first aspartate-rich domain, the FARM. The residues colored in orange are the B-domain residues that in closed-form proteins bind to MgB and in cyclases are the NTE/DSE residues.<sup>1-2</sup> More detailed stick-structure views of the ligand and active site residues are shown in Figures 2e-h and show key Mg-ligand and H-bond interactions. Note that all of these stick-diagram ligand-and-interaction views are from the back of the structures shown in Figure 2a-d since there are key residues that are present on the back-face of the pocket, as detailed below, that are more readily visualized from this perspective.

As to the mechanism or sequence of ligand binding and the relevance of “open”, ligand-free structures: in cells, the concentration of Mg<sup>2+</sup> is high, typically ~15-18 mM<sup>21</sup>, but almost all is bound to ATP ( $K_d$ ~78  $\mu$ M, ATP concentration ~5 mM), phospholipids, other proteins, nucleic acids, nucleotides and chromatin, for example, so the cytosolic free [Mg<sup>2+</sup>] is only ~

0.5-1 mM.<sup>21</sup> Consequently, cyclases and prenyltransferases in cells are very unlikely to bind Mg<sup>2+</sup> in the absence of a “helper” ligand. That is, it is only on binding of an isoprenoid diphosphate, such as FPP, that metal binding ensues. The farnesyl chain is thus expected to bind first with its hydrophobic side-chain entering the hydrophobic cavity in the protein and this is what is observed with the FSPP structures shown in Figure 2b,f. Mg<sup>2+</sup> are not expected to bind strongly to carboxylates alone, but they can bind strongly to diphosphate groups due to a chelate effect and it is the combined interaction between Mg<sup>2+</sup>, the diphosphate and the Asp residues that facilitates Mg<sup>2+</sup> binding. It is important to note here that in the FSPP structure the diphosphate headgroup as well as the 1st isoprene groups have higher B-factors than the protein atoms or the second and third more distal isoprene units, as can be seen in Figure S4a, and are thus more disordered. However, that is not unexpected since there are no Mg<sup>2+</sup> present to pin down the diphosphate group.

The (hydrophobic) farnesyl side-chain thus first enters the hydrophobic cavity in SaSES in an extended, low-energy conformation. After the side-chain has entered, we propose that a polar diphosphate (or here, a diphosphate analog PO<sup>-</sup> group) binds to the 100% conserved Arg-457 with a P-O...N<sup>H</sup> distance of ~2.7 Å, Figure 2f (PDB ID code 6A1E). This Arg, in turn, exhibits a strong electrostatic (salt-bridge; Coulomb) interaction between its N<sup>H</sup> and the Asp-461 O<sup>H</sup> (N<sup>H</sup>...O<sup>H</sup> ~3.0 Å) which helps orient the diphosphate for subsequent MgAC binding. This RD pair is strictly conserved in all  $\alpha$ ,  $\alpha\beta$  and  $\alpha\beta\gamma$  cyclases (with “active”  $\alpha$  domains), based on JPred4<sup>22</sup> sequence alignments, Figures S7-S9, and is essential for catalysis. Following binding of the diphosphate (or diphosphate analog) group, MgA and MgC bind. There are 3 Mg<sup>2+</sup> in fully-liganded  $\alpha$ C cyclase structures, but we propose that—at least in our systems—MgA and MgC bind first, Figures 2c, g. They





**Figure 3.** Comparison between FPP and FpCCl<sub>2</sub>p ligand-bound SaSES structures in the active site region. (a) SaSES1.FpCCl<sub>2</sub>p (PDB ID code 6A1D) indicating Cl-Asn, Arg interactions. Distances shown are in Å units. FpCCl<sub>2</sub>p is shown in cyan. (b) SaSES1.FpCCl<sub>2</sub>p superimposed on SaSES2.FPP (pink, PDB ID code 6A2C). (c) SaSES1.FpCCl<sub>2</sub>p showing halogen-bond interaction with Asn-460. Values shown are in halogen bond angles, in degrees.

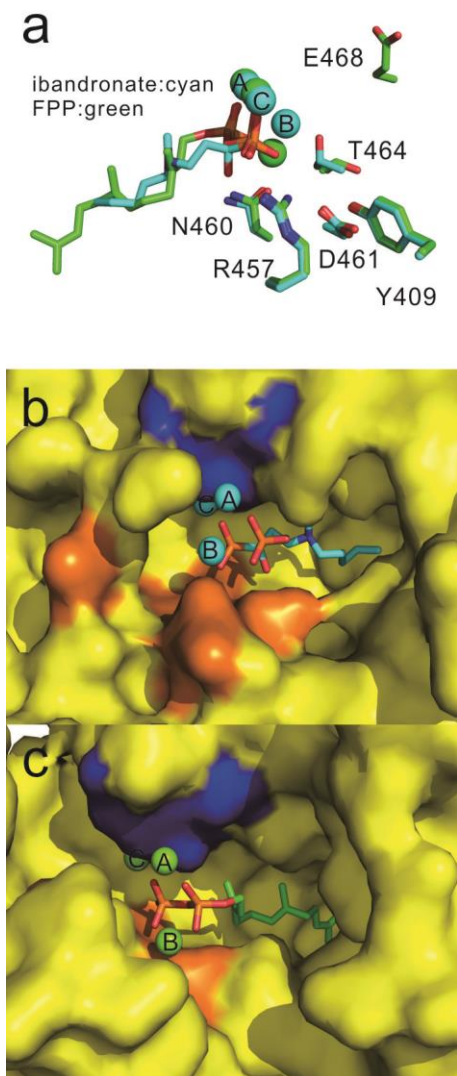
both bind to the DDXXD domain, here, by interacting with Asp-316 and Asp-320. The DDXXD domain exhibits a very similar overall conformation in both the apo and FSPP-bound structures, the main difference being a change in the C<sup>α</sup>C<sup>β</sup>C<sup>γ</sup>O<sup>δ1</sup> torsion, and the ligand-bound conformation of these residues is that found in closed, fully ligand-bound (substrate+MgABC) structures. In addition to binding to these two Asps, MgAC also bind to single oxygens on each phosphonate group, Figures 2c, g (here, in FpCCl<sub>2</sub>p). The third Mg<sup>2+</sup>, MgB, then binds and interacts with the diphosphate group in FPP as well as with Asn-460, Figures 2d, h (with this N/D being present in all αC cyclases). In the case of the FpCCl<sub>2</sub>p structure with 2 Mg<sup>2+</sup>, it appears that one of the Cl atoms interacts with Arg-457 and Asn-460, Figure 3a, thereby blocking a potential MgB binding (Figure 3b), suggesting that MgB binding is relatively weak, at least in the open state. This makes sense because there is only one MgB protein-ligand interaction and that is with a neutral residue, Asn-460, whereas MgAC binds to two anionic groups, Asp-316 and Asp-320, expected to form a stronger, Coulombic interaction. The Cl atom that appears to block MgB binding is at 3.2 Å from Asn-460 (which would bind to MgB), the C-Cl...O bond angle is 164.8°, and the Cl-O-C angle is 107.5°, Figures 3a,c. These values are those expected for a halogen bond (≤ 3.27 Å for Cl...O; C-Cl...O ≈ 165° and Cl...O-C ≈ 120°) in biological molecules.<sup>23</sup> In addition, a charge-dipole interaction with Arg-457 (at 3.4 Å) is expected.

This sequence of metal binding to an α-fold protein is also of course of general interest for αHT as well as αHH prenyl-transferases. Indeed, as shown in Figure S10, in three structures of FPPS,<sup>24</sup> only MgA and MgC were observed, bound to DMAPP, GPP or FPP, the 3<sup>rd</sup> Mg<sup>2+</sup>, MgB, being absent. This is consistent with our proposal that MgAC bind first. While we might be biased in citing these three structures, we note that when viewing the 20 (αC, αHT and αHH) structures discussed previously in our proposal for the structure of the αβγ proteins<sup>5</sup> it can be seen (Figure S11) that of the 4 proteins containing 1 or 2 Mg (i.e. not 0 or 3 Mg), in 3 out of 4 cases it is only MgA and/or MgC that bind, Figure S11. Plus, 19/20 GGPPS structures with bound bisphosphonates contain only MgAC.<sup>25</sup>

What is also particularly interesting about all four structures: apo, +FSPP, +FpCCl<sub>2</sub>p/MgAC and +FPP/MgABC is that in every case, the FPP (or analog) ligand binds to the “open” or “apo” form of the protein, with the highly conserved

Thr-464, Glu-468 seen in other class I cyclases being located in a more distant loop, and in some cases, these residues are not observed at all, due to disorder. Sabinene also binds to this open pocket. While in a formal sense we cannot comment on the “closed” forms of SaSES1, 2 or SaSS—we have no structures—it is of interest that all of the eight structures we obtained exhibit an “open” fold with the very highly conserved DDXXD (AC-domain) and Arg-452, Asn-460 and Asp-461 residues adopting the same or very similar conformations in each structure. Plus, these “local” conformations are the same or very similar to the conformations found in other fully liganded, closed-form proteins. The only obvious difference between ligand-bound “open” structures—in particular that of FPP+MgABC—and closed class I cyclase structures is, then, that Thr-464, Glu-468 (part of the canonical DTE-motif) binding to MgB is not observed. We thus propose that it is the binding of these groups to MgB which closes the pocket, forming the catalytically-competent structure, and that MgB binding is generally a late event in many αC and αHT proteins.

As to the reasons for the observation of “open” crystal structures: while it is possible that our deletion of 27-31 N-terminal residues (from the non-catalytic β-domain) that was required for crystallization may have favored formation of open structures, we do obtain very similar activity to full length protein (SaSES1  $K_{cat}/K_m = 0.28 \times 10^5 \mu\text{M}^{-1} \text{min}^{-1}$  Table S3, versus a reported<sup>9</sup> value of  $0.3 \times 10^5$ ; and for SaSES2,  $0.29 \times 10^5 \mu\text{M}^{-1} \text{min}^{-1}$ , Table S3, versus a reported value of  $0.09 \times 10^5 \mu\text{M}^{-1} \text{min}^{-1}$ ). Moreover, the recently reported structure<sup>20</sup> of a full length apo-SaSES1 as well as of a bisphosphonate ligand-bound form that was produced by co-crystallization (PDB ID codes 609Q and 609P) were both “open” structures, meaning that the open structures seen in our case need not be attributed to truncation or to use of ligand-soaking. That raises the question as to why the SaSES1+ibandronate structure is not closed, but this might not be unexpected since as can be seen in the SaSES1+ibandronate/SaSES2+FPP superposition, Figure 4a, although the farnesyl and ibandronate side-chains are in close register, MgB in the (co-crystallized) ibandronate structure is displaced from the position of MgB in the SaSES2/FPP structure and the pocket apparently cannot close with this ligand. Surface views of both structures are shown in Figures 4b,c.

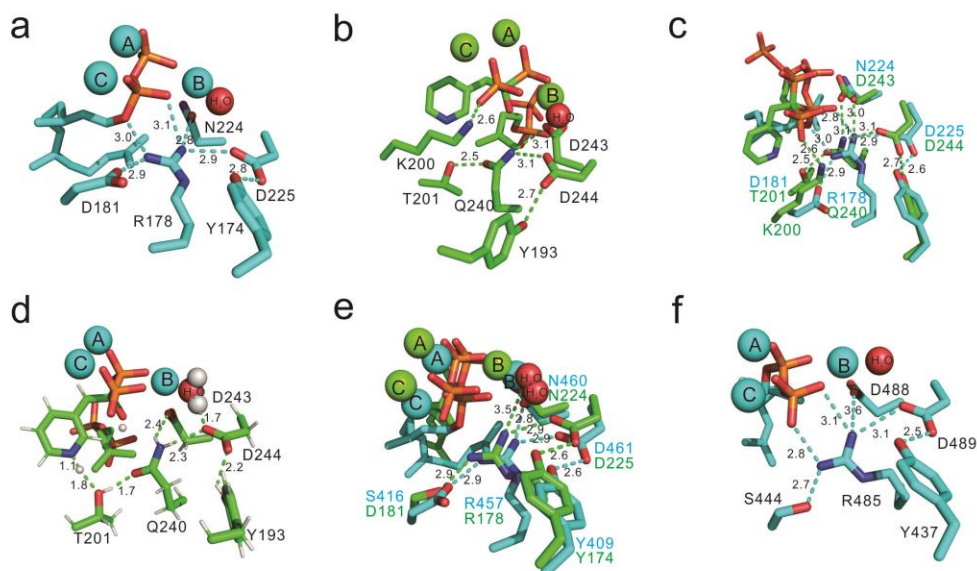


**Figure 4.** Comparison between the structures of FPP and ibandronate bound to SaSES. (a) Alignment of SaSES1.ibandronate (cyan, PDB ID code 6O9P) and SaSES2.FPP (green, PDB ID code 6A2C). (b) Surface view of SaSES1.ibandronate pocket showing open structure (PDB ID code 6O9P). (c) Surface view of SaSES2.FPP pocket showing open structure (PDB ID code 6A2C).

**Catalytic mechanisms.** Next, we consider some closed structures since they help clarify catalytic mechanisms. We show in Figures S12a,b one such “closed” structure, of dihydro-FPP bound to selinadiene synthase (PDB ID code 4OKZ) in which Ser-228 (corresponding to Thr-464 in SaSES1), Glu-232 (corresponding to Glu-468 in SaSES1) and H<sub>2</sub>O bind to MgB and the pocket is closed and no Mg<sup>2+</sup> or dihydro-FPP species are visible, the conformations of the “DDXXE” and RND (Arg-178, Asn-224 and Asp-225) groups being similar to those we find in SaSES (for the corresponding DDXXD and RND groups). A closed pocket structure is also seen with MgABC and PPi bound to aristolochene synthase (Figure S12c,d) but binding to solely MgB and PPi reveals an open pocket, Figure S12e,f. Consistent with most previous work, we propose that diphosphate ionization occurs due to bidentate chelation to Mg<sup>2+</sup>, the highly conserved Arg-457 in SaSES1 facilitating released-diphosphate stabilization. What might appear controversial is that in earlier work on selinadiene synthase<sup>12</sup> it

was proposed that there was an “effector triad” of Arg-178, Asp-181 and Gly-182 (selinadiene synthase numbering) that is completely structurally conserved in all bacterial, fungal and plant class I terpene cyclases, with Arg-178 acting as a pyrophosphate sensor. However, as shown in the sequence alignment in Figure S9, this motif is relatively poorly conserved in terpene cyclases because Asp-181 and Gly-182 are hypervariable, and it is only Arg-178 in this triad (corresponding to Arg457 in SaSES) that is essentially totally conserved (and is part of what we denote as the RD motif). Asp-181 is typically Asp, Thr or Val in bacterial class I terpene cyclases, as shown in the JPred4 structure alignment in Figure S9, while Gly-182 can also be a, Phe or Ser, Figure S9. Moreover, as can be seen in the sequence alignments based on selinadiene synthase, Figure S9, an  $\alpha$ -fold protein; of SaSES1, an  $\alpha\beta$ -fold protein, Figure S7; and of taxadiene synthase, Figure S8, an  $\alpha\beta\gamma$  protein, it is apparent that this Arg is almost always present together with a very highly-conserved Asp, which is typically located at  $\sim 3$  Å from the Arg, as shown for 8 class I cyclases in Figures S13a-h, and as discussed in more detail below. In selinadiene synthase and the other  $\alpha$ -fold-only proteins, the Arg (Arg-178 in selinadiene synthase, Figure 5a) is on a separate helix than that of the conserved Asp (Asp-225, Figure 5a), but it is just 3 residues away from the corresponding Asp in the  $\alpha\beta$  and  $\alpha\beta\gamma$  proteins (with catalytic  $\alpha$  domains). We thus propose that in addition to the DDXXD/DDXXE (AC-domain) and NSE/DTE (B-domain) Mg<sup>2+</sup>-binding domains found in all  $\alpha$ C cyclases, the RD motif forms part of a “D-domain”, a H-bond network that plays a key role in catalysis. As we discuss in more detail below, in  $\alpha$ C cyclases the D-domain residues are the Arg-Asp(H<sub>2</sub>O/MgB)-Tyr triad while in the  $\alpha$ HT prenyltransferases such as FPPS the Arg is a Gln and the Tyr can also be a His so the D-domain is Gln-Asp(H<sub>2</sub>O/MgB)-Tyr/His. Then, in selinadiene synthase, Asp-181 and Gly-182 would help determine substrate binding conformations and final product structure.

The structures we have investigated reflect the structures that would be found in solution (in cells), prior to pocket closure. In the closed structure, ionization and cyclization occurs and is expected to involve formation of pi-stacked carbocation/ $\pi$ -bonds. In SaSES, it has been proposed<sup>26</sup> that the FPP diphosphate group first ionizes, forming a carbocation, then diphosphate rebinds, the C2,3 bond in nerolidyl diphosphate rotates, then diphosphate re-ionizes, forming the allylic, nerolidyl cation required for 1,6-cyclization, Figure 1. However, based on the highly-extended farnesyl side-chains seen in our structures (Figures S4) and based on model building, it is apparent that a C2,3 rotation alone is not sufficient to bring a *cisoid* nerolidyl carbocation into an appropriate “stacked” alignment with the C6,7 double bond. To see how the final product might form—or at least, to provide ideas for future QM/MM work—we next obtained the structure of sabinene, a model compound which lacks the additional isopentenyl group found in the sesquisabinenes, bound to SaSES2. What is of interest in this structure, Figures S2d and S14, is that sabinene binds to the hydrophobic pocket (Figure S14a) with its terminal methyl group (C10) only  $\sim 1.1$  Å from where the corresponding (C15) Me in FPP is located in the SaSES/FPP structure, as illustrated in the diagram shown in Figure S14b. This leads to the idea that in addition to the C2,3 rotation, rotations about C3,4; C4,5 and C5,6 are all required in order to form an initial cation- $\pi$  complex, leading to the bisaboyl cation. Since the actual isomer formed by SaSES1 is not known, both



**Figure 5.** Structure comparisons between HsFPPS and class I terpene cyclases illustrating the D-domain H-bond networks. (a) Structure of selinadiene synthase + dihydrofarnesyl diphosphate and 3  $\text{Mg}^{2+}$  (PDB ID code 4OKZ). The  $\text{H}_2\text{O}$  bound to MgB and Asp-225 is shown as a red sphere. (b) Structure of HsFPPS + IPP + risenedronate (PDB ID code 5CG6). (c) Superposition of the structures shown in (a) and (b). The H-bond distances in the two structures are very similar. The alignment was made by using the two complete protein structures, not selected amino-acids or the  $[\text{Mg}^{2+}]_3$  cluster. (d) The neutron diffraction structure of ( $^2\text{H}_2\text{O}$ -exchanged) HsFPPS + IPP + risenedronate (PDB ID code 5CG6). Unlike the “heavy” (O,N) atom distances shown in (a)–(c) and in (e) and (f), the distances shown in (d) are  $^2\text{H}\cdots\text{O}$  or  $^2\text{H}\cdots\text{N}$  distances, so are shorter. (e) Structure superposition of SaSES2+FPP (PDB ID code 6A2C) and selinadiene synthase (PDB ID code 4OKZ). The O $\cdots\text{N}$  and O $\cdots\text{O}$  distances are very similar in both structures—even though the SaSES2 structure is an “open” one (that still actually contains both FPP and 3  $\text{Mg}^{2+}$ ) while the selinadiene structure is fully liganded and is closed. (f) Structure of limonene synthase + 2-fluorogeranyl diphosphate + 3  $\text{Mg}^{2+}$ . This structure (PDB ID code 5UV1) is also open. Red spheres are water molecules bound to MgB; the grey spheres in (d) are  $^2\text{H}$  in  $^2\text{H}_2\text{O}$  bound to MgB.

isomeric species are shown in Figures S14c,d, and after such rotations, C1 and C6 can be as close as 1.8 Å, facilitating cyclization, and the methyl group is  $\sim 3.8$  Å from a diphosphate oxygen, facilitating  $\text{H}^+$ -removal. The structure alignments between sabinene and FpCCl<sub>2</sub>p (in both SaSES structures) are very similar to that found in the alignment with FPP, but as expected are in worse agreement with the FSPP structure, which is less well resolved in the diphosphate headgroup/1<sup>st</sup> isoprenyl group region (since it lacks any  $\text{Mg}^{2+}$ ). While we of course do not have experimental structures of reactive intermediates, the comparisons between the sabinene and farnesyl-containing structures are of interest since they indicate that sabinene occupies the FPP pocket, is closely aligned with FPP, and has a relatively short distance to what would be a released diphosphate group. Moreover, it is clear that multiple rotations are required in order to bring C1 and C6 into close proximity—not just a C2,3 rotation. Interestingly, there is no requirement for any “massive” re-orientation of the farnesyl group in this reaction since, as shown in Figures 1i and S14, the reaction involves a series of rotations in the nerolidyl cation, required in order to form a productive species, and the atoms involved are basically only the first 6 out of 15 in the farnesyl group. This is unlike the situation found in many other cyclases where there are very large conformational changes required for cyclization (in e.g. 1,10- cyclizations) and in these systems, it is clear that that farnesyl groups bind in an initial, bent conformation, as discussed further below.

We also found that essentially any increase in the size of residues (shown in Figure S15) close to the terminal part of the farnesyl side-chain: Tyr288→Trp; Glu305→W and

Gly309→A,N,Y,W in SaSES1 and 2 (10 mutants in total) resulted in a  $\sim 10\times$  increase in  $K_m$  and a corresponding  $\sim 10\times$  decrease in  $K_{cat}/K_m$ , Table S3, consistent with a steric clash with the farnesyl side-chain. The E305→A mutant had the same activity as the WT protein (since the side-chain is smaller), but the E305→I mutant also had decreased activity. E305 is  $\sim 3$  Å from the nearest farnesyl group carbons but has a short (2.2 Å), strong H-bond with the backbone carbonyl in Leu-442 whose methyl groups are  $\sim 3.6$ –4 Å from the farnesyl group so it is possible that loss of this H-bond interaction results in a change in the conformation of the farnesyl group, leading to decreased activity. In any case, these mutagenesis results indicate that these residues form part of the farnesyl-binding pocket in SaSSB1, in solution.

Finally, after  $\text{H}^+$ -elimination to prevent back-reaction, product formation ensues and instead of an isoprenoid diphosphate substrate capable of binding to 3  $\text{Mg}^{2+}$  via strong electrostatic interactions, the product terpene binds much less strongly, solely via London dispersion force (attractive) interactions and can readily diffuse out of the protein. But how are  $\text{H}^+$  eliminated from the active site? Are there other highly conserved residues that might be involved in the cycle? Indeed, based on inspection of the alignments shown in Figures S7–9 it is clear that the RD dyad is actually more highly conserved than is the NSE/DTE motif, so it might play a role in  $\text{H}^+$ -elimination. Also, how are the structures of the class I cyclases related to those of the  $\alpha\text{HT}$  and  $\alpha\text{HH}$  prenyltransferases—are there mechanistic clues? The folds are similar as is the involvement of 3  $\text{Mg}^{2+}$  in catalysis, but are there other key structural and functional similarities?



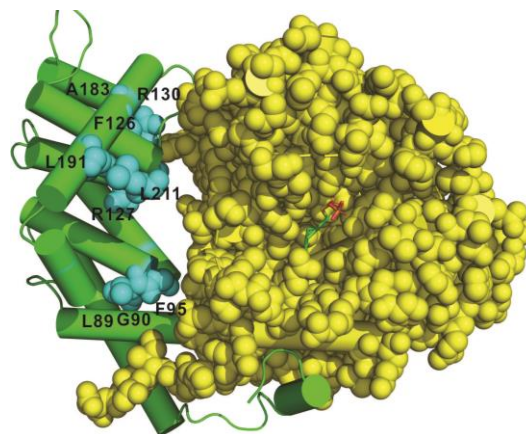
**Structural and functional comparisons between cyclases, lyases and prenyltransferases.** In previous bioinformatics work<sup>5</sup> we noted that the  $\alpha$ C cyclase domains (in  $\alpha$ ,  $\alpha\beta$  and  $\alpha\beta\gamma$  proteins) had quite similar overall folds to those found in the  $\alpha$ HT *trans*-prenyltransferases, enzymes such as FPPS, geranylgeranyl diphosphate synthase (GGPPS), heptaprenyl diphosphate synthase (HepPPS) and octaprenyl diphosphate synthase (OPPS), but were less similar to the  $\alpha$ HH proteins (CrtM and SQS). Very many of these  $\alpha$ HT structures exhibit an open fold into which substrates or substrate-like ligands can bind. For example, in FPPS from *Staphylococcus aureus*, the structure in the absence of any ligands is completely open (PDB ID code 1RTR)<sup>27</sup>, while in *E. coli* FPPS in the presence of IPP+DMASPP (Figure S11n) (PDB ID code 1RQI)<sup>27</sup>, or with IPP+ risedronate,<sup>28</sup> it is essentially fully closed. Thus, both ligand-free structures (expected to be abundant in cells) as well as closed structures that reflect catalytically-competent or inhibitor bound structures, are apparent—basically as found with the cyclases. But what, then, are the main structural differences between the  $\alpha$ C cyclases and the  $\alpha$ HT prenyltransferases—other than the lack of an IPP-binding site? The C $\alpha$  rmsd values between the cyclases and prenyltransferases are only moderate ( $\sim 4$  Å),<sup>5</sup> but are there more local structural similarities that might be important in catalysis, in both systems?

While it might be thought that the main differences would originate in the nature of the metal-binding ligands, specifically two sets of DDXXD-like ligands in the prenyltransferases but a DDXXD and an NSE/DTE (also called NTE/DSE) pair in the cyclases, in early work<sup>29</sup> it was shown that the sesquiterpene cyclases  $\delta$ -selinene synthase and  $\gamma$ -humulene synthase actually contained two sets of DDXXD residues, lacking the NSE/DTE motif. Similar results (and a structure) were then reported for  $\delta$ -cadinene synthase.<sup>30</sup> Moreover, it has been pointed out that FPPS can make terpenes, albeit at a very slow rate.<sup>31</sup> These results are of interest since they show that some professional  $\alpha$ C cyclases function with two DDXXD-like domains. Based on an initial inspection of the  $\delta$ -selinene,  $\gamma$ -humulene and  $\delta$ -cadinene sequences and our X-ray structures, it appeared that the DDXXD domains might actually be part of a larger conserved domain: RXX[DDXXD/E]XXXE, the RDE residues being associated with cyclase activity. However, the “RDE” dyad is also present in terpenoid lyases such as isoprene synthase (RXX[NDXXS]XXXE) and  $\beta$ -farnesene synthase (RXX[DDXXS]XXXE), suggesting perhaps a common role in catalysis.

In order to learn more about  $\alpha$ C cyclases,  $\alpha$ HT as well as  $\alpha$ HH prenyltransferase mechanisms, we next used the SCORECONS program<sup>32</sup> which ranks residues in terms of their essential nature, an approach we used previously with a small set of residues with FPPS<sup>33</sup> and *Staphylococcus aureus* dehydroqualene synthase (SaCrtM<sup>7</sup>). A SCORECONS score of 1.000 means a most highly conserved residue, while a zero means a totally non-conserved residue. In order to provide robust, unbiased and easy-to-read outputs, we used JPred4<sup>22</sup> alignments based on the amino-acid sequences of SaSES1 (an example of an  $\alpha$ C cyclase containing an  $\alpha\beta$ -domain), of *Allicyclobacillus acidocaldarius* squalene-hopene cyclase (AaSHC, an example of  $\beta\gamma$ C cyclase), of human FPPS (HsFPPS, an  $\alpha$ HT prenyltransferase) and *S. aureus* CrtM, an  $\alpha$ HH prenyltransferase. There were 1008 related sequences returned by the JPred4 program based on the SaSES1 target,

483 for the AaSHC target, 938 for the HsFPPS target and 1016 for the SaCrtM target. We then selected the top sequences from the JPred4 alignments that were most similar to the target sequences that yielded just one residue with a score of 1.000: the most highly conserved residue. For example, the top 331 sequences out of 1008 for SaSES1 yielded just one residue with a score of 1.000. The diversity scores (as reported by the SCORECONS program) for the alignments were SaSES1, 100.0%; AaSHC, 99.2%, HsFPPS, 99.9% and SaCrtM, 100.0%, meaning very high diversity and consequently, informative alignments. If much smaller data sets are used then there will be many scores of 1.000, conversely, if much larger sets are used then all scores drop. In brief then: we used very large and diverse sets of sequences for each protein. SCORECONS results are shown in Table 1.

In the case of the SaSES1-based results, as can be seen in Table 1, the most highly conserved residue in the set of 331 proteins is an Arg, Arg-457 in SaSES1, part of the D-domain. The 2<sup>nd</sup>, 3<sup>rd</sup> and 5<sup>th</sup> most highly conserved residues correspond to the Asps that bind MgAC and make up part of the AC domain: Asp-316, Asp-320 and Asp-317. The next most highly conserved residue is Asp-461, part of the D-domain. This Asp is not directly bonded to any Mg<sup>2+</sup>, but is bonded via a bridging water to MgB in our and most other cyclase structures. All of these Asp residues are in the so-called catalytic or  $\alpha$ -domain. What comes as a surprise is the observation that there are also several very highly conserved residues such as Arg-130, Leu-191, Arg-279, Phe-95 and Leu-211 that are in the non-catalytic or  $\beta$ -domain, Table 1. On inspection of the SaSES1 structure, it can be seen that most of these  $\beta$ -domain residues are located close to the interface with the catalytic or  $\alpha$ -domain, as shown in Figure 6. Since there have been no reports of the catalytic activity of any  $\alpha$ -domain protein based on an  $\alpha\beta$ -protein sequence, it is apparent that the  $\beta$ -domain is required for catalytic activity, presumably facilitating proper protein folding and/or enhancing stability, with the presence of these conserved  $\beta$ -domain residues indicating that they are of particular importance for protein activity. So, the SCORECONS results for the  $\alpha\beta$  proteins indicate that both catalytic as well as more purely structural information is obtained using this bioinformatics approach. The results for the  $\alpha\beta$  proteins are reminiscent of those we reported earlier<sup>34</sup>



**Figure 6.** Structure of SaSES1 (PDB ID code 6A1E) showing conserved residues (in cyan) in the  $\beta$ -domain that are close to the  $\alpha/\beta$  domain interface. The  $\alpha$ -domain residues are shown as yellow spheres, the green cylinders represent the  $\beta$ -domain.



**Table 1.** SCORECONS results for  $\alpha$ C cyclase,  $\alpha$ HT and  $\alpha$ HH prenyltransferase and  $\beta$ C cyclase proteins.

Rank	$\alpha$ C (SaSES1)			$\alpha$ HT (HsFPPS)			$\alpha$ HH (SaCrtM)			$\beta$ C AaSHC		
	Score	Residue	Type	Score	Residue	Type	Score	Residue	Type	Score	Residue	Type
1	1.000	317	D	1.000	104	D	1.000	129	Y	1.000	76	G
2	0.997	320	D	0.996	103	D	0.996	176	D	0.986	78	W
3	0.996	316	D	0.995	240	Q	0.996	172	D	0.982	415	G
4	0.994	457	R	0.983	93	E	0.995	165	Q	0.977	312	W
5	0.980	279	R	0.982	236	G	0.983	114	D	0.975	376	D
6	0.978	461	D	0.980	112	R	0.980	168	N	0.966	32	W
7	0.976	404	P	0.980	239	F	0.980	138	G	0.966	485	W
8	0.975	266	W	0.980	107	D	0.972	183	Y	0.966	489	W
9	0.964	302	R	0.979	200	K	0.971	181	R	0.966	533	W
10	0.964	394	E	0.979	244	D	0.969	48	D	0.966	556	T
11	0.961	327	E	0.979	261	D	0.968	41	Y	0.966	171	R
12	0.961	540	D	0.967	113	R	0.968	134	A	0.966	479	Q
13	0.959	397	W	0.965	100	L	0.967	45	R	0.966	30	G
14	0.947	284	Q	0.965	243	D	0.939	185	S	0.966	483	G
15	0.944	536	Y	0.963	57	K	0.910	163	S	0.966	531	G
16	0.941	347	P	0.959	56	G	0.866	22	F	0.966	532	G
17	0.936	504	W	0.952	60	R	0.859	171	R	0.966	589	G
18	0.934	346	L	0.934	201	T	0.851	137	V	0.966	535	E
19	0.931	281	G	0.906	114	G	0.831	157	A	0.959	609	Y
20	0.928	338	W	0.891	351	R	0.815	52	D	0.958	495	Y

on the long chain  $\alpha$ HT prenyltransferases such as heptaprenyl diphosphate synthase in which there are 2 proteins: a large, catalytic protein and a smaller regulatory protein, both of which are required for activity, and in these systems there are conserved residues in the interface between the two proteins.<sup>34</sup>

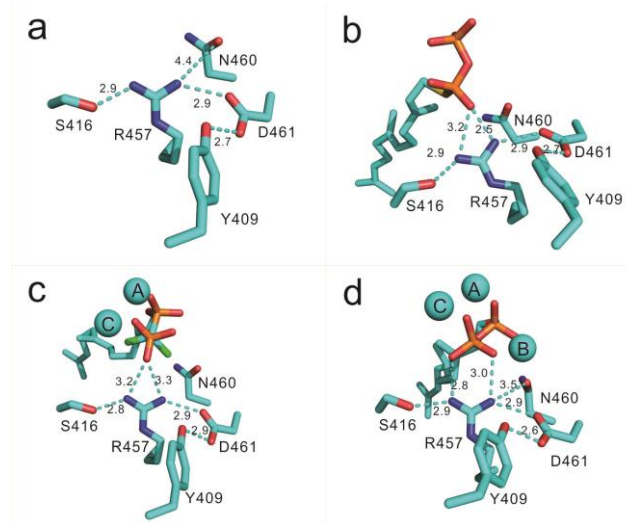
In contrast to the results we obtain with the  $\alpha\beta$  proteins—in which the most highly conserved residues play a direct role in catalysis—in the C cyclase squalene-hopene cyclase, almost all of the top scoring residues, Table 1, are Trp and Gly and they appear to play a structural role with only Asp-376, ranked 5<sup>th</sup> in the top 20 residues, being directly involved in cyclization. We find that there is moderate structural similarity between the  $\beta$ -domain in SaSES1 and the  $\beta$ -domain in AaSHC (a 4.44 Å C $\alpha$  rmsd over 221 residues), with four helices in SaSES1 overlapping 4 helices in AaSHC (Trp-312 to Ala-324; Val375 to Leu-388, Glu-446-Phe-459 and Tyr-493 to Gly-508, in AaSHC), but there is no structural similarity between the  $\beta$ -domain in SaSES1 and the  $\gamma$ -domain in AaSHC (as determined by using the VMD program<sup>35</sup>). Also, the Trp-Asp motifs are absent in SaSES1, and the conserved  $\beta$ -domain residues found in SaSES1 are not present in AaSHC, suggesting the convergent evolution of these  $\beta$ -domains.

In the case of the results based on HsFPPS, the results are generally similar to those found with the catalytic site in the SaSES1-based alignment with the first aspartate rich (AC) domain being very highly conserved but now, there is another very highly conserved residue, Gln-240 (HsFPPS numbering) whose role we shall discuss below. What is notable about the HsFPPS-based alignment is that it is likely that all highly conserved residues have a primarily catalytic as opposed to a purely structural role.

Next, we investigate the structures of the some  $\alpha$ C cyclases and  $\alpha$ HT prenyltransferases.

On close inspection of our apo, +FSPP, +FpCCl<sub>2</sub>p and +FPP structures, Figure 7, it can be seen that there is a H-bond network between the highly conserved Arg-457, the highly con-

served Asp-461, as well as with a Tyr (Tyr-409) H-bonded to Asp-461, and a 4th residue, Ser-416, that H-bonds to Arg-457, forming an extended H-bond network. The four structures have very similar Ser-Arg-Asp-Tyr conformations—not unexpected since the liganded structures were obtained by soaking. Notably, the Arg-Asp-Tyr conformations as well as the Arg-PO- distances are all very similar to those found in other closed, fully-liganded class I cyclases, Figure S13, in which there are likely H-bond networks between the ligand diphosphate group PO- (or in released PPI), the Arg guanidine group,



**Figure 7.** Structures of the Ser, Arg, Asp, Tyr residues near the MgB binding site in apo and ligand- bound sesquisabinene synthases. (a) Apo-SaSES1 (PDB ID code 6A1I). (b) SaSES1 + FSPP (PDB ID code 6A1E). (c) SaSES1 + FpCCl<sub>2</sub>p + MgAC (PDB ID code 6A1D). (d) SaSES2 + FPP + MgABC (PDB ID code 6A2C). The H-bond network (D-domain residue) distances are in Å.

the Asp carboxyl and the Tyr OH group with an average  $3.2 \pm 0.66$  Å distance between donor and acceptor atoms. What may seem puzzling here is that the SCORECONS program does not rank Tyr-409 highly (ranked 28). However, this is because in some cases this Tyr is replaced by a His (that can also H-bond to the Asp), an effect that is also seen with FPPS, as discussed below.

What is of even more interest is that we find that the  $\alpha$ C cyclase structures are highly superposable in the B- and D-domains on the H-bond network found in the  $\alpha$ HT prenyltransferase FPPS, as shown in Figure 5a-c. Figure 5a shows the structure of selinadiene synthase (PDB ID code 4OKZ), chosen because it has a high resolution structure with a bound ligand and 3 Mg<sup>2+</sup>, and is in the closed form; Figure 5b that of (closed) human FPPS (HsFPPS; PDB ID code 5CG6) and Figure 5c, their superposition, based on a full protein structure alignment. There is a  $1.17 \pm 0.36$  Å rmsd between the (3x3=9) terminal side-chain atoms in Arg-178, Asn-224 and Asp-225 in selinadiene synthase with the corresponding groups (Gln-240, Asp-243 and Asp-244) in HsFPPS—much smaller than the 4.24 Å C $\alpha$  rmsd for the overall structures, which reflect the more global cyclase/prenyltransferase structural differences (the prenyltransferases having of course a second, IPP, binding site). Notably, these residues are all essential for efficient catalysis. The Tyr (Tyr-174) in selinadiene synthase that is seen in almost all class I cyclases is located very close to the Tyr-193 in HsFPPS, in the superposition, Figure 5c. There appear, therefore, to be very similar H-bond networks in the two proteins involving residues that are not directly bonded to MgA,B or C. In the  $\alpha$ C cyclases: Thr, Arg, Asp, Tyr and in the  $\alpha$ HT prenyltransferases, Thr, Gln, Asp, Tyr. As discussed below, the Thr in the cyclases is actually the most common residue in this location. When we investigated a larger set of prenyltransferase/cyclase structures (listed below) we found a 1.4 Å rmsd for the 20 atoms (which span ~9 Å in the crystal structures) noted above—less than a single C-C bond length (1.5 Å) variation on average—versus a 3.4 Å C $\alpha$  rmsd for all protein residues. The proteins investigated and their local/global C $\alpha$  rmsd values were as follows: HsFPPS/limonene synthase ( $1.15 \pm 0.42$  Å/ $3.48$  Å); HsFPPS/selinadiene synthase ( $1.17 \pm 0.36$  Å/ $4.24$  Å); ScGGPPS/taxadiene synthase ( $1.35 \pm 0.19$  Å/ $2.30$  Å); HsFPPS/taxadiene synthase ( $1.59 \pm 0.20$  Å/ $3.37$  Å) and EcOPPS/limonene synthase ( $1.176 \pm 1.0$  Å/ $3.54$  Å). These structural similarities are of great interest since the Arg/Asp and Gln/Asp residues have been shown previously to be very important in catalysis<sup>12, 36-37</sup> as well as being very highly conserved, based on the computational results (Table 1), but they have no obvious role since they are not bound to Mg.

We thus next inspected the neutron diffraction<sup>28</sup> structure of HsFPPS with bound IPP and a bisphosphonate inhibitor, risedronate, since neutron diffraction (of <sup>2</sup>H-exchanged samples) can reveal <sup>2</sup>H positions (in e.g. <sup>2</sup>H<sub>2</sub>O, or a <sup>2</sup>H-exchanged amide side-chain) that might suggest the presence of a strong H-bond network—a potential “proton wire” for H<sup>+</sup>-elimination. In earlier work<sup>38-39</sup> we found evidence for hydrogen-bond networks amongst highly conserved, catalytically essential residues in numerous head-to-head, head-to-middle and *cis*-head-to tail prenyltransferases in which we proposed that neutral species (Asn, Gln, Ser, Tyr), not bonded to Mg<sup>2+</sup>, could act as proton shuttles. Plus, in previous work<sup>40</sup> we proposed that the protonated form of risedronate was a prenyl carbocation-diphosphate mimic or isostere, so comparisons

between the H-bond networks in the  $\alpha$ C and  $\alpha$ HT prenyltransferases is of interest.

As shown in Figure 5d, the neutron diffraction structure of FPPS+risedronate+IPP (PDB ID code 5CG6) shows that the pyridine group in risedronate is indeed protonated and it donates a H-bond to Thr-201 (NH<sup>+</sup>-O<sup>1</sup> 1.8 Å); the Thr-OH proton H-bonds to Gln-240 (OH<sup>+</sup>...O<sup>61</sup> 1.7 Å); the Gln amide H-bonds to the anionic Asp-244 (NH<sup>+</sup>...O<sup>2</sup> 2.3 Å) and the Asp-244 carboxylate oxygen is also H-bonded to the Tyr-193 OH (O<sup>1</sup>...HO 2.2 Å), the backbone H<sup>N</sup> of Trp-269 and the <sup>2</sup>H<sub>2</sub>O that is bonded to MgB. The 4<sup>th</sup> residue here, Tyr, is not a particularly highly-conserved residue (as apparent in the SCORECONS results shown in Table 1), but this is because in some  $\alpha$ HT prenyltransferases this residue is a His, as in e.g. *E. coli* FPPS (EcFPPS) (PDB ID code 1RQI), where there is a H-bond between the (non-metal-bound) Asp-245 and His-199. Indeed, all the heavy-atom distances of interest are almost the same in HsFPPS and EcFPPS.

The neutron diffraction results (on FPPS) thus suggest the presence of very similar H-bond networks in  $\alpha$ HT prenyltransferases and the  $\alpha$ C cyclases involving, for example, in selinadiene synthase, the highly conserved and essential Arg-178, Asn-224 and Asp-225 since as can be seen in Figure 5c, all the heavy atom distances are very similar in FPPS and in the cyclase, for the interactions noted above. Essentially the same distances are also seen in e.g. SaSES2+FPP, an open structure, Figure 5e, as well as in other open (but liganded) structures, such as that of limonene synthase, Figure 5f. The Arg-178/Asp-225 dyad in the cyclase and the Gln-240/Asp-244 dyad in FPPS (and in GGPPS and OPPS) thus appear to play a key role in stabilizing MgB, via H-bonding to the Asp/Asn bonded to MgB, as well as by stabilizing the Asp-H<sub>2</sub>O-MgB interaction.

Where the cyclase and FPPS structures differ in the catalytic (D-domain, *versus* IPP-binding) region is in the nature of the other residues that bind to the conserved Arg (found in the cyclases) or Gln (found in FPPS). In the cyclases, ~53% of the residues are Thr, while ~67% are Thr, Ser or Asn, all of which can H-bond to Arg. Plus, the Arg can also H-bond to prenyl diphosphate oxygens (as seen in our structures), or to PPI. In human FPPS, there is a Thr (Thr-201; PDB ID code 5cg6) that has its hydroxyl hydrogen 1.7 Å from the highly conserved Gln-240 (see e.g. Figure 5b) and this Thr is very highly conserved, although in a few cases is substituted by Ser. Another difference between the cyclases and FPPS is that in FPPS (and in the related prenyltransferases) there is now an additional highly conserved and catalytically-essential pair of residues: Lys-200 and Phe-239<sup>37, 41</sup> (HsFPPS numbering system). Lys-200 can interact with prenyl diphosphates or diphosphate and appears to play the same role as does Arg-178 (in selinadiene synthase), interacting via Coulomb interactions with PO<sup>-</sup> moieties. It is thus possible that Arg-178 (and related residues in the  $\alpha\beta$  and  $\alpha\beta\gamma$  cyclases) may have evolved from an ancestral prenyltransferase at least in part via a Gln → Arg mutation. For Gln, there are two codons, CAA and CAG. For Arg, there are six: CGT, CGC, CGA, CGG, AGA and AGG. There are thus two ways of effecting a Q→R mutation: CAA→CGA, or CAG→CGG. In EcFPPS, the Gln codon used is CAG<sup>42</sup> while that used for Arg (in e.g. SaSQS2) is CGG. This would be consistent with a single A→G mutation. Similarly, in limonene synthase (from *Mentha spicata*), the Arg codon is again CGG.<sup>43</sup> While it is not possible to say whether such a mutation actually occurred during the evolution of the

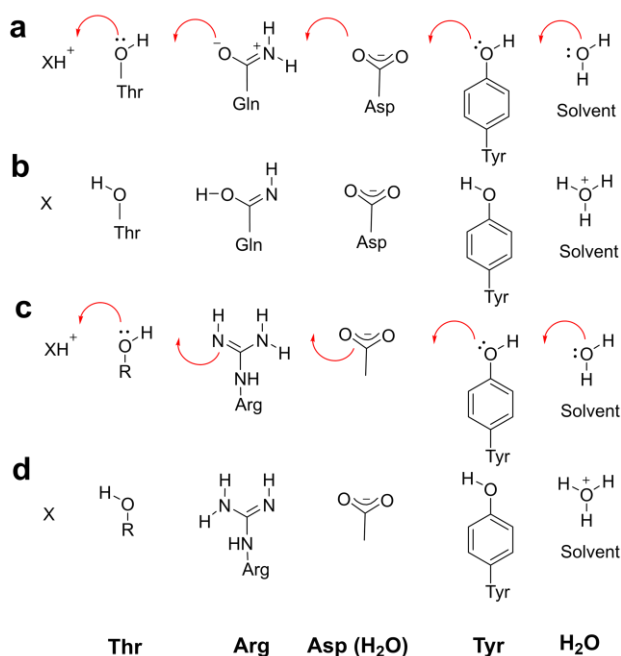
cyclases, the observation that these highly conserved and essential residues occupy similar spatial positions indicates a similar role in catalysis.

There are, therefore, very similar highly conserved H-bonded clusters of residues in the  $\alpha$ C cyclases and the  $\alpha$ HT prenyltransferases that are not directly bonded to Mg and are “outside” the standard NSE/DTE/DDXXD-like motifs. In the cyclases: Thr, Arg, Asp, Tyr, and in the prenyltransferases, Thr, Gln, Asp, Tyr. There are, of course, several hundred amino-acid differences between e.g. FPPS and the cyclases and we find that mutation of the essential Gln-240 in FPPS to an Arg, or to K200Q/Q240R or K200A/Q240R double mutants did not result in cyclase activity, rather, there was a loss in FPPS activity, which is not unexpected given the essential nature of Gln-240.

The structural, bioinformatics and mutagenesis results suggest that the D-domain residues may play a role in  $H^+$ -elimination, as shown in Figure 8, since they are highly conserved, do not bind directly to  $Mg^{2+}$ , and with Arg, Gln and Asp (not bound to  $Mg^{2+}$ ) are essential for activity and are involved in H-bond networks near the active site. As noted above, in previous work<sup>40</sup> we proposed that ring-protonated risedronate acted as a carbocation-diphosphate reactive

**Ligand side-chain binding in  $\alpha$ C cyclases and  $\alpha$ HT prenyltransferases.** In the previous section we considered binding of metal ions in the AC and B domains,  $Mg^{2+}$ -ligand interactions, intermediate isostere, so it is of interest to see that in the HsFPPS neutron structure, the risedronate pyridinium  $^2H^+$  is 1.8 Å from the Thr O $\gamma$  since the Thr-Gln-Asp-Tyr tetrad could provide a path for  $H^+$ -elimination from cationic species in the active site to solvent water, Figure 8a. There is, therefore, a need to remove a proton and transport it to solvent water (in both prenyltransferases as well as in cyclases) while at the same time protecting the active site from solvent water. In FPPS the active site pocket is protected from direct  $H_2O$  access by the highly conserved and catalytically essential<sup>37</sup> Phe-220 (*Bacillus stearothermophilus* experiment and numbering), leading to product formation as shown in Figures 8a,b. In the class I cyclases,  $H^+$ -elimination would be as in Figures 8c, d. Now, Arg acts as a base, as proposed for numerous other enzymes<sup>44</sup>, its “pK<sub>a</sub>” being lowered due to close (~4 Å) proximity to the divalent MgB. As to the nature of the cationic species  $XH^+$  (Figure 8), this could in principle be a protonated diphosphate, or a carbocation species since it seems unlikely that diphosphate alone would be a universal base that would be able to act in 1'-4 condensations as well as in 1,6-, 1,10- and 1,11-cyclizations—because it will most likely remain close to the MgABC cluster. But carbocations do need to be deprotonated by something—and the D-domain “tetrads” discussed above are one possibility, although clearly, differentiating in a quantitative manner their “structural” as opposed to a “catalytic” role—in  $H^+$ -elimination—will be challenging. together with the nature and the location of the D-domain H-bond networks. Here, we focus on a comparison between the ligand side-chain positions in  $\alpha$ C and  $\alpha$ HT proteins, and in the next section, a similar consideration of the  $\alpha$ C and  $\alpha$ HH proteins.

In previous work<sup>25</sup> we noted that in  $\alpha$ HT proteins there were four domains that could be involved in substrate, product, and inhibitor binding: **a**, **b**, **c** and **d**, shown schematically in Figure 9a. The **a** domain residues are involved in binding to MgABC and allylic substrate *diphosphate* binding in FPPS as

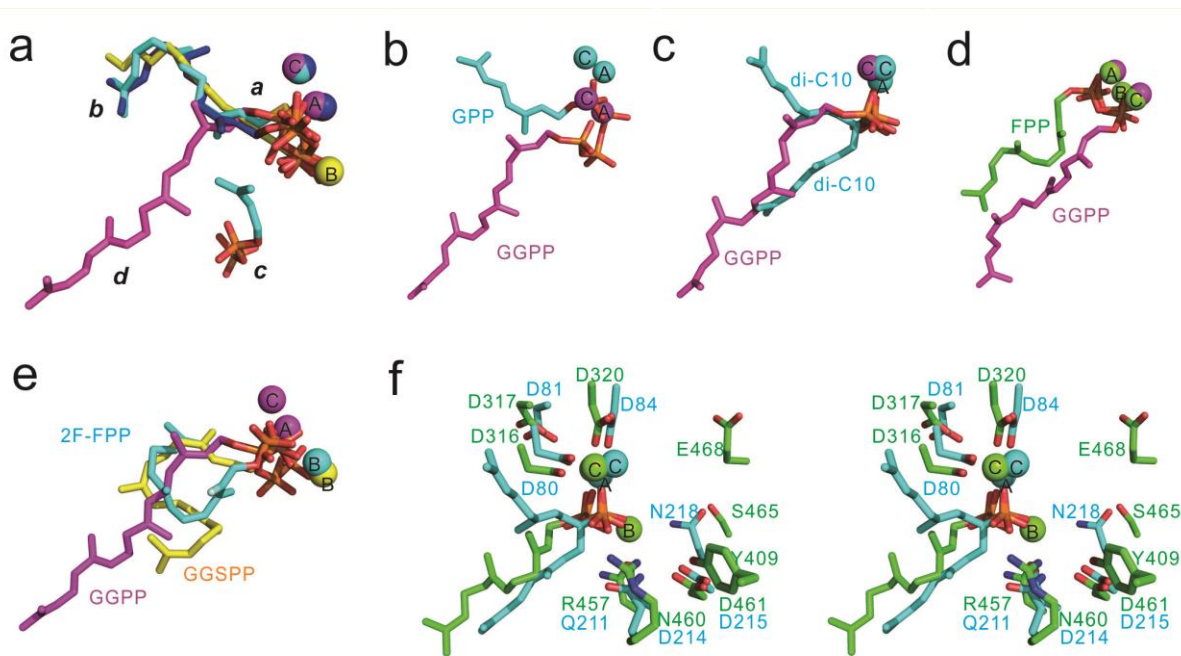


**Figure 8.** “Proton-wire” mechanisms for  $H^+$ -elimination from the active site in head-to-tail *trans*- prenyltransferases and class I terpene cyclases based on X-ray and neutron crystallographic structures, bioinformatics, and site-directed mutagenesis. (a) Initial starting point for  $H^+$ -transfer to **Thr**, **Gln**, **Asp**, **Tyr** and  $H_2O$  in HsFPPS. Thr, Gln and Asp are essentially totally conserved residues. (b) Tautomer and protonation states after  $H^+$ -transfer—from either a protonated diphosphate, or a carbocation reactive intermediate (both shown as  $XH^+$ ). Gln is the most highly conserved residue in all head-to-tail *trans*-prenyltransferases and the structures shown in (a), (b) represent the zwitterionic form of the amide side-chain in (a), or the imidic acidic form in (b), which can tautomerize back to the form shown in (a). (c)  $H^+$ -elimination pathway involving **Thr**, **Arg**, **Asp** and **Tyr** in the class I terpene cyclases showing the initial state. (d) After  $H^+$ -elimination. For simplicity, the intermediate protonated forms of Asp are not shown. The Asp that is shown is bound via  $H_2O$  to MgB and this MgB- $H_2O$ -Asp motif is found in both the prenyltransferases as well as in the class I terpene cyclases.

well as in GGPPS. The **b** domain residues are involved in allylic *side-chain* binding in FPPS and GGPPS, Figure 9a, as well as bisphosphonate-inhibitor binding in FPPS. The **c** domain residues bind the IPP diphosphate group, while the **d** domain residues bind the GGPP product (in GGPPS), as well as long-chain bisphosphonate GGPPS inhibitors.<sup>25</sup> When considering ~20 substrates, products and inhibitors (in GGPPS and FPPS) we found that ligands could bind to **ab**, **ac**, **bc** as well as in one case, **cd** sites.<sup>25</sup> In addition, we found that with 1, 1-bisphosphonates containing long side-chains, such as 1, 1-digeranyl bisphosphonate, the long-chains bound to both **ab** as well as **ad** sites in a “V-shaped” manner. Here, we use this **abcd** binding site model as a guide to compare how ligands bind to  $\alpha$ C cyclases, as well as to the  $\alpha$ HH prenyltransferases.

We show in Figure 9b a superposition of GPP (cyan) binding to FPPS (**ab** site binding) as well as of GGPP (pink) binding to GGPPS (**ad** site binding). Figure 9c shows a superposition of 1, 1-digeranyl bisphosphonate (cyan; PDB ID code



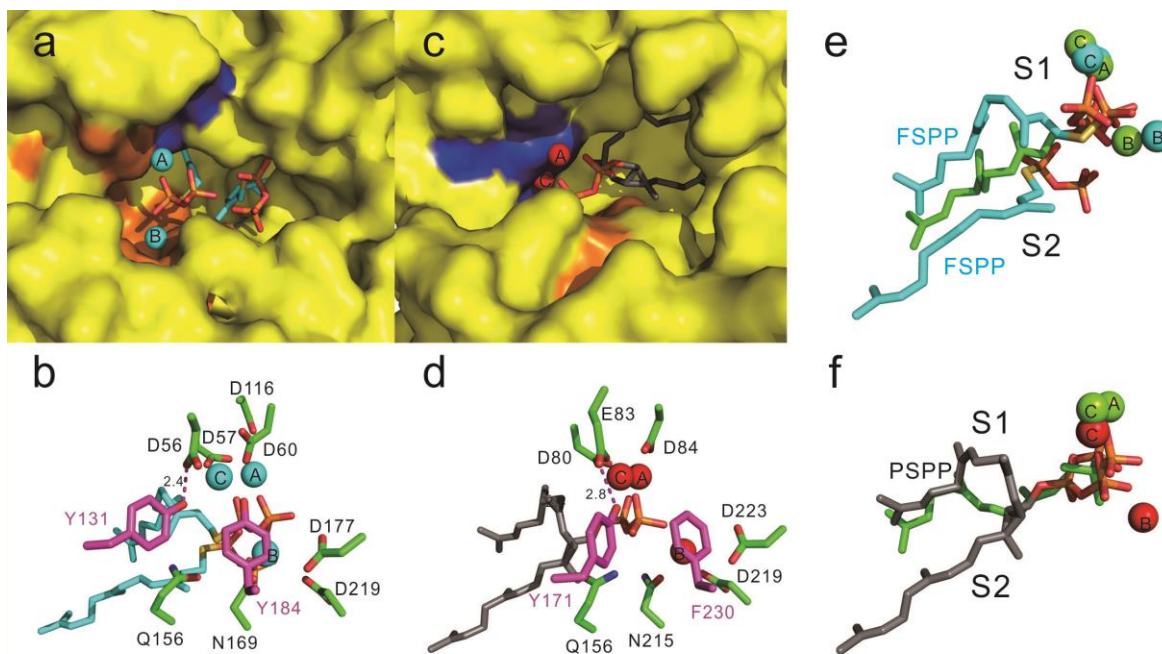


**Figure 9.** Comparison between ligand side-chain binding sites in  $\alpha$ C cyclases and  $\alpha$ HT prenyltransferases. (a) The four ligand binding sites (**a**, **b**, **c**, and **d**) found in GGPPS and/or FPPS ( $\alpha$ HT proteins). GGPP (pink) in ScGGPPS (PDB ID code 2Z4V); GSPP and IPP (cyan) in ScGGPPS (PDB ID code 2E8T); n-octyl-1,1-bisphosphonate (yellow) in ScGGPPS (PDB ID code 2Z52) and FPP (blue) in ScGGPPS (PDB ID code 2E90). (b) GGPP (pink) in ScGGPPS (PDB ID code 2Z4V) plus GPP (cyan) in *Gallus gallus* FPPS (PDB ID code 1UBW). Note that MgB is absent. (c) 1,1-Digeranyl bisphosphonate (cyan) bound to ScGGPPS (PDB ID code 2Z4W) together with GGPP (pink) bound to ScGGPPS (PDB ID code 2Z4V). (d) FPP (green) bound to SaSES2 (PDB ID code 6A2C) and GGPP (pink) bound to ScGGPPS (PDB ID code 2Z4V). (e) 2-F FPP (cyan) bound to AtAS (PDB ID code 3BNY); GGSP (yellow) bound to SmCotB2 (PDB ID code 5GUE) and GGPP (pink) bound to ScGGPPS (PDB ID code 2Z4V). (f) Stereo-view of 1,1-digeranyl bisphosphonate (cyan) bound to ScGGPPS (PDB ID code 2Z4W) aligned to FPP (green) bound to SaSES2 (PDB ID code 6A2C), together with AC, B, and D-domain ligands.

2Z4W) together with the GGPP product (pink; PDB ID code 2Z4V), both bound to GGPPS. Clearly, the GPP binding site is at  $\sim 90^\circ$  to the GGPP site, and the side-chains in the 1,1-bisphosphonate bind to both (**ab**, **ad**) sites. The question then arises as to whether the farnesyl-containing species FSPP, FpCCl<sub>2</sub>p and FPP bound to the  $\alpha$ C cyclase SaSES adopt binding poses that are similar to either of these **ab** or **ad** ligand-binding motifs. As can be seen in the superposition of FPP in SaSES2 (PDB ID code 6A2C) with GGPP in GGPPS (PDB ID code 2Z4V), Figure 9d, it is apparent that the farnesyl side-chains in ligands bound to SaSES2 occupy an **ad** site. As shown in Figure 9e, this **ad** site is also occupied with “bent” side-chains, e.g. of 2-F-farnesyl diphosphate in *Aspergillus terreus* aristolochene synthase (cyan, PDB ID code 3BNY)<sup>45</sup> as well as *S-thiolo*-geranylgeranyl diphosphate in CotB2 (yellow, PDB ID code 5GUE),<sup>46</sup> as shown in the superposition (Figure 9e) with GGPP (pink) in GGPPS. A stereo-view of the SaSES2/FPP and digeranyl bisphosphonate structures is shown in Figure 9f, and clearly indicates that the “FPPS” substrate site is not occupied. The isoprenoid side-chain in the  $\alpha$ HT proteins can thus occupy two sites: one is located between helices D and F and is involved in regulating product chain length<sup>47</sup> while the second occupies the product<sup>48</sup> site and may be involved in feedback inhibition (as shown for FPP binding to the allosteric site in FPPS<sup>49</sup> and it is this second, **ad** site (between helices C and G) into which cyclase substrates bind (and where amino-acid side-chains control substrate/product conformations). So, while there are many similarities in the AC, B and D domains between the  $\alpha$ C cyclases

and the  $\alpha$ HT prenyltransferases, the ligand side-chain positions differ with allylic  $\alpha$ HT prenyltransferase substrates binding to **ab** (helices D, F), while the  $\alpha$ C cyclase substrates bind to a site that is very similar to the **ad** site seen in GGPPS which is contoured, in most cases, to facilitate cyclization although in SaSES (product or inhibitor), the chains are quite extended.

It is also interesting to note that when inspecting numerous other  $\alpha$ C cyclase structures, it is clear that most contain either 3 Mg<sup>2+</sup> (MgABC), or are apo-structures. However, in early work it was shown that in *Aspergillus terreus* aristolochene synthase containing a 2-fluorofarnesyl ligand<sup>45</sup>, there was 1 Mg<sup>2+</sup> present, MgB (PDB ID code 3BNY), although this Mg<sup>2+</sup> was only found in one of the four chains. In each chain, the farnesyl-like ligand side-chain was highly curved into a “C”-like conformation, independent of the presence of Mg<sup>2+</sup> and the structures (obtained by soaking) were quite “open”, Figure S16a. In the MgB-containing chain D, MgB binds to the NSE triad, Figure S16b. In more recent work<sup>46</sup>, the structure of CotB2 (from *Streptomyces melanosporofaciens*) with a bound GGSP ligand plus one Mg<sup>2+</sup>, MgB, was reported<sup>46</sup>, and in this structure, the GGSP ligand adopted a highly bent “S”-shaped conformation (PDB ID code 5GUE), but was more closed, due presumably to use of co-crystallization, Figure S16c-e. The local structures and MgB locations were nevertheless remarkably similar to each other, and include the same H-bond network as discussed above with the conserved D-



**Figure 10.** Comparisons between two  $\alpha$ HH prenyltransferases, HsSQS and EhCrtM, with the  $\alpha$ C cyclase SaSES2. (a) Surface view of FSPP (cyan) bound to HsSQS (PDB ID 5IYS) showing both ligands bound to an open pocket. (b) Protein-ligand interactions in HsSQS-FSPP. The FSPP molecules are shown in cyan, the pink amino acids are the two Tyr, one of which H-bonds to Asp-116. (c) Surface view of PSPP (gray) bound to EhCrtM (PDB ID code 3WEI) showing PSPP binding to an open pocket. (d) Protein-ligand interactions in EhCrtM-PSPP. Tyr-171 H-bonds to Asp-80. The two aromatic residues (Tyr-171, Phe-230) shown in pink occupy the same positions as Tyr-131 and Tyr-184 in HsSQS. Most D-domain residues found in the  $\alpha$ HT prenyltransferases and  $\alpha$ C cyclases are absent in HsSQS and EhCrtM. (e) Overlay of FSPP (cyan) ligands in HsSQS with FPP (green) bound to SaSES2. The FPP headgroup in SaSES2 is bound in a similar manner to that of the S1 FSPP in HsSQS. (f) Overlay of PSPP (gray) ligand in EhCrtM with FPP (green) in SaSES2. The FPP in SaSES2 occupies a position that is similar to that seen with the tail of the S1-derived farnesyl side-chain in PSPP, in addition to binding to the MgABC cluster.

domain Asp and Tyr (Figure S16c-e), including an Asp bound to the conserved Arg, Figure S16c-e, although whether this single MgB binding motif is related to the presence of 1,10- and 1,11- cyclizations remains to be determined.

**A Comparison Between the  $\alpha$ C Cyclases,  $\alpha$ HH and  $\alpha$ HT Prenyltransferases:** Finally, we investigate the question: How similar are the structures of the  $\alpha$ C,  $\alpha$ HH and  $\alpha$ HT proteins? The  $\alpha$ HH prenyltransferases are involved in very complex condensation, cyclization/ring-opening and in some cases, reduction reactions. For example, in CrtM, two FPP molecules initially react to form presqualene diphosphate (PSPP), Figure 1b. One of these FPP ionizes (in the so-called S1 site) to form a farnesyl cation which is then attacked by the C2-C3 double bond in the second FPP (in the S2 site) to form PSPP with release of PPi. The PSPP diphosphate is thus initially located in the S2 site, then moves back to the S1 site, ionizes, the cyclopropyl cation rearranges and eliminates  $H^+$  to form the product, dehydrosqualene. In squalene synthase, the carbocation formed in the second “half-reaction” is reduced by NADPH (or NADH) to form the product, squalene. In hopanoid biosynthesis, HpnD forms PSPP which is then used as the substrate for HpnC, resulting in formation of 12-OH squalene, which is then reduced by a separate reductase to form squalene, the substrate for squalene-hopane cyclase. CrtM, SQS and HpnD are thus in a sense cyclases because they form a cyclic compound, while HpnC is not, although it is structurally similar to CrtM and SQS.

In early work, we and others reported numerous CrtM and SQS structures<sup>6-8</sup> plus, an HpnC structure has been deposited (PDB ID code 4HD1). All exhibit highly  $\alpha$ -helical folds and while in some cases three  $Mg^{2+}$  were observed, they were *not* the canonical MgABC seen in the  $\alpha$ C and  $\alpha$ HT proteins—the ones required for ionization. Rather, there was a  $Mg^{2+}$  bound to the S2 site FPP or PSPP diphosphate group. Also, the most highly conserved residues were Tyr and Gln, as shown in Table 1. More recently, CrtM (PDB ID code 5IYS)<sup>50</sup> and SQS (PDB ID code 3WEI)<sup>51</sup> structures have been reported that do in fact contain the anticipated MgABC (or MnABC) motifs. The structures are of *Enterococcus hirae* CrtM (EhCrtM) with two FSPP and 3  $Mg^{2+}$  bound,<sup>50</sup> and of an HsSQS Y73A mutant with PSPP and 3  $Mn^{2+}$  bound<sup>51</sup>, structures that represent the substrates involved in the 1<sup>st</sup> and 2<sup>nd</sup> half-reactions.

In dehydrosqualene synthase, two FSPP molecules bind to an open structure, Figure 10a, but only one binds to the MgABC cluster where it interacts with all 3  $Mg^{2+}$ , Figure 10b. This is the ionization site, S1. The diphosphate group in the second FSPP is not involved in binding to  $Mg^{2+}$ . The human squalene synthase structure (with PSPP and 3  $Mn^{2+}$ ) is also an open one, Figure 10c, and the same “MgABC” site (but with  $Mn^{2+}$ , which is also catalytically active) is involved in binding of PSPP to SQS in the second half-reaction, the ring-opening rearrangement (and reduction). The highly conserved Asp seen in the D-domain in the  $\alpha$ C and  $\alpha$ HT proteins (part of the H-bond network) is absent in both EhCrtM and HsSQS. This Asp is also absent in e.g. SaCrtM, *Alicyclobacillus acidocaldarius*

12-hydroxysqualene synthase (HpnC), *Neisseria meningitidis* PSPP (HpnD) and *Arabidopsis thaliana* phytoene synthase<sup>38</sup> as is the conserved Tyr (or His) seen in the  $\alpha$ C and  $\alpha$ HT proteins that forms part of the D-domain H-bond network, as well as the residues that H-bond to the conserved Gln (Gln-165 in SaCrtM, Table 1). The D-domain H-bond network is thus largely absent in the  $\alpha$ HH prenyltransferases.

Based on the structural and SCORECONS sequence results, it is apparent that the most highly conserved residue in the  $\alpha$ HH synthases is a tyrosine (Tyr-129 in SaCrtM, Table 1) and this residue is found in all other  $\alpha$ HH prenyltransferases.<sup>38</sup> In HsSQS, this Tyr (Tyr-171) is involved in a 2.8 Å H-bond interaction with Glu-83, which interacts with MgC, Figure 10d, and is highly conserved in alignments based on HsSQS (Figure S17), but is not present in CrtM, HpnC, HpnD or PHYS.<sup>38</sup> Since this Tyr is the most highly conserved residue in CrtM (and is essential in both CrtM and SQS), it seemed plausible that there might be an alternate residue involved in H-bonding to Tyr-129 in CrtM (and the other  $\alpha$ HH proteins). On inspection of the EhCrtM structure (PDB ID code 5IYS), it can be seen that the corresponding Tyr-131 interacts (d=2.9 Å) with Asp-116 Figure 10b, which is H-bonded to an H<sub>2</sub>O bound MgC. In an alignment based on EhCrtM (Figure S18), it can be seen that this Asp is very highly conserved (and corresponds to the Asp-114 seen in SaCrtM, Table 1) and it is also found in HpnC, HpnD and PHYS.<sup>50</sup> We therefore conclude that there are highly conserved Tyr-Glu or Tyr-Asp motifs in the HH prenyltransferases that play a role in stabilizing the MgAC domain, and it is known that Y129 (SaCrtM) or Y171 (HsSQS) are essential for catalytic activity in both the first- and second-half reactions and propose to call this the D\*-domain. As we suggested previously,<sup>38</sup> this domain may also be involved in H<sup>+</sup>-elimination in the  $\alpha$ HH prenyltransferases since this Tyr is partly solvent exposed in both CrtM and SQS.<sup>38</sup> The question to answer is then: how are the farnesyl chains located in the  $\alpha$ HH proteins *versus* their positions in the cyclases as well as in  $\alpha$ HT prenyltransferase such as GGPPS and FPPS?

When the SaSES complex structures are compared with the  $\alpha$ HH protein structures, the superimpositions shown in Figures 10e,f show that the FPP headgroup in SaSES2 binds in a similar manner to that seen with the S1 site FSPP in EhCrtM, as well as with that of PSPP bound to HsSQS, as expected based on the presence of the MgABC cluster. Its side-chain more closely overlays the S1-derived side-chain in PSPP than the S2-derived side-chain, and there is no similarity to the side-chain binding seen in the allylic (DMAPP, GPP) species that bind to the S1 site (**ab**) in FPPS or GGPPS,  $\alpha$ HT prenyltransferases—even though the AC, B and D domains are very similar. Thus, while there are some structural similarities between SaSES and the  $\alpha$ HH prenyltransferases, there is much closer similarity between the  $\alpha$ C cyclase SaSES with the  $\alpha$ HT prenyltransferases, although with the bound ligands, this similarity is only seen with the **ad** (product/inhibitor) binding site and not the **ab** substrate-binding site.

**Conclusions.** The results we have presented above are of interest for several reasons. By solving 8 structures of sesquiterpene and santalene synthases in apo forms, and with the sesquiterpenes, bound to FPP, FSPP, FpCCl<sub>2</sub>p or sabinene, we propose a sequential binding mechanism in which the (hydrophobic) farnesyl side-chain binds first, to the hydrophobic cavity, since there are not expected to be any Mg<sup>2+</sup> present.

Then, its diphosphate interacts with a conserved Arg; MgA,C then bind to the conserved DDXXD-like domain and to the diphosphate; MgB then also binds to the diphosphate and an N/D residue in the B-domain—all in the open conformation. Given the low concentration of free Mg<sup>2+</sup> in cells, the apo structures reflect the relevant forms present in cells, with Thr, Glu binding to MgB closing the pocket to form the catalytic conformation. Bond rotations (about C2-C3; C3-C4 and C5-C6) are required to convert the nerolidyl carbocation into a productive conformation for 1,6-condensation, the structure of sabinene being closely superimposable with that of the first prenyl group in the farnesyl side-chain. We also show that there are close structural similarities between apo (open), liganded (open) and liganded (closed) class I cyclase and (closed)  $\alpha$ HT prenyltransferase structures with a ~1.4 Å rmsd for the most highly conserved amino-acids *versus* a ~3.4 Å Ca rmsd for all atoms. For the  $\alpha$ HT prenyltransferases these residues are Thr, Gln, Asp, Tyr and for the  $\alpha$ C cyclases, Thr, Arg, Asp, Tyr. The close (local) structural similarity between the cyclases and  $\alpha$ HT prenyltransferases is due to the presence of the MgABC cluster, required for ionization, while the larger global rmsd can be attributed to the presence of a second, Lys/Arg-rich binding site pocket for IPP in the prenyltransferases which is absent in the cyclases. We propose that the (typically) Thr, Gln/Arg, Asp and Tyr residues are involved in stabilizing MgB via H<sub>2</sub>O, and in H<sup>+</sup>-elimination via the “proton wire” H-bond network seen in the neutron structure of IPP+protonated risedronate (a carbonium ion-diphosphate analog) bound to FPPS. There are, therefore, a third set of residues in the  $\alpha$ C cyclases and in the  $\alpha$ HT prenyltransferases—the D-domain—that are in fact the most highly conserved residues and are essential for catalysis. The observation here of H-bond networks in highly conserved residues that are involved in catalysis in  $\alpha$ HT prenyltransferases is reminiscent of that seen in head-to-middle and *cis*-head-to-tail prenyltransferases in which neutral species (Asn, Gln, Ser, Tyr), not bonded to Mg<sup>2+</sup>, again play a key role in catalysis. We also show that while  $\alpha$ HH prenyltransferases such as dehydrosqualene synthase and squalene synthase have obvious ( $\alpha$ -helical) structural similarities to  $\alpha$ C and  $\alpha$ HT structures, there are several major differences. First, the D-domain Thr, Gln/Arg, Asp and Tyr H-bond network is missing and only the Gln is present. Second, the most highly conserved residue is a Tyr that H-bonds to a conserved Asp or Glu that is bonded to MgC. This Tyr is found in CrtM, SQS, HpnC, HpnD and PHYS. Third, we show that in the  $\alpha$ C cyclases, substrate side-chains bind to a site that is similar to the **ad** pocket first identified in the structure of GGPP bound to GGPPS (the so-called product or inhibitor site), an  $\alpha$ HT prenyltransferase, not to the **ab** domain—the substrate pocket seen in  $\alpha$ HT prenyltransferases such as FPPS, GGPPS and OPPS where in these proteins, product chain length is regulated by steric interactions with the protein in the **ab** site region. In the  $\alpha$ HH prenyltransferases, the same **ad** domain is occupied by e.g. the farnesyl side-chain in FPP (or FSPP) or by the (S1-derived) PSPP side-chain, reflecting perhaps, the cyclase nature of the first half-reaction in proteins such as CrtM and SQS. However, in many  $\alpha$ C cyclases, this pocket is truncated or contoured to enable substrates to bind in “bent” configurations, facilitating cyclization, although in SaSES, the farnesyl groups bind in an (initially) extended form. Finally, we report the discovery that there are numerous highly conserved residues in the  $\beta$ -domain (that interfaces



with the catalytic  $\alpha$ -domain) in SaSES and many other  $\alpha\beta$  cyclases that play a more structural or regulatory role, similar to that seen with the small subunit in the heterodimeric long-chain  $\alpha$ HT prenyltransferases.

## ASSOCIATED CONTENT

### Supporting Information

The supporting information is available free of charge on the ACS Publications website at doi: xxxx.

Supporting Information file containing Figures, Tables, protein expression, purification, crystallization and structure determination, and diphosphate release assay for enzyme kinetics.

## AUTHOR INFORMATION

### Corresponding Author

E-mail: guoreyting@gmail.com (R-T. G.)

E-mail: eoldfield@illinois.edu (E. O.)

## ORCID

Rey-Ting Guo: 0000-0002-2779-7115

Eric Oldfield: 0000-0002-0996-7352

Satish Malwal: 0000-0001-7606-1932

### Author Contributions

# These authors contributed equally.

## ACKNOWLEDGMENT

This work has been supported by the National Natural Science Foundation of China (grants 31870790, and 31570130); KFZD-SW-215-01 from CAS; CAS Interdisciplinary Innovation Team; Taiwan Young Visiting Scholar Funding, CAS and by the United States Public Health Service (National Institutes of Health grants GM065307 and CA158191). Synchrotron data was obtained at beam lines TPS-05A, BL13C1 and BL15A1 of the NSRRC (National Synchrotron Radiation Research Center, Taiwan).

## REFERENCES

- Christianson, D. W., Structural biology and chemistry of the terpenoid cyclases. *Chem. Rev.* 2006, 106 (8), 3412-3442.
- Christianson, D. W., Structural and chemical biology of terpenoid cyclases. *Chem. Rev.* 2017, 117 (17), 11570-11648.
- J., B.; Cooper, C. M.; Purchase, R., *Natural Products Desk Reference*; CRC Press, Taylor & Francis Group: Boca Raton. 2016, p 235.
- Oldfield, E.; Lin, F. Y., Terpene biosynthesis: modularity rules. *Angew Chem. Int. Ed. Engl.* 2012, 51 (5), 1124-1137.
- Cao, R.; Zhang, Y.; Mann, F. M.; Huang, C.; Mukkamala, D.; Hudock, M. P.; Mead, M. E.; Pristic, S.; Wang, K.; Lin, F. Y.; Chang, T. K.; Peters, R. J.; Oldfield, E., Diterpene cyclases and the nature of the isoprene fold. *Proteins* 2010, 78 (11), 2417-2432.
- Liu, C. I.; Liu, G. Y.; Song, Y.; Yin, F.; Hensler, M. E.; Jeng, W. Y.; Nizet, V.; Wang, A. H.; Oldfield, E., A cholesterol biosynthesis inhibitor blocks *Staphylococcus aureus* virulence. *Science* 2008, 319 (5868), 1391-1394.
- Lin, F. Y.; Liu, C. I.; Liu, Y. L.; Zhang, Y.; Wang, K.; Jeng, W. Y.; Ko, T. P.; Cao, R.; Wang, A. H.; Oldfield, E., Mechanism of action

and inhibition of dehydrosqualene synthase. *Proc. Natl. Acad. Sci. U. S. A.* 2010, 107 (50), 21337-21342.

- Pandit, J.; Danley, D. E.; Schulte, G. K.; Mazzalupo, S.; Pauly, T. A.; Hayward, C. M.; Hamaoka, E. S.; Thompson, J. F.; Harwood, H. J., Jr., Crystal structure of human squalene synthase. A key enzyme in cholesterol biosynthesis. *J. Biol. Chem.* 2000, 275 (39), 30610-30617.
- Srivastava, P. L.; Daramwar, P. P.; Krithika, R.; Pandreka, A.; Shankar, S. S.; Thulasiram, H. V., Functional characterization of novel sesquiterpene synthases from Indian sandalwood, *Santalum album*. *Sci. Rep.* 2015, 5, 10095.
- Wendt, K. U.; Poralla, K.; Schulz, G. E., Structure and function of a squalene cyclase. *Science* 1997, 277 (5333), 1811-1815.
- Tarshis, L. C.; Yan, M.; Poulter, C. D.; Sacchettini, J. C., Crystal structure of recombinant farnesyl diphosphate synthase at 2.6-Å resolution. *Biochemistry* 1994, 33 (36), 10871-10877.
- Baer, P.; Rabe, P.; Fischer, K.; Citron, C. A.; Klapschinski, T. A.; Groll, M.; Dickschat, J. S., Induced-fit mechanism in class I terpene cyclases. *Angew Chem. Int. Ed. Engl.* 2014, 53 (29), 7652-7656.
- Trapp, S. C.; Croteau, R. B., Genomic organization of plant terpene synthases and molecular evolutionary implications. *Genetics* 2001, 158 (2), 811-832.
- Koksal, M.; Jin, Y.; Coates, R. M.; Croteau, R.; Christianson, D. W., Taxadiene synthase structure and evolution of modular architecture in terpene biosynthesis. *Nature* 2011, 469 (7328), 116-120.
- Koksal, M.; Hu, H.; Coates, R. M.; Peters, R. J.; Christianson, D. W., Structure and mechanism of the diterpene cyclase ent-copalyl diphosphate synthase. *Nat. Chem. Biol.* 2011, 7 (7), 431-413.
- Zhou, K.; Gao, Y.; Hoy, J. A.; Mann, F. M.; Honzatko, R. B.; Peters, R. J., Insights into diterpene cyclization from structure of bifunctional abietadiene synthase from *Abies grandis*. *J. Biol. Chem.* 2012, 287 (9), 6840-6850.
- Koksal, M.; Zimmer, I.; Schnitzler, J. P.; Christianson, D. W., Structure of isoprene synthase illuminates the chemical mechanism of teragram atmospheric carbon emission. *J. Mol. Biol.* 2010, 402 (2), 363-373.
- Whittington, D. A.; Wise, M. L.; Urbansky, M.; Coates, R. M.; Croteau, R. B.; Christianson, D. W., Bornyl diphosphate synthase: structure and strategy for carbocation manipulation by a terpenoid cyclase. *Proc. Natl. Acad. Sci. U. S. A.* 2002, 99 (24), 15375-15380.
- Krisinell, E.; Henrick, K., Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr.* 2004, D60 (12), 2256-2268.
- Blank, P. N.; Shinsky, S. A.; Christianson, D. W., Structure of sesquisabinene synthase 1, a terpenoid cyclase that generates a strained [3.1.0] bridged-bicyclic product. *ACS Chem. Biol.* 2019, 14 (5), 1011-1019.
- Romani, A., Regulation of magnesium homeostasis and transport in mammalian cells. *Arch. Biochem. Biophys.* 2007, 458 (1), 90-102.
- Drozdzetskiy, A.; Cole, C.; Procter, J.; Barton, G. J., JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* 2015, 43, W389-W394.
- Auffinger, P.; Hays, F. A.; Westhof, E.; Ho, P. S., Halogen bonds in biological molecules. *Proc. Natl. Acad. Sci. U. S. A.* 2004, 101 (48), 16789-16794.
- Tarshis, L. C.; Proteau, P. J.; Kellogg, B. A.; Sacchettini, J. C.; Poulter, C. D., Regulation of product chain length by isoprenyl diphosphate synthases. *Proc. Natl. Acad. Sci. U. S. A.* 1996, 93 (26), 15018-15023.
- C, K. M. C.; Hudock, M. P.; Zhang, Y.; Guo, R. T.; Cao, R.; No, J. H.; Liang, P. H.; Ko, T. P.; Chang, T. H.; Chang, S. C.; Song, Y.; Axelson, J.; Kumar, A.; Wang, A. H.; Oldfield, E., Inhibition of geranylgeranyl diphosphate synthase by bisphosphonates: a crystallographic and computational investigation. *J. Med. Chem.* 2008, 51 (18), 5594-5607.
- Gatto, N.; Vattekkatte, A.; Köllner, T.; Degenhardt, J.; Gershenzon, J.; Boland, W., Isotope sensitive branching and kinetic isotope effects to analyse multiproduct terpenoid synthases from *Zea mays*. *Chem. Commun.* 2015, 51 (18), 3797-3800.
- Hosfield, D. J.; Zhang, Y.; Dougan, D. R.; Broun, A.; Tari, L. W.; Swanson, R. V.; Finn, J., Structural basis for bisphosphonate-mediated inhibition of isoprenoid biosynthesis. *J. Biol. Chem.* 2004, 279 (10), 8526-8529.
- Yokoyama, T.; Mizuguchi, M.; Ostermann, A.; Kusaka, K.; Niimura, N.; Schrader, T. E.; Tanaka, I., Protonation state and hydration

- of bisphosphonate bound to farnesyl pyrophosphate synthase. *J. Med. Chem.* 2015, 58 (18), 7549-7556.
29. Little, D. B.; Croteau, R. B., Alteration of product formation by directed mutagenesis and truncation of the multiple-product sesquiterpene synthases  $\delta$ -selinene synthase and  $\gamma$ -humulene synthase. *Arch. Biochem. Biophys.* 2002, 402 (1), 120-135.
30. Gennadios, H. A.; Gonzalez, V.; Di Costanzo, L.; Li, A.; Yu, F.; Miller, D. J.; Allemann, R. K.; Christianson, D. W., Crystal structure of (+)- $\delta$ -cadinene synthase from *Gossypium arboreum* and evolutionary divergence of metal binding motifs for catalysis. *Biochemistry* 2009, 48 (26), 6175-6183.
31. Saito, A.; Rilling, H. C., The formation of cyclic sesquiterpenes from farnesyl pyrophosphate by prenyltransferase. *Arch. Biochem. Biophys.* 1981, 208 (2), 508-811.
32. Valdar, W. S., Scoring residue conservation. *Proteins* 2002, 48 (2), 227-241.
33. Liu, Y. L.; Lindert, S.; Zhu, W.; Wang, K.; McCammon, J. A.; Oldfield, E., Taxodione and arenarone inhibit farnesyl diphosphate synthase by binding to the isopentenyl diphosphate site. *Proc. Natl. Acad. Sci. U. S. A.* 2014, 111 (25), E2530-2539.
34. Desai, J.; Liu, Y. L.; Wei, H.; Liu, W.; Ko, T. P.; Guo, R. T.; Oldfield, E., Structure, function, and inhibition of *Staphylococcus aureus* Heptaprenyl Diphosphate Synthase. *ChemMedChem* 2016, 11 (17), 1915-23.
35. Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *J. Mol. Graph.* 1996, 14 (1), 33-8, 27-8.
36. Chang, K. M.; Chen, S. H.; Kuo, C. J.; Chang, C. K.; Guo, R. T.; Yang, J. M.; Liang, P. H., Roles of amino acids in the *Escherichia coli* octaprenyl diphosphate synthase active site probed by structure-guided site-directed mutagenesis. *Biochemistry* 2012, 51 (16), 3412-3419.
37. Koyama, T.; Tajima, M.; Nishino, T.; Ogura, K., Significance of Phe-220 and Gln-221 in the catalytic mechanism of farnesyl diphosphate synthase of *Bacillus stearothermophilus*. *Biochem. Biophys. Res. Commun.* 1995, 212 (2), 681-686.
38. Malwal, S. R.; Gao, J.; Hu, X.; Yang, Y.; Liu, W.; Huang, J.-W.; Ko, T.-P.; Li, L.; Chen, C.-C.; O'Dowd, B.; Khade, R. L.; Zhang, Y.; Zhang, Y.; Oldfield, E.; Guo, R.-T., Catalytic Role of conserved asparagine, glutamine, serine, and tyrosine residues in isoprenoid biosynthesis enzymes. *ACS Catal.* 2018, 8 (5), 4299-4312.
39. Gao, J.; Ko, T. P.; Chen, L.; Malwal, S. R.; Zhang, J.; Hu, X.; Qu, F.; Liu, W.; Huang, J. W.; Cheng, Y. S.; Chen, C. C.; Yang, Y.; Zhang, Y.; Oldfield, E.; Guo, R. T., "Head-to-middle" and "head-to-tail" cis-prenyl transferases: structure of isoesquilavandulyl diphosphate synthase. *Angew. Chem. Int. Ed. Engl.* 2018, 57 (3), 683-687.
40. Martin, M. B.; Arnold, W.; Heath, H. T., 3rd; Urbina, J. A.; Oldfield, E., Nitrogen-containing bisphosphonates as carbocation transition state analogs for isoprenoid biosynthesis. *Biochem. Biophys. Res. Commun.* 1999, 263 (3), 754-758.
41. Koyama, T.; Tajima, M.; Sano, H.; Doi, T.; Koike-Takeshita, A.; Obata, S.; Nishino, T.; Ogura, K., Identification of significant residues in the substrate binding site of *Bacillus stearothermophilus* farnesyl diphosphate synthase. *Biochemistry* 1996, 35 (29), 9533-9538.
42. Fujisaki, S.; Hara, H.; Nishimura, Y.; Horiuchi, K.; Nishino, T., Cloning and nucleotide sequence of the ispA gene responsible for farnesyl diphosphate synthase activity in *Escherichia coli*. *J. Biochem.* 1990, 108 (6), 995-1000.
43. Croteau, R. B.; Colby, S. M., DNA encoding limonene synthase from *Mentha spicata*. *US Patent 5871988* 1999.
44. Guillen Schlippe, Y. V.; Hedstrom, L., A twisted base? The role of arginine in enzyme-catalyzed proton abstractions. *Arch. Biochem. Biophys.* 2005, 433 (1), 266-278.
45. Shishova, E. Y.; Yu, F.; Miller, D. J.; Faraldos, J. A.; Zhao, Y.; Coates, R. M.; Allemann, R. K.; Cane, D. E.; Christianson, D. W., X-ray crystallographic studies of substrate binding to aristolochene synthase suggest a metal ion binding sequence for catalysis. *J. Biol. Chem.* 2008, 283 (22), 15431-15439.
46. Tomita, T.; Kim, S. Y.; Teramoto, K.; Meguro, A.; Ozaki, T.; Yoshida, A.; Motoyoshi, Y.; Mori, N.; Ishigami, K.; Watanabe, H.; Nishiyama, M.; Kuzuyama, T., Structural insights into the cotb2-catalyzed cyclization of geranylgeranyl diphosphate to the diterpene cyclooctat-9-en-7-ol. *ACS Chem. Biol.* 2017, 12 (6), 1621-1628.
47. Guo, R. T.; Kuo, C. J.; Chou, C. C.; Ko, T. P.; Shr, H. L.; Liang, P. H.; Wang, A. H., Crystal structure of octaprenyl pyrophosphate synthase from hyperthermophilic *Thermotoga maritima* and mechanism of product chain length determination. *J. Biol. Chem.* 2004, 279 (6), 4903-4912.
48. Kavanagh, K. L.; Dunford, J. E.; Bunkoczi, G.; Russell, R. G.; Oppermann, U., The crystal structure of human geranylgeranyl pyrophosphate synthase reveals a novel hexameric arrangement and inhibitory product binding. *J. Biol. Chem.* 2006, 281 (31), 22004-22012.
49. Gritzalis, D.; Park, J.; Chiu, W.; Cho, H.; Lin, Y.-S.; De Schutter, J. W.; Lacbay, C. M.; Zielinski, M.; Berghuis, A. M.; Tsantrizos, Y. S., Probing the molecular and structural elements of ligands binding to the active site versus an allosteric pocket of the human farnesyl pyrophosphate synthase. *Bioorg. Med. Chem. Lett.* 2015, 25 (5), 1117-1123.
50. Schwalen, C. J.; Feng, X.; Liu, W.; B. O. D.; Ko, T. P.; Shin, C. J.; Guo, R. T.; Mitchell, D. A.; Oldfield, E., Head-to-head prenyl synthases in pathogenic bacteria. *Chembiochem* 2017, 18 (11), 985-991.
51. Liu, C. I.; Jeng, W. Y.; Chang, W. J.; Shih, M. F.; Ko, T. P.; Wang, A. H., Structural insights into the catalytic mechanism of human squalene synthase. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* 2014, 70, 231-241.

---

Solving the structures of eight terpene cyclases leads to the discovery of remarkable structural and functional similarities between the active site regions of class I cyclases and head-to-tail *trans*-prenyl transferases, such as farnesyl diphosphate synthase.

