Structural studies on Sesquisabinene Synthase 1: Enzyme involved in Terpene Biosynthesis Pathway

by

Sneha Singh

10BB14A26043

A thesis submitted to the

Academy of Scientific & Innovative Research
for the award of the degree of
DOCTOR OF PHILOSOPHY

in

SCIENCE

Under the supervision of

Dr. Kiran Kulkarni

(Supervisor)

Dr. H.V. Thulasiram

(Co-Supervisor)

CSIR-National Chemical Laboratory, Pune

Academy of Scientific and Innovative Research AcSIR
Headquarters, CSIR-HRDC campus

Sector 19, Kamla Nehru Nagar,
Ghaziabad, U.P. – 201 002, India

May, 2021

Certificate

This is to certify that the work incorporated in this Ph.D. thesis entitled, "*Structural studies on Sesquisabinene Synthase 1: Enzyme involved in Terpene Biosynthesis Pathway",* submitted by *Sneha Singh* to the Academy of Scientific and Innovative Research (AcSIR) in fulfillment of the requirements for the award of the Degree of *Doctor of Philosophy in Science*, embodies original research work carried-out by the student. We, further certify that this work has not been submitted to any other University or Institution in part or full for the award of any degree or diploma. Research material(s) obtained from other source(s) and used in this research work has/have been duly acknowledged in the thesis. Image(s), illustration(s), figure(s), table(s) etc., used in the thesis from other source(s), have also been duly cited and acknowledged.

Sneha Singh (Student) Date: 13.05.2021	Dr. H.V. Thulasiram (Research Co-Supervisor) Date: 13.05.2021	Dr. Kiran Kulkarni (Research Supervisor) Date: 13.05.2021

STATEMENTS OF ACADEMIC INTEGRITY

I Sneha Singh, a Ph.D. student of the Academy of Scientific and Innovative Research (AcSIR) with Registration No. 10BB14A26043 hereby undertake that, the thesis entitled "Structural studies on Sesquisabinene Synthase 1: Enzyme involved in Terpene Biosynthesis Pathway" has been prepared by me and that the document reports original work carried out by me and is free of any plagiarism in compliance with the UGC Regulations on *"Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions (2018)"* and the CSIR Guidelines for "*Ethics in Research and in Governance (2020)".*

Sneha

Signature of the Student
Date : 13.05.2021
Place : Pune

It is hereby certified that the work done by the student, under our supervision, is plagiarism-free in accordance with the UGC Regulations on *"Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions (2018)"* and the CSIR Guidelines for "*Ethics in Research and in Governance (2020)".*

Thulasiram H.V.

Signature of the Co-Supervisor
Name : Dr. H.V. Thulasiram
Date : 13.05.2021
Place : Pune

Signature of the Supervisor
Name : Dr. Kiran Kulkarni
Date : 13.05.2021
Place : Pune

Dedicated to……

Mummy, Papa and my lovely Sis who were always there for me and encouraged me throughout this whole journey

Acknowledgement

First and foremost, I thank the Almighty for this beautiful opportunity of pursuing PhD. It was a dream for me and You made it happen. Without Your grace, I am nothing and with Your blessings I have everything I wish for.

The journey of my PhD has helped me to grow as a researcher and overcome my fears to face various challenges. My long tenure as a research scholar is enriched by the help of a number of people. Here I would like to take this opportunity to thank the people who have in no small way made the research in this thesis possible.

It is my great privilege to extend my heartfelt and sincere gratitude to my guide, Dr. Kiran Kulkarni, who has supported me throughout my journey with his patience, knowledge and belief in me. He has taught me very interesting and difficult aspects of Structural Biology. I am really grateful for his constant encouragement, guidance and efforts. His enthusiasm for pursuing daunting problems at the highest levels of scientific integrity has been a constant source of inspiration, excitement, advice and guidance throughout my study. His words of motivation have always helped me to overcome the tough times in my PhD pursuit. I hope that during these years I have been able to absorb even a tiny bit of his enthusiasm and passion towards the field of Science.

My sincere thanks to my co-guide, Dr. H.V. Thulasiram, who introduced me to this work of Terpenes by allowing me to study the structural aspect of the project. He offered me complete support and helped me in overcoming the research problems related to the biochemical aspects of the project. I was lucky to have him as my co-supervisor due to which I was able to learn new techniques and gain knowledge of a completely different field.

I place on record my heartfelt gratitude to the members of my Doctoral Advisory Committee, Dr. Jomon Joseph, Dr. Mahesh Kulkarni and Dr. Nandini Devi for their continuous support, guidance and suggestions that helped me to widen my research from various perspectives.

I am also thankful to Council of Scientific and Industrial Research for providing me

scholarship, Dr. Ashish Lele (Director, NCL) and Dr. Narendra Kadoo (Head, Biochemical Sciences Division) for allowing me to carry out my research and providing all necessary infrastructure and facilities.

I would also like to acknowledge Dr. H.V. Thulasiram for providing reagents and GC-MS facility for my work. My sincere thanks to Dr. Prabhakar Srivastava who initiated work of this project of sesquiterpene synthases and helped me initially to understand the project. I am also grateful to Dr. Durba Sengupta for delivering inputs on MD simulations. I would also like to thank Ms. Aiswarya Pawar for her help with R scripts.

My special thanks to Dr. Ravindra Makde, Dr. Biplab Ghosh and Dr. Ashwani Kumar for their help during my experiments at RRCAT Indus II synchrotron, Indore. I am also thankful to Dr. Juha Huiskonen, University of Helsinki for hosting me in his lab as a visiting student and providing insights to the experiments related to cryo electron microscopy. I would like to acknowledge Dr. Gayathri Pananghat & Dr. Saikrishana Kayarat from IISER, Pune and Dr. Janesh Kumar & Dr. Radha Chauhan from NCCS, Pune for providing access to their crystallization set up facility, initially. I am also thankful to Dr. Somnath Dutta and Mr. Anil Kumar for their help in preparing cryo EM grids and data collection at IISC, Bangalore. I would also like to express my gratitude to Dr. Maurizio Polentarutti for diffraction data collection at XRD1 beamline facility, ELETTRA synchrotron.

I am also thankful to Dr. Dhanasekaran Shanmugam, Dr. Mahesh Kulkarni, Dr. Ashok Giri, Dr. Narendra Kadoo, Dr. Subhashchandrabose Chinnathambi and Dr. Koteswara Rao for allowing me to use their lab facilities and reagents.

I express my gratitude from the bottom of my heart to my senior, Dr. Anandsukeerthi. He has guided me by sharing his experiences and knowledge that helped me to troubleshoot various scientific problems. I am also grateful to the constant support and suggestions by my wonderful senior, Ashwini. I also take this opportunity to thank my labmates, Debopriya, Zenia, Sharmila, Aishwarya, Ravi, Ankita and Prateeksha for making the lab environment joyful and making it my second home. I cherish all the memories we share. I would also like to thank the past trainees and project assistants for their help and support.

I would also like to acknowledge Gopal, Nalini, Rajeshwari, Prachi,

Dr. Parag, Dr. Debjyoti, Dr. Amit, Shrikant, Amarnath, Abhishek, Sindhuri, Dr. Rahul, Dr. Meenakshi, Dr. Rupali, Dr. Anurag, Tejashri, Tushar, Abhishek, Madhura, Dr. Deepanjan, Dr. Yashpal, Dr. Amrita, Gauri, Shakuntala, Shiva, Yogendra, Babasaheb, Vaishnavi, Santosh, Dr. Amol, Sonal and others for their scientific inputs and help with the reagents and the facilities. I am also thankful to all the non-teaching staff members.

I am also heartily grateful to my wonderful friends, Gopal, Nalini, Rajeshwari, Prachi, Dr. Amit and Dr. Parag for all their support, care and making me feel at home. I can't thank you enough for all the support you have given me. A special thanks to Sangeeta, Dr. Ananth, Aarohi and Dr. Ezaz for their unconditional help.

I am fortunate to have been surrounded by a wonderful group of friends, Upasana, Pallavi, Abhilasha and Likith. They have cultivated my growth in embarking this exciting journey and always stood by me and supported me.

I would like to thank a very special person in my life, Shankar who has always been a pillar of support for me. He encouraged me to start and pursue this beautiful and learning experience and has always extended support during the toughest times of my tenure. Thank you for believing in me more than me, bearing me and providing your unconditional love and support.

I also wish to thank my Uncle, Ayodhya Tiwari, for his constant support in all types of situations, love and the belief he had in me. Last and not at all the least, a big thank you to my wall of strength and support, my parents, Usha Singh and Vijay Singh, and my sister, Urvashi. I can't thank you enough for all the support you all have given me. Thank you for all the sacrifices you made in order to make my PhD tenure smoother and life so beautiful. You all are my inspiration and encouragement. Without your support and belief in me, this day of submitting my thesis would not have been possible. Thank you for being supportive and patient.

- Sneha Singh

Education is not the
learning of facts
It's rather the training
of the Mind to
THINK

- Albert Einstein

Table of Contents

List of Figures

List of Tables

List of Schemes

Abbreviations

Å	Angstrom
AaBOS	α-bisabolol synthase
ADP	Atomic displacement parameter
bp	Base pair
sCCD	Charge-coupled device
CoA	Coenzyme A
C-terminal	Carboxyl terminal
DCCM	Dynamical cross-correlation matrix
dIC	Dynamics based independent component
DMAPP	Dimethylallyl diphosphate
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleoside triphosphate
dSCA	Dynamics based statistical coupling analysis
DTT	Dithiothreitol
dU	Deoxy uridine nucleotide
e/Å	Electron/Angstrom
EDTA	Ethylenediaminetetraacetic acid
FPP	Farnesyl pyrophosphate
GC-MS	Gas chromatography- Mass spectrometry
GGPP	Geranylgeranyl pyrophosphate
GPP	Geranyl pyrophosphate
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HI	Hydrophobicity index
His tag	Histidine tag
HPC	Hybrid photon counting
IC	Independent component
IPP	Isopentenyl diphosphate
IPTG	Isopropyl 1-thio-D-galactopyranoside
K	Kelvin
kDa	Kilodalton
LB	Luria Bertani
MD	Molecular dynamics
MEP	Methylerythritol phosphate
Mg	Magnesium
mg	Milligram
Min	Minute (s)
mL	Millilitre
mAU	Milli absorbance unit
mM	Millimolar
mm	Millimetre

MR	Molecular replacement
MSA	Multiple sequence alignment
MVA	Mevalonic acid
NaCl	Sodium chloride
ng	Nanogram
Ni-NTA	Nickel-nitrilotriacetic acid
NIST	National Institute of Standards and Technology
nL	Nanolitre
nm	Nanometre
ns	Nanosecond
N-terminal	Amino terminal
OD	Optical density
PCR	Polymerase chain reaction
PDB	Protein data bank
PMR	Product modulating residue
PMSF	Phenylmethylsulfonyl fluoride
RE	Restriction endonuclease
RMSD	Root mean square deviation
RNA	Ribonucleic acid
rpm	Revolutions per minute
S.album	*Santalum album*
SaBS	Bisabolene synthase
SaFDS	Farnesyl diphosphate synthase
SaSQS1	Sesquisabinene synthase 1
SaSS	Santalene synthase
SCA	Statistical coupling analysis
SDS	Sodium Dodecyl Sulfate
Sec	Second (s)
sIC	Sequence based independent component
sSCA	Sequence based statistical coupling analysis
SSQ	Sesquiterpene synthase
SUMO	Small Ubiquitin-like Modifier
TAE	Tris-acetate-EDTA
TEAS	5-epi-aristolochene synthase
TEV	Tobacco etch virus
U	Unit
USER	Uracil-Specific Excision Reagent
V/cm	Volt/centimetre
VI	Vicinity index
µg	Microgram
µL	Microlitre
µM	Micromolar

Chapter 1

Introduction

Secondary metabolites synthesized by plants are an essential assembly of natural products, which play an important role in various biological activities of plants, like defense system against biotic and abiotic stresses. These secondary metabolites are also largely used as food and aromatic additives, ingredient for medicines, attractants for pollinators and signaling molecules[1]. Plant secondary metabolites are categorized based on their chemical structure [1] and grouped as flavonoids and phenolic acids, alkaloids, steroids and terpenoids (Fig.1.1) [1,2]. These compounds are responsible for activation and augmentation of defense in plants and help the plants to adapt to varying and inconsistent surroundings. Flavonoids play crucial role in signaling, pathogenesis in plants and are derived from acetyl-CoA[3]. Alkaloids contain nitrogen and are found in 20% plants, which play important role in defense mechanism of plants against pathogens[4]. Isoprenoids are the largest class amongst these, which have various roles in plants as hormones, electron carriers, photosynthetic pigments along with defense and communication roles. Due to the utility of these metabolites in essential oils, cosmetics, chemicals, pharmaceuticals and in nutraceuticals, they have gained commercial importance[5].

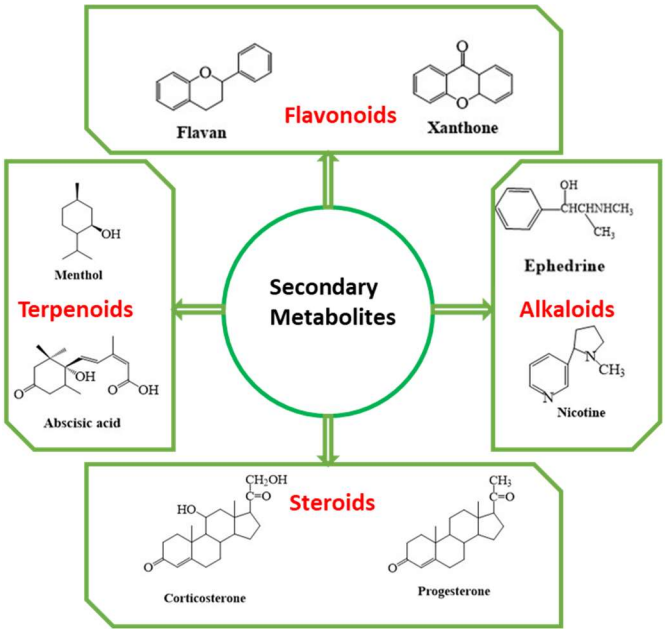


Fig.1.1. Different types of plant secondary metabolites. *Categories of secondary metabolites of plants with their few examples and chemical structures.*

The extraction and purification of secondary metabolites directly from plant sources has a lot of limitations like the process is tedious and time consuming and also the product yield is quite low [6]. Thus, overcoming these limitations to enhance the yield and strengthen the process pose challenge to the researchers. A number of studies have emphasized on searching of high yielding species and optimization of growth conditions of plants [7,8]. However, this process also requires a lot of time and resource investment and also is dependent on a number of environmental and other conditions.

1.1 Isoprenoids:

Terpenes or Isoprenoids are a diverse group of natural products found in plants, fungi and bacteria. In plants, these compounds play role as secondary metabolites like carotenoids and chlorophyll, quinones and also in plant defense mechanism and communication[9]. These secondary metabolites are also used in various pharmaceuticals like taxol, which is used as an anti-cancer drug[10], artemisinin which is an anti-malarial drug[11], etc. Terpenoids are a group of structurally diverse natural substances with approximately 80,000 compounds[12].

According to the isoprene rule by Rutzicka & Wallach, IPP (isopentenyl diphosphate) and DMAPP (dimethylallyl diphosphate) are the two universal precursors for all the terpenoids[13]. IPP and DMAPP are the five-carbon blocks, which after "head-to-tail" condensation form prenyl diphosphate intermediates, which leads to the biosynthesis of terpenes. The C_5 isoprenoid units undergo condensation to form linear chains of monoterpenes (C_{10}), sesquiterpenes (C_{15}), diterpenes (C_{20}), triterpenes (C_{30}), tetraterpenes (C_{40}) and polyterpenes ($>C_{40}$)[13], which further go through a series of cyclization, rearrangement and oxidation reactions.

For the production of IPP and DMAPP, plants have two independent pathways; MVA (mevalonic acid) pathway, which is located in cytosol and the MEP (methylerythritol phosphate) pathway, which is active in plastid (Fig.1.2). The precursors for biosynthesis of sesquiterpenes, phytosterols, polyprenols and triterpenoids are provided predominantly by the MVA pathway, while the MEP pathway is involved in the biosynthesis of isoprene, monoterpenes, diterpenes and carotenoids. The MVA pathway comprises of six steps, initiated from the condensation of two molecules of acetyl-CoA to subsequently form IPP, followed by maintenance of a balance between IPP and DMAPP by IPP isomerase. The MEP pathway on the other hand involves transformation of glyceraldehyde-3-phosphate and pyruvic acid to DMAPP and IPP (1:5). Prenyltranferases catalyze the conversion of IPP and DMAPP to longer

chain precursors like GPP (geranyl pyrophosphate, C_{10}), FPP (farnesyl pyrophosphate, C_{15}) and GGPP (geranylgeranyl pyrophosphate, C_{20}), which are precursors of monoterpenes, sesquiterpenes and diterpenes, respectively. In plants, monoterpenes and diterpenes are synthesized in the plastid, however, synthesis of sesquiterpenes and triterpenes takes place in the cytosol. Due to increase in the production of sustainable plant-based medicines and the surge in the application of terpenes in synthesis of alternative fuels, there is a remarkable progress in engineering of biosynthetic pathways involved in production of terpenes in microorganisms and plants[14]. Engineering of these metabolic pathways help to produce terpenoids in huge quantities avoiding the need for extraction of these compounds from their natural sources. Thus, the bioengineering method helps in overcoming the limitations of natural extraction which include low yield of these compounds, impurities and huge consumption of natural resources.

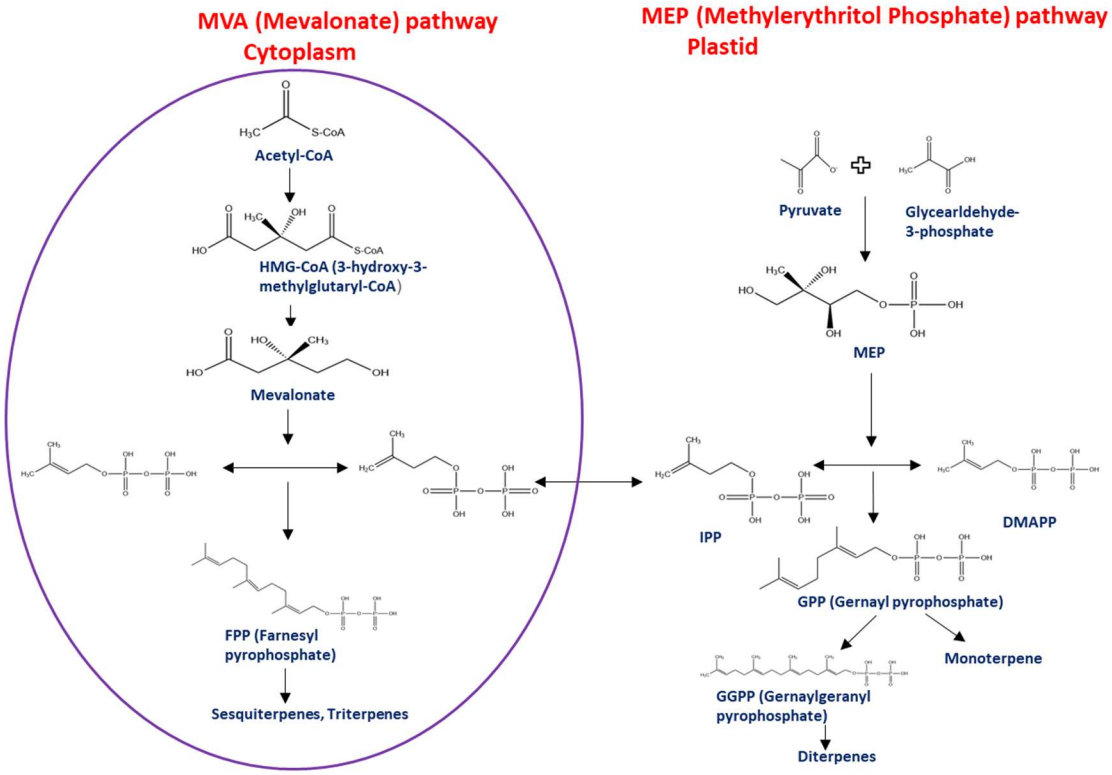


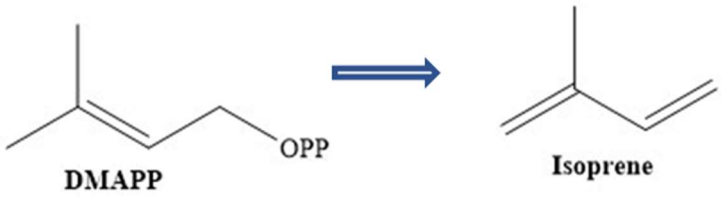
Fig.1.2. Overview of isoprenoid biosynthesis pathway. *The MVA and MEP pathways lead to the biosynthesis of universal precursors of terpenes, IPP (isopentenyl diphosphate) and DMAPP (dimethylallyl diphosphate). The MVA pathway, circled, forms FPP which cyclizes to form sesquiterpenes and squalene act as the precursor for triterpenes.*

1.1.1 Classification of Isoprenoids:

According to the number of isoprene units present, terpenes can be classified into the following groups:

1.1.1.1 Hemiterpene:

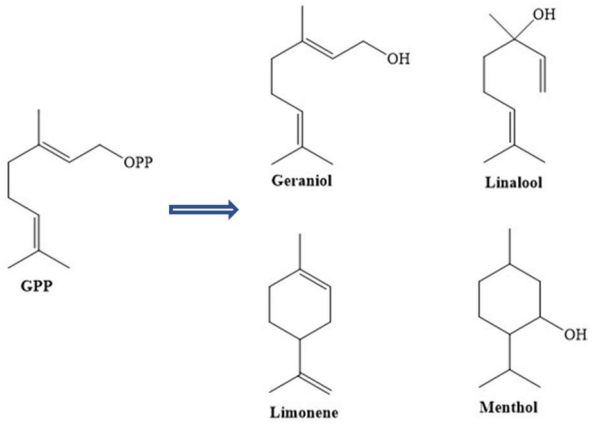
This group of isoprenoids consist of a single isoprene unit. Hemiterpenoids are their oxygenated derivatives. Isoprenes are produced in huge amounts from various plant species like ferns and mosses. DMAPP after diphosphate elimination forms isoprene[15]. Isoprene plays crucial role in protecting plants from abiotic stress, aids plant survival in rapid temperature changes and their emission from plants accounts for a remarkable ratio of atmospheric hydrocarbon[16].



Scheme 1. *Isoprene biosynthesis. Pyrophosphate group is represented as OPP.*

1.1.1.2 Monoterpene:

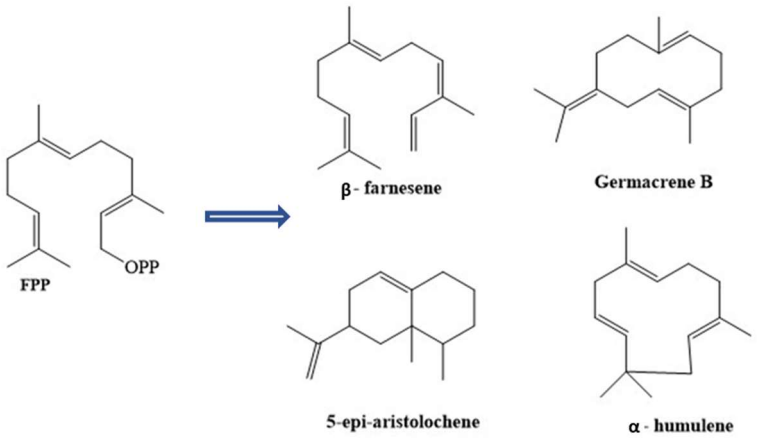
Monoterpenes are found in secretory tissues like oil glands of plants, fungi and insects. It comprises of two isoprene units ($C_{10}H_{16}$). They are volatile compounds and have wide applications in fragrance and flavor industry. They are also used in pharmaceutical sector as antioxidants, antibacterials, antifungals and anticancer agents[17] . Monoterpenes can be acyclic, monocyclic and bicyclic. Monoterpene synthases catalyze the conversion of GPP to monoterpenes. Although in plants they are synthesized in plastids through MEP pathway, in yeasts and other higher organisms they are synthesized through MVA pathway.



Scheme 2. Monoterpene biosynthesis. *Examples showing conversion of GPP to monoterpenes. Pyrophosphate group is represented as OPP.*

1.1.1.3 Sesquiterpene:

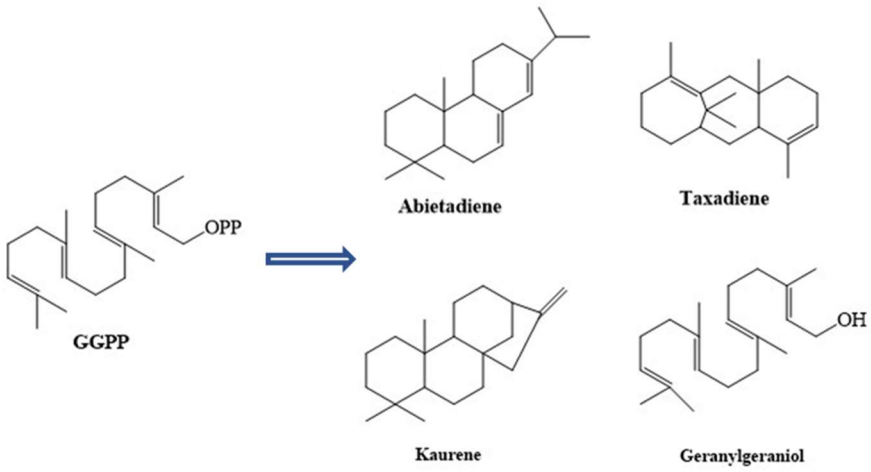
Sesquiterpenes consist of three isoprene units and constitutes the most diverse group of terpenes. The molecular formula for sesquiterpenes is $C_{15}H_{24}$. The group consists an array of 7000 molecules which form more than 300 stereochemical discrete hydrocarbon skeletons[18]. Linear substrate, FPP, is cyclized by Sesquiterpene Synthases to form sesquiterpenes. These are one of the most crucial components of essential oils from plants. The huge structural diversity of sesquiterpenes is due to existence of three double bonds and long carbon chain of FPP. Sesquiterpenes can be acyclic, monocyclic, bicyclic and tricyclic. Sesquiterpenes are synthesized in cytosol through MVA pathway. They are used as flavor and fragrance agents and they have potential to act as substitute for petroleum derived fuels[19]. These compounds have a possible role as an anticancer and antimalarial agent [20,21].



Scheme 3. Sesquiterpene biosynthesis. *Examples showing conversion of FPP to sesquiterpenes. Pyrophosphate group is represented as OPP.*

1.1.1.4 Diterpene:

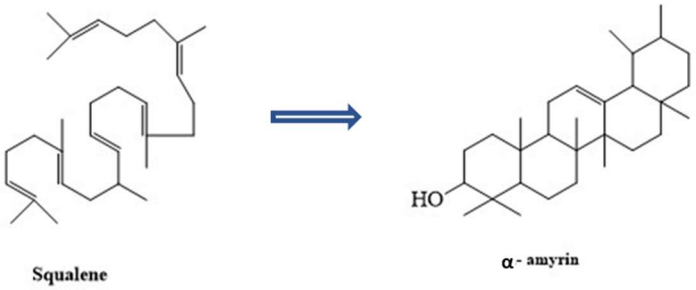
Diterpenes are synthesized from GGPP and consists of four isoprene units ($C_{20}H_{32}$). These compounds could be linear, bicyclic, tricyclic, tetracyclic and macrocyclic. Multiple pharmacological utilities, such as being anti-fungal, anti-bacterial, anti-inflammatory and anti-leishmanial have been suggested for diterpenes. Taxol is one such diterpene, which is wieldy used as an anti-cancer agent[10].



Scheme 4. Diterpene biosynthesis. *Examples showing conversion of GGPP to diterpenes. Pyrophosphate group is represented as OPP.*

1.1.1.5 Triterpene:

Triterpenes are synthesized from squalene and consist of six isoprene units with a molecular formula $C_{30}H_{48}$. These compounds are involved in regulation of permeability and fluidity of proteins and lipids[22]. Sterols are involved in synthesis of a number of compounds involved in crucial cellular processes in animals and in plants, which aid in synthesis of brassinosteroids. These are essential for normal growth and development of plants[23].

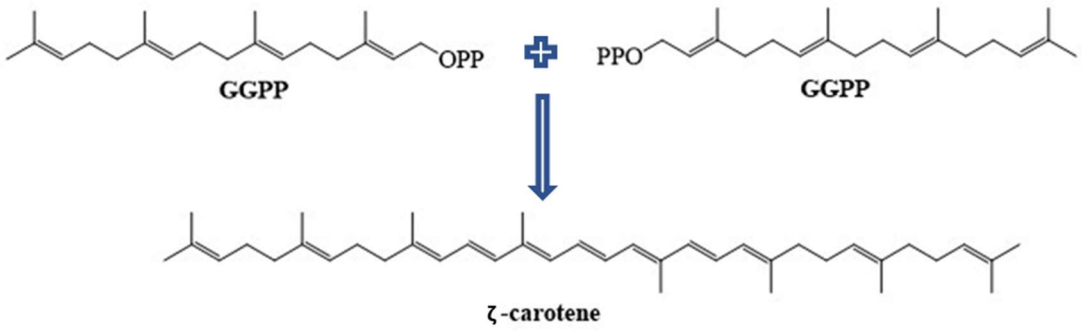


Scheme 5. Triterpene biosynthesis. *Examples showing conversion of squalene to triterpenes.*

1.1.1.6 Tetraterpene:

Tetraterpenes contain eight isoprene units ($C_{40}H_{56}$), which are found in plants, fungi and bacteria. Carotenoids are important members of this group of terpenes which are synthesized

in plants and bacteria through MEP pathway, however, in fungi they are synthesized through the MVA pathway[24]. Oxygenated and mono-oxygenated derivatives of carotenoids are known as xanthophylls and carotenes, respectively. Carotenoids help the photosynthetic organisms to absorb light and xanthophylls prevent chlorophyll bleaching by harvesting intense light[25]. Carotenoids possess antioxidant properties and ameliorate the severity of chronic diseases[26]. They have commercial importance as they are widely used as animal supplements, cosmetics and food colorants.



Scheme 6. Tetraterpene biosynthesis. *Examples showing reaction of two molecules of GGPP to form tetraterpene. Pyrophosphate group is represented as OPP.*

1.1.1.7 Polyterpene:

Multiple isoprene units form polyterpene, which has a molecular formula $(C_5H_8)_n$. Prenyltransferases catalyze the condensation of IPP groups on FPP successively to form a huge chain of polyprenyl diphosphates (C_{55}-C_{100}). A well-known polyisoprenoid is rubber which is synthesized by more than 2000 plants[27] and has a number of applications in automobile and other industries.

1.2 Terpene Synthases:

The enormous diversity of terpene carbon skeletons is mainly due to a class of enzymes that catalyze them, terpene synthases. These enzymes catalyze the conversion of acyclic prenyl diphosphates and squalene to various cyclic and acyclic products. A number of terpene synthases act on the same substrate to produce different products or in some cases multiple products creating product diversity. These enzymes have substrate preference which modulate its folding, stabilize the transient carbocation intermediates and control the quenching of these intermediates. Structural studies of these enzymes from lower prokaryotic to higher plant level have shown that the enzymes do not exhibit sequence identity. However, they adopt a common homologous fold to obtain such wide product diversity with conformational and stereochemical

accuracy[28,29]. The primary sequence identity among bacterial, fungal and plant terpene synthases is less than 15%.

Depending on the method of generation of initial carbocation or the way the catalysis is activated, the terpene synthases can be divided into two categories[12,30], the type I and type II terpene synthases. The type I terpene synthases are characterized based on the generation of an allylic carbocation due to abstraction of a diphosphate group (Fig1.3A). Whereas the terpene synthases in which carbocation abstraction happens due to protonation of a double bond of the substrate are termed as type II (Fig.1.3B). As there is no requirement of a diphosphate group for activation in the case of type II synthases, presence of diphosphate moiety in the substrates is not essential. The type I enzymes interact with three Mg^{2+} ions forming a trinuclear cluster that interacts with the diphosphate group and drives electrophilic ionization[31]. However, the Mg^{2+} ion binding in the case of type II enzymes is not found to be crucial for initiation of the catalysis and intermediate formation[32]. After the generation of carbocation, both the classes of terpene synthases generate a wide structural diversity of terpenes through a common chemical strategy (Fig.1.4), which include stabilization of carbocation, rearrangement reactions and quenching[12,30]. As the carbocation intermediate is highly reactive, it needs to be protected from the solvent. Since the carbocation intermediates are hydrophobic in nature the enzyme is lined with hydrophobic amino acids, which help protecting them from the bulk solvent.

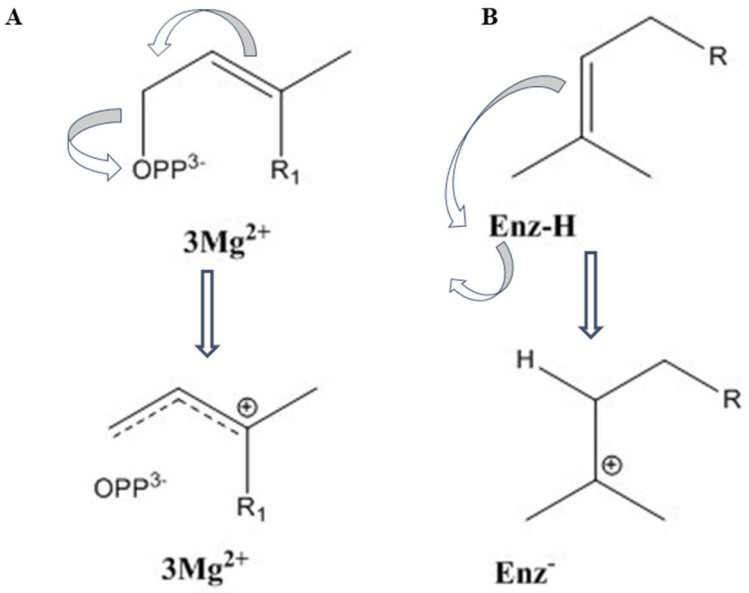


Fig.1.3. Classification of terpene synthases depending on their initial catalytic mechanism. *(A). Type I synthases initiate catalysis after ionization of the diphosphate substrate which coordinates with the metal ions (Mg^{2+}) leading to formation of carbocation. (B). Type II synthases initiate catalysis by proton addition to a carbon-carbon double bond of the substrate and this proton is donated by a general acid of the enzyme (Enz).*

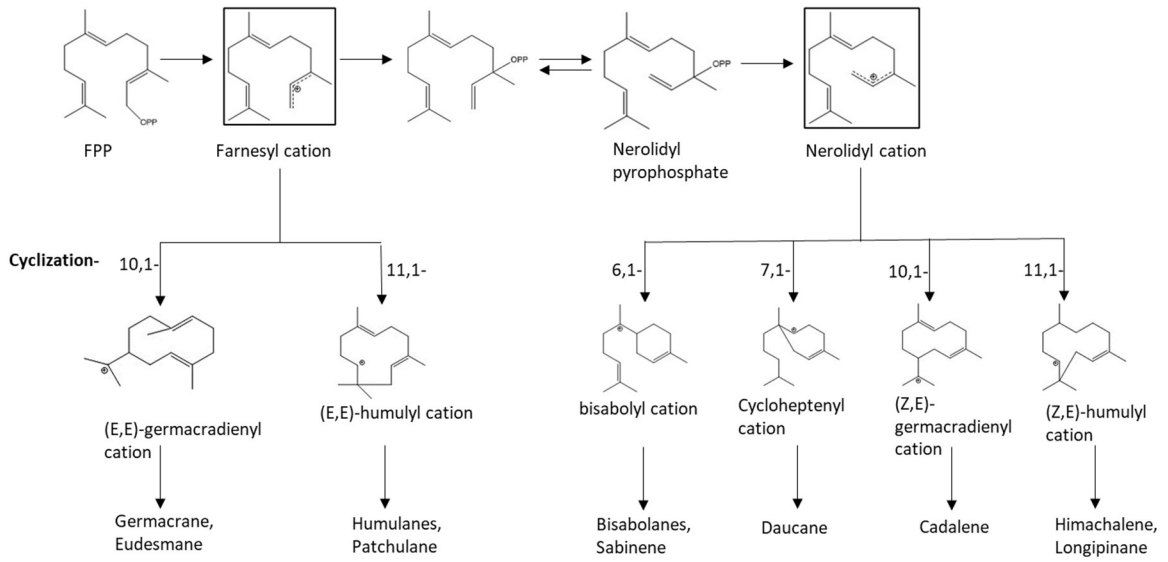


Fig.1.4. An example of generation of various carbocations formed due to different possible types of cyclisation of farnesyl cation. *Farnesyl cation formed after release of diphosphate anion undergoes isomerization and subsequently various cyclization and rearrangements leading to formation of different intermediates and thus sesquiterpenes.*

Carbocation rearrangements occur in a stepwise manner which includes hydride shifts, cation-alkene cyclization and a number of alkyl shifts like ring contraction and expansion and methyl shifts[12,33]. In certain cases proton transfers[33,34] are also included. The low energy carbocations are stabilized through a number of charge delocalization methods like, charge-dipole, charge-charge and charge-quadruple interactions[35]. In some terpene synthases, stabilization of transient cations takes place due to π-cation interaction between the side chains of aromatic amino acids like Phenylalanine, Tyrosine, Tryptophan and the carbocation[12]. The mechanism through which the terpene synthases control the cyclization reactions is still not clearly understood, however, few studies suggest mechanisms that include intrinsic reactivity of the substrate[36], templating of active site contour[37] and induced fits[38]. The cyclization reaction is terminated by the quenching of the final carbocation. An alkene is formed due to deprotonation of a carbon atom while solvent capture yields an alcohol and intramolecular hydroxyl capture yields an ether. The orientation and location of the water molecule and intermediate needs to be regulated by the enzyme, in case of water quenching [39].

1.2.1 Conserved motifs and their roles:

Terpene synthases are characterized by extremely conserved aspartate rich motifs[12]. The functions of putative terpene synthases from gene clusters, genomes and sequence databases

can be confirmed through the presence of these highly conserved motifs[40,41]. Depending on whether these signature motifs are present in type I or type II synthases, they have different functions. Type I synthases bind with the three Mg^{2+} ions for diphosphate detachment by using two distinct aspartate motifs. The prenyl transferase FPP synthase has two DDxxD motifs which bind to the metal[42]. Similarly, terpene synthases have two metal binding motifs, viz., **D**Dxx**D** and (**N**,**D**)D(L,I,V)x(**S**,**T**)xxx**E** or the NSE/DTE motif (Fig.1.5)[12]. The residues indicated in bold are involved in the metal binding. These two motifs are generally found to co-exist in almost all type I terpene synthases. However, these aspartate motifs play different roles in type II terpene synthases. The middle aspartate acts as a Bronsted Acid, which is required for protonation of double bond of the substrate[43] (Fig.1.6). The role of the motif in type II synthases was first identified in squalene-hopene cyclase (SHC) which protonates a double bond in squalene to initiate the cyclisation process[44,45]. Substrates of a number of type II terpene synthases possess diphosphate group which binds to Mg^{2+} [32], however, the Mg^{2+} does not interact with aspartate rich motif, as it does in the case of type I terpene synthases.

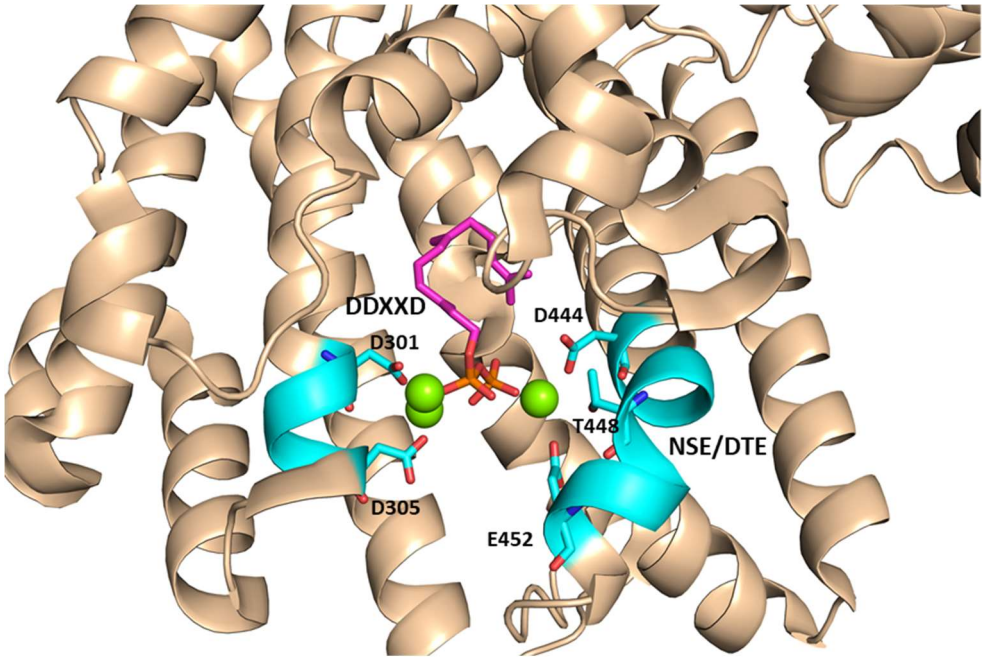


Fig.1.5. Structure of 5-epi-aristolochene synthase (PDB ID: 5IK0) mapped with conserved motifs. *The conserved DDXXD and NSE/DTE motifs are highlighted in cyan. The residues of both the motifs which coordinate with the metal ions are labelled. The Mg^{2+} ions and the ligand are shown in ball and stick, respectively.*

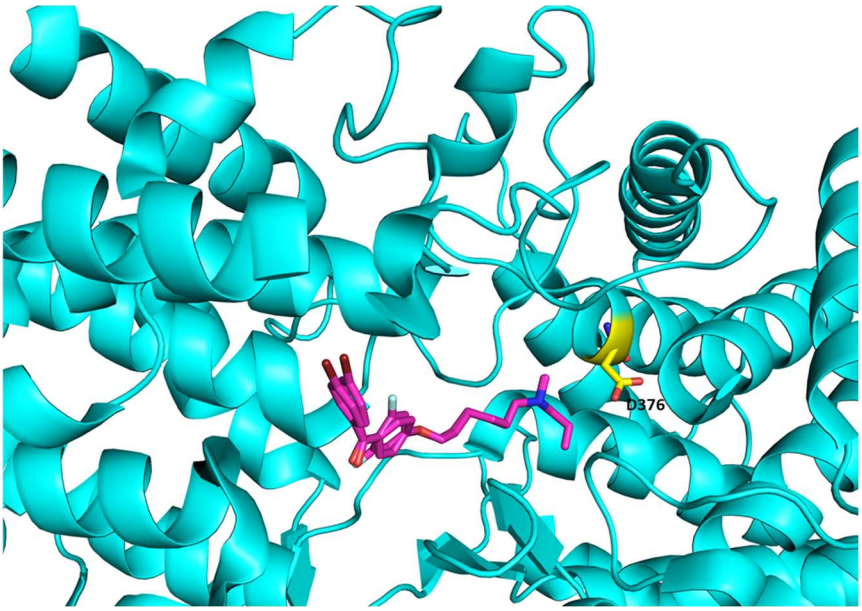


Fig.1.6. Structure of Squalene hopene cyclase (PDB ID: 1GSZ), which is a member of type II synthases. *The middle aspartate of the conserved DDXXD motif is highlighted in yellow color which acts as an acid for protonation of the substrate. The ligand is represented as sticks.*

Till date, structures of more than 30 terpene synthases, including mainly mono-, sesqui- and diterpene synthases, from bacteria, fungi and plants have been studied[12]. Type I synthases are characterized by α domain fold or the isoprenoid fold which consists of mostly 10-12 anti-parallel α-helices and was first studied in FPP synthase (Fig.1.7)[42]. Both prenyl transferases and type I synthases possess the same fold indicating the conserved ionization mechanism in both the groups of enzymes. The central region of the α-helical bundle is hydrophobic and contains the three Mg^{2+} ions and the conserved NSE/DTE and DDxxD motifs. On the contrary, the type II terpene synthases consist of β and γ domains which form a bi-modal fold. The hydrophobic cavity of this group of enzymes with the acidic DxDD motif is present at the interface of the two β and γ domains. The presence of these two types of α and βγ fold has been attributed to the gene duplication and fusion as there are two ancestral 4-helix bundles in the α domain and remarkable sequence and structural homology between β and γ domains. However, these domains show no homology with the α domain[46]. Presence of different combinations of α, β and γ domains in type I and II synthases[12,46] might have led to the evolution of bifunctional terpene synthases[46]. Furthermore, both type I and II synthases exhibit different tertiary structures with mono-, di- and tri-domains. For example, plant type I terpene synthases may consist of αα, αβ, α and αβγ folds (Fig.1.8A, B), while type II synthases may adopt βγ or αβγ folds (50) (Fig.1.8B). The catalytically active domains are characterized by asp-rich motifs and Mg^{2+} interacting motifs[47,48].

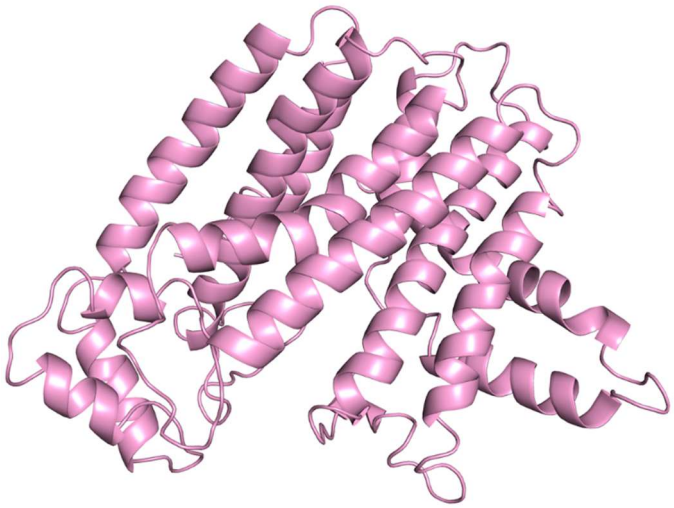


Fig.1.7. A depiction of Farnesyl diphosphate synthase from Gallus gallus (PDB ID: 1FPS). *FPP synthase belongs to type I synthase family consisting of only α domain, also known as isoprenoid fold.*

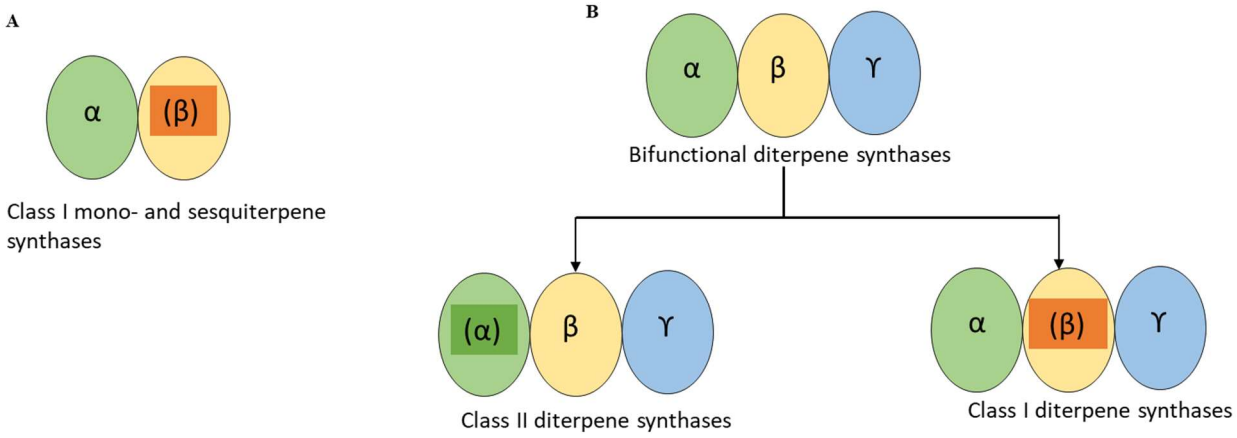


Fig.1.8. Different architectures of type I and type II plant terpene synthases. The domains, α,β & γ, are represented with different colors and the domains with parentheses are the inactive domains, which do not possess active site with DDXXD and NSE/DTE motifs. *(A) Type I terpene synthase with an active α domain and mainly consists of mono- and sesquiterpene synthases. (B) Diterpene synthases which consists of all the three domains and depending on the type of inactive domain is categorized as type I and type II.*

1.3 Sesquiterpene Synthases and their product specificity:

Several structural and biochemical studies on sesquiterpene synthases have provided a wealth of information on the mechanism of action of these class of enzymes. The product chemo diversity of sesquiterpene synthases could consist of single product with high fidelity from a synthase or formation of multiple products due to multiple cyclisation cascades of the enzyme (Fig.1.9).

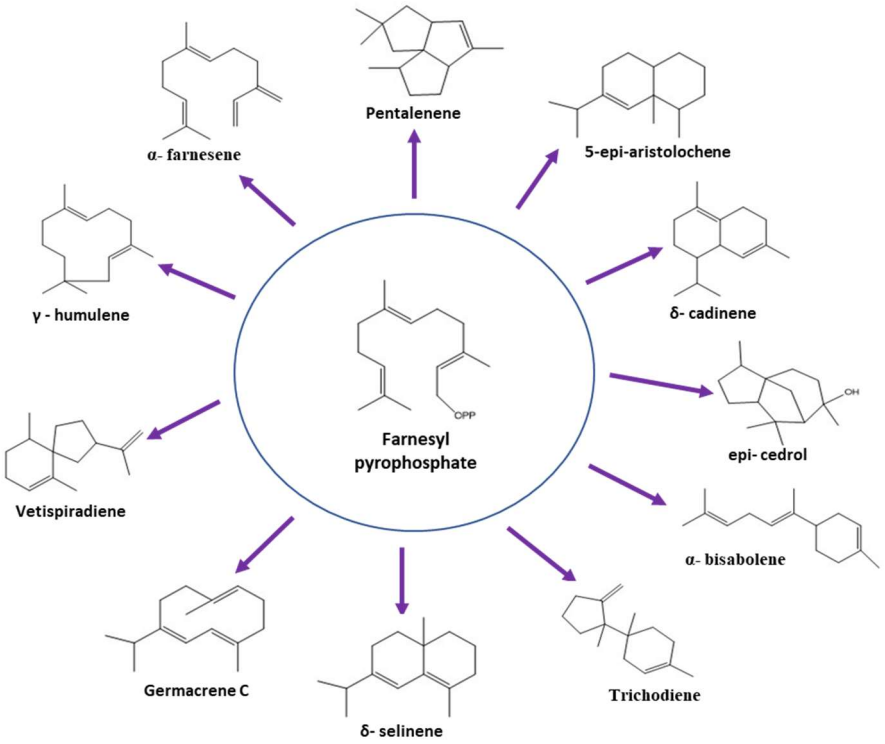


Fig.1.9. Product chemo-diversity of sesquiterpenes. *Sesquiterpene synthases catalyzing the cyclization of a single substrate, farnesyl pyrophosphate, into various cyclic and acyclic products.*

The "plasticity residues" which have the potential to alter this product specificity of these sesquiterpene synthases might be present within or outside the active site[49,50]. Minute differences in the active site or in its vicinity could result in an altered product profile[50,51]. Recognizing and understanding the functional plasticity residues in these synthases could aid in the engineering of the enzymes with regard to product specificity, catalytic efficiency and thermostability[50,51]. Also, a number of mutational studies of some of these enzymes have led to the identification of these plasticity residues, responsible for product specificity of terpene synthases[49,52–56]. Examples of identifying product specificity determinants of few sesquiterpene synthases are discussed below.

1.3.1 5-epi-aristolochene Synthase:

Structures of the *apo* and ligand bound form of 5-epi-aristolochene synthases from *Nicotiana tabacum* have provided information about the αβ helical domain organization of the protein. Comparisons of these structures suggest ordering of two loops in the protein, known as A-C and J-K loop. The ordered A-C and J-K loop in ligand-bound protein leads to closure of the active site due to the H-bond network between R266 of A-C loop, residues of J-K loop and the ligand, protecting the reactive carbocations from the solvent (Fig.1.10A)[56]. Comparative studies of 5-epi-aristolochene synthase and its close homolog, premnaspirodiene synthase from *Hyoscyamus muticus*, have led to the identification of divergent residues amongst the two

proteins[55] by using their contact map. These residues when mutated resulted in the switching of product specificity of both the enzymes. For example, a set of nine mutations (A274T, V291A, V372I, T402S, Y406L, S436N, I438T, I439L, V516I) in 5-epi-aristolochene synthase switched its product from 5-epi-aristolochene to that of premnaspirodiene (Fig.1.10B).

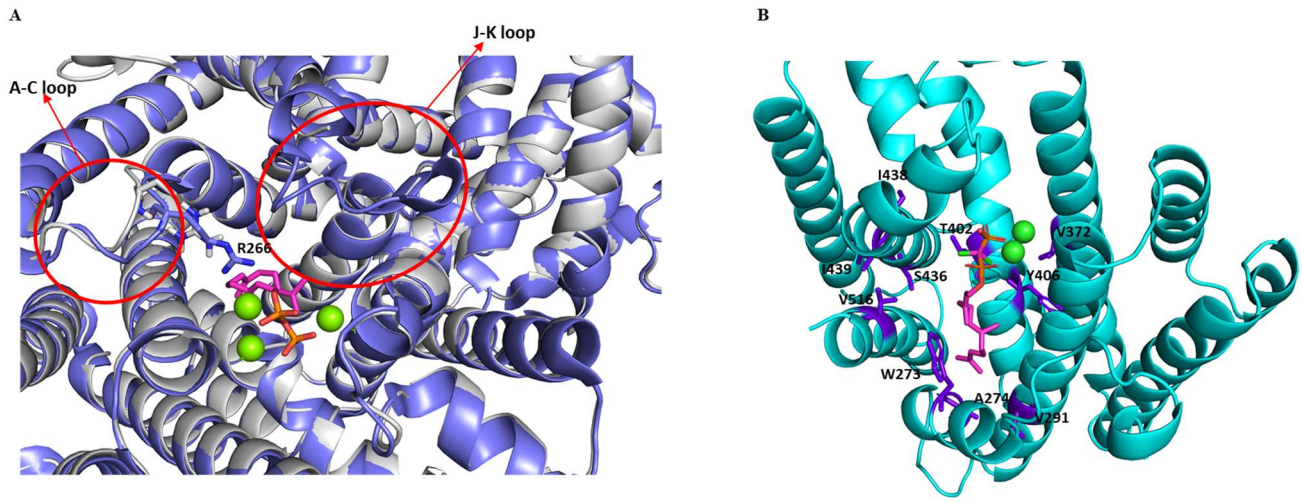


Fig.1.10. Mutational and conformational analysis of 5-epi-aristolochene synthase (TEAS) from Nicotiana tabacum. *(A) Conformational changes in A-C and J-K loop upon ligand binding which leads to formation of H-bond network between R266, J-K loop and ligand rendering closing of the active site. Structure of apo form (PDB ID: 5EAS) and ligand bound TEAS (PDB ID: 5IK0) are shown in grey and violet colors, respectively. (B) Divergent residues in the active site of TEAS which when mutated changes the product specificity of the enzyme. The Mg^{2+} ions and the ligand are shown in ball and stick, respectively.*

1.3.2 Epi-isozizaene Synthase:

Epi-isozizaene synthase catalyzes the cyclisation of FPP to form epi-isozizaene and minor products like α-bisabolene, sesquisabinene A, zizaene, etc**.** From the structure of the enzyme, it was proposed that the generation of initial carbocation could be due the Mg^{2+} dependent pyrophosphate cleavage of the substrate. Mutational and structural studies of epi-isozizaene synthase with BTAC (Benzyltriethyl ammonium cation), the aza analogue of the intermediate bisabolyl cation, have underlined a number of aromatic and aliphatic residues at the active site contour that play an important role in the product specificity. These residues include the aromatic triad F95, F96 and F198 and other aromatic and aliphatic residues lining the active site, including W203, W325, F332, L72, V329 and A236 (Fig.1.11)[52,57]. The aromatic triad F95, F96 and F198 stabilizes the carbocation transition states through cation-π interactions. Mutagenesis of these residues showed altered modes of stabilization of intermediates (farnesyl and bisabolyl) and also altered templates for FPP cyclization, thus, leading to production of non-cognate sesquiterpenes. This might be either due to redirection of the cyclization towards

intermediates other than the main intermediate, (7S) homobisabolyl cation, or due to early quenching of the positive charge of the carbocation.

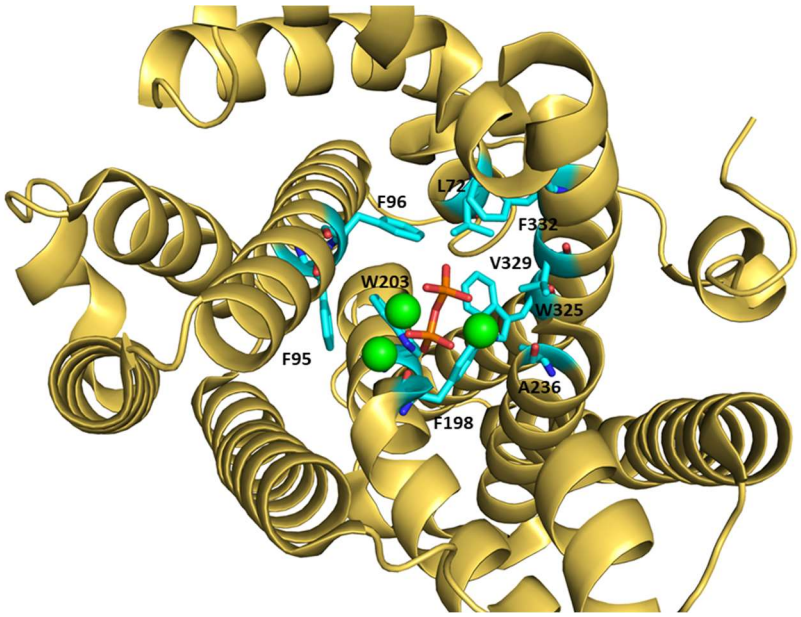


Fig.1.11. Product specificity determinants of Epi-isozizaene synthase (PDB ID: 3KB9). *The aliphatic and aromatic residues, which line the active site contour and have shown to play role in product specificity are highlighted with cyan color. The Mg^{2+} ions and the ligand are shown in ball and stick, respectively.*

1.3.3 δ-cadinene synthase:

Structure of unliganded and inhibitor (2-fluorofarnesyl diphosphate) bound delta-cadinene synthase from *Gossypium arboreum* has been reported[58]. This enzyme catalyzes the cyclization of FPP leading to the biosynthesis of gossypol, which is implicated in the defense of the plants from bacterial and fungal pathogens. δ-cadinene synthase shows sequence divergence in the conserved aspartate rich motif. Instead of NSE/DTE it has $D^{451}DVAE^{455}$ which coordinates with the Mg^{2+} (Fig.1.12). However, it has the βα domain, similar to that of other plant terpene synthases. Mutational studies of residues of the two Asp-rich motifs highlight their role in the enzyme catalysis and formation of intermediates, which includes nerolidyl diphosphate (NPP).

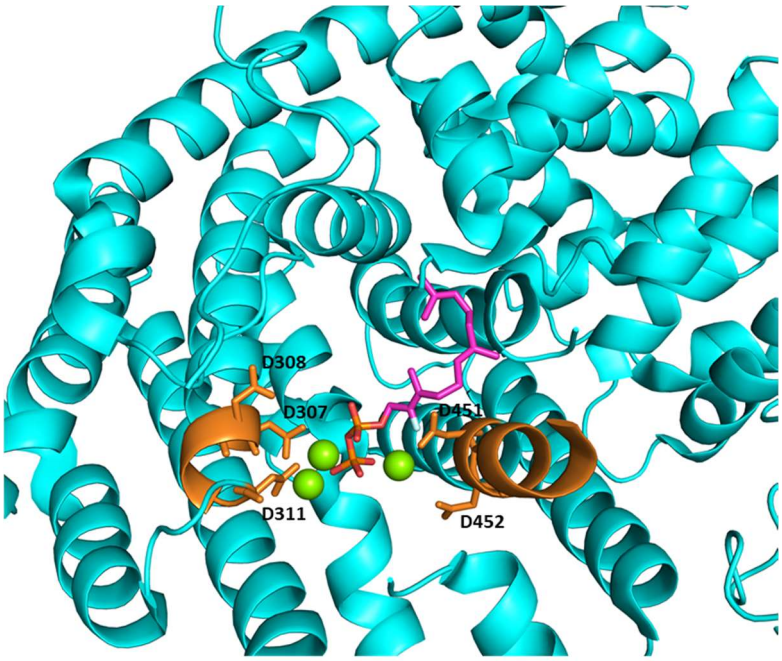


Fig.1.12. Crystal structure of Delta-cadinene synthase (PDB ID: 3G4F) with the active site motifs. *The enzyme possesses an additional Asp-rich motif, instead of NSE/DTE motif, which coordinates with metal ions. The* Mg^{2+} *ions and the ligand are shown in ball and stick, respectively.*

1.3.4 α-bisabolol synthase:

Structural studies of *apo* α-bisabolol synthase and its mutant from *Artemisia annua* along with domain swapping and site-directed mutagenesis experiments have aided in identification of residues of BOS motif of the enzyme that plays crucial role in the product specificity[49]. α-bisabolol synthase has significant sequence similarity (82%) with Amorpha-4,11-diene synthase and both the enzymes produce bisabolyl cation as their common intermediate. Domain swapping experiments helped in recognizing the regions of both the proteins which play role in the formation of distinct products by the enzymes. These experiments involve generation of chimeric enzymes with reciprocal replacement of corresponding sites for both the proteins. This was followed by measuring the activities of the chimeric enzymes to identify the fragments essential for the activity. These active fragments were further subdivided to obtain minimum segment or residues that mediate product specificity. Thus, the comparative analysis of both the enzymes led to the identification of key residues which might play role in catalysis and product specificity of the enzyme (Fig.1.13).

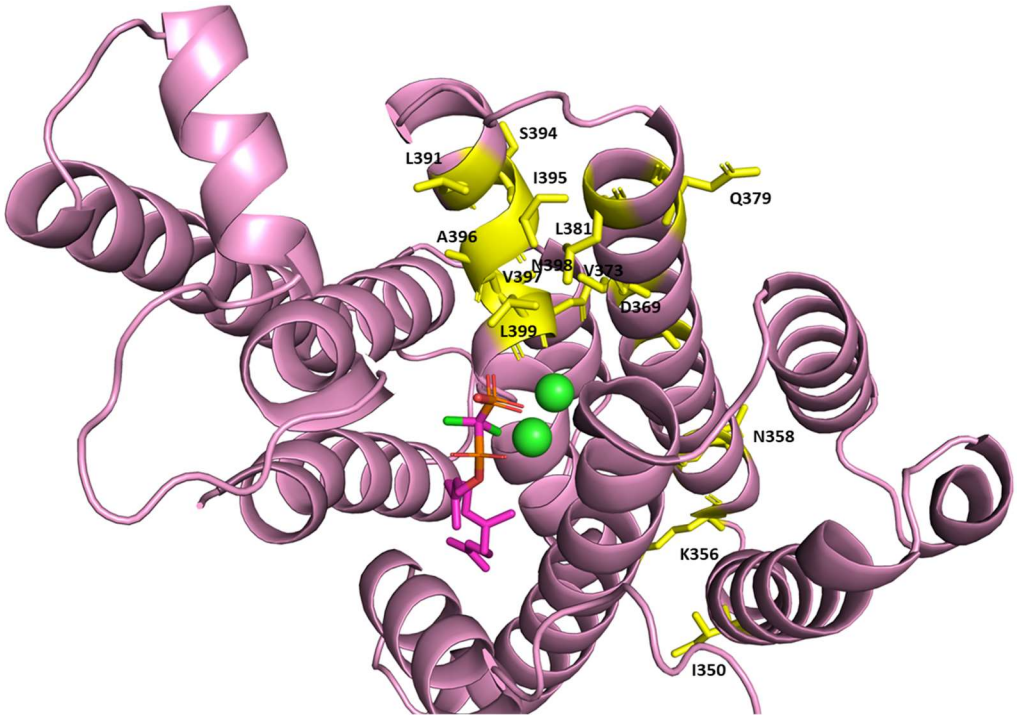


Fig.1.13. Structure of α -bisabolol synthase (PDB ID: 4FJQ) docked with FPP and Mg^{2+} ions. *The residues of the active site which change the product specificity of the enzyme leading to production of an additional product, γ –humulene. The Mg^{2+} ions and the ligand are shown in ball and stick, respectively.*

1.3.5 β-farnesene synthase:

In *Artemisia annua*, two important sesquiterpene synthases, Amorpha-4,11-diene synthase and β-farnesene synthase catalyze the cyclization of FPP to a cyclic (amorpha-4,11-diene) and an acyclic (β-farnesene) product. These two proteins share a sequence identity of 49% [59]. From protein models 24 variable amino acid positions (which lie within a radius of 6 Å from the center of the active site) yielding 2^{24} possible combinations of mutants that could influence the product were identified. Based on the epistatic interactions between substitutions, 754 mutants clustered in nine groups were tested for their role in substrate specificity [59]. From this analysis three residues (Y402, Y430 and V467), which play crucial role in modulating product specificity and cyclization were identified (Fig.1.14).

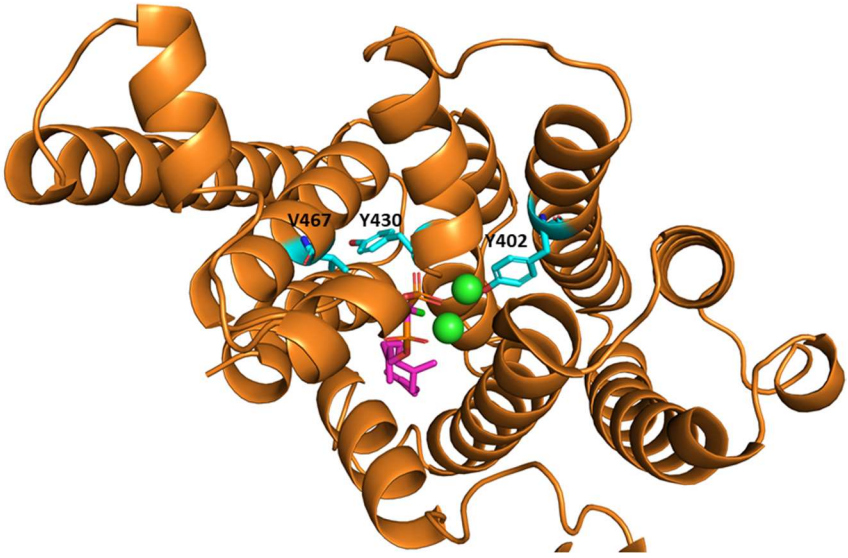


Fig.1.14. Modelled structure of β -farnesene synthase. *Mutation of the residues Y402 and Y430 lead to change in the product formation from acyclic to cyclic and mutation of V467 reverses this effect. The Mg^{2+} ions and the ligand are shown in ball and stick, respectively.*

1.3.6 Selinadiene synthase:

Structural studies of Selinadiene synthase (SdS), which catalyzes FPP into selina-4-diene and germacrene B as the major and minor products, provided insights on an "effector triad", R178, D181 and G182 that influences the substrate binding and reengagements. It has been seen that, upon substrate binding, R178 shifts towards the ligand and forms hydrogen bond with pyrophosphate which is further stabilized by D181 (Fig.1.15A). This ternary complex leads to conformational change in the helix, which brings catalytic carbonyl oxygen of G182 near the active site to trigger the cleavage and release of pyrophosphate. Mutation of the residue R178 and other aromatic or hydrophobic residues near the binding pocket of the enzyme like F55, F79 and Y152 (Fig.1.15B) leads to change in the product specificity of the enzyme, yielding germacrene B as the major product [38].

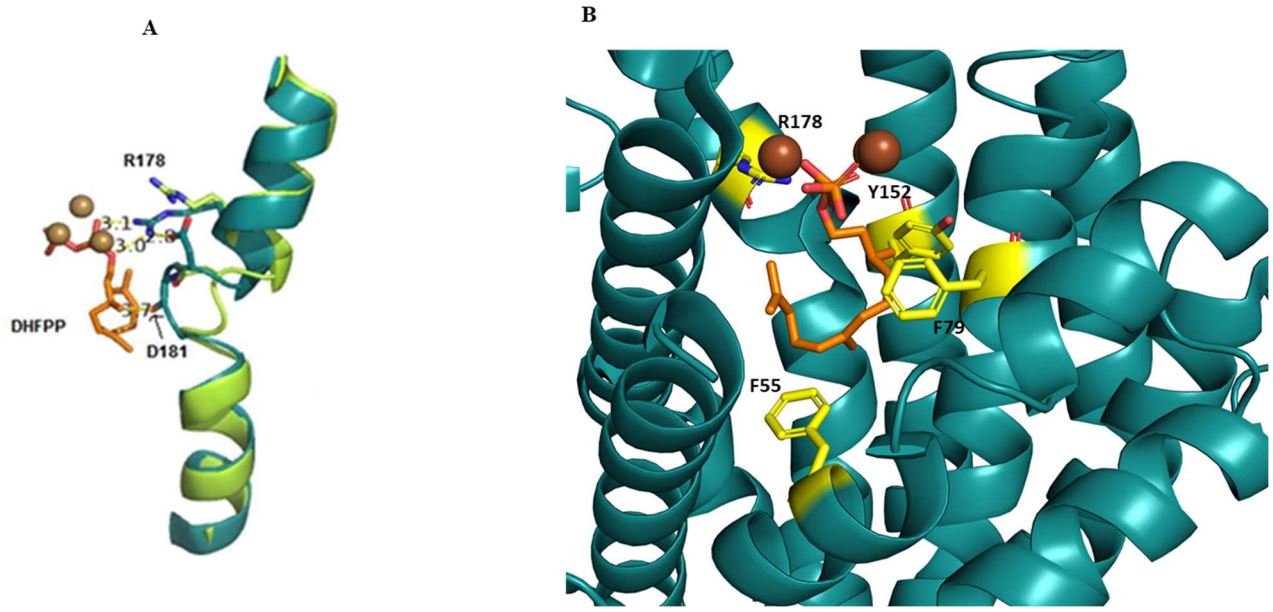


Fig.1.15. 'Effector triad' of Selinadiene synthase. *(A) Superposition of ligand bound (PDB ID: 4OKZ) (Deepsteal) and apo form (PDB ID: 4OKM) (green) of Selinadiene synthase showing rearrangement of 'effector triad' leading to H-bond formation between R178-D181-substrate. (B) Product specificity determinants highlighted in yellow color. The Mg^{2+} ions and the ligand are shown in ball and stick, respectively.*

1.4 Sandalwood and its oil:

Santalum belongs to the family Santalaceae which consists of approximately 500 species, of which 19 belong to the genus *Santalum* [60]. They grow slowly and are hemiparasite trees. These species are distributed throughout the tropical and temperate regions of India, Pacific Islands and Australia. Indian sandalwood or East Indian sandalwood (*Santalum album*) and Australian sandalwood (*Santalum spicata*) are of very high economic value due to their commercial utility. Indian sandalwood has a number of applications in various industries like cosmetic, aroma and perfume and traditional medicine and is also used for religious purposes[60]. The use of Indian sandalwood dates back to ancient times for carving idols, boxes, cabinets and tables. It was used for various fire rituals of Hindus and Buddhists and also to carve temples and idols of gods and goddesses[60]. Sandalwood trade with the East has a historical relevance dating back to 5th century BC due to its fragrant heartwood and oil[60]. Apart from religious importance, sandalwood has medicinal applications too. The paste of its wood is used as an ointment.

Essential oils of sandalwood comprise of mixtures of a number of aldehydes, terpenoids, ketones, alcohols, esters and aromatic compounds. These oils have various applications in food flavoring, perfume and pharmaceutical industries[61]. One such oil is sandalwood oil which is volatile in nature and is obtained from the dried wood of the trunks and roots of the tree. It is

used in perfumes, aromatherapy, cosmetics, incense sticks and pharmaceuticals. Other benefits of the oil include antiviral, bactericidal, anti-inflammatory, antipyretic and anticarcinogenic activity.

1.5 Sesquiterpene biosynthesis in Indian sandalwood:

The sandalwood oil comprises of more than 100 compounds. The major constituents which contribute to the fragrance are α-santalol (≥60% of total santalol), β-santalol (≥33% of total santalol) and 2-fufuryl pyrrole[62]. Other components of the oil include sesquiterpene hydrocarbons like α-santalene, β-santalene, epi- β-santalene, α-bisabolol, β-bisabolene, α-curcumene, β-curcumene and γ-curcumene[62]. The other reported constituents include santene, teresantol, dihydro-β-agarofuran, santalone, teresantalic acid, santanol and tricycloekasantalal[62]. α-santalol and β-santalol contribute mostly to biological benefits of the oil and have shown to exhibit neuroleptic and chemo-preventive properties in bioassay systems[63]. In *S. album*, the genes involved in biosynthesis of santalene are expressed highly at the transition zone of sapwood and heartwood. Cyclisation of FPP to yield α- and β-santalenes by santalene synthase is the first step in santalol biosynthesis. The most important set of enzymes that are responsible for terpene biosynthesis in sandalwood include santalene synthase, β-bisabolene synthase and sesquisabinene synthase. These enzymes catalyse a complex reaction which produces an array of sesquiterpenes including α-santalene, β-santalene, epi- β-santalene, (E)- β-farnesene, exo- α-bergametone, sesquisabinene and β-bisabolene, employing a common substrate FPP[64]. Further, cytochrome P450 system converts these santalenes to santalols by hydroxylation of santalene derivatives at their *cis*-methyl position. To understand the biosynthesis of sesquiterpenes in Indian sandalwood and the mechanism of action of the enzymes involved, Prabhakar et al. performed gene isolation and functional characterization of farnesyl diphosphate synthase (SaFDS), bisabolene synthase (SaBS), santalene synthase (SaSS) and two isoforms of sesquisabinene synthases (sesquisabinene synthase 1, SaSQS1 and sesquisabinene synthase 2, SaSQS2)[19]. They carried heterologous expression of the two isoforms, SaSQS1 and SaSQS2, in microbial system and compared their degree of expression and kinetic parameters. The two isoforms are 82.8% identical. Of the two isoforms, SaSQS1 is kinetically more active. The 1701 bp nucleotide coding for 566 amino acids was found to have a molecular weight of 63 kDa. GC-MS analysis of SaSQS1 indicated that the enzyme converts FPP to sesquisabinene (>93%), β-sesquiphellandrene (~5%) and unidentified metabolite (~2%). In this thesis, structural and

mechanistic aspects of SaSQS1 and the factors that influence the type of product formed, are discussed.

1.6 Statement of Problem:

Terpenes are the secondary metabolites which have a number of biotechnological applications like pharmaceuticals, fragrance & cosmetics, pesticides, food additives and biofuel. Sesquiterpene synthases catalyze the cyclization of FPP to a wide array of one such terpene, sesquiterpenes. Recent structural and mutational studies of sesquiterpene synthases from various organisms have provided insights to the mechanism of action of these enzymes and also their product specificity. These studies employ extensive testing of mutants in order to determine the product defining residues of the enzyme. Also, the rationale underlying the structure based mutational studies markedly varies.

In the present study, we aim to understand the mechanism of action and product specificity of a sesquiterpene synthase, SaSQS1 from Indian sandalwood. It catalyzes the cyclization and conversion of FPP to sesquisabinene, which is a key component of commercially important sandalwood oil. Similar to other sesquiterpene synthases, SaSQS1 also forms a distinct product, though utilizes the same substrate FPP. However, the structural basis of product specificity of SaSQS1 is not known. Understanding the mechanistic aspect and product specificity of the enzyme will help to identify the residues affecting the formation and type of products. This information would further help in devising a general approach to identify the product specificity of the sesquiterpene synthases. The outcome of this study will have significant impact on developing recombinant techniques to produce these highly useful molecules at the industrial scale.

1.7 Objectives:

The objective of the present work is to understand the structural and mechanistic aspects of an important sesquiterpene synthase, SaSQS1. Employing structural, biochemical and computational approaches we try to address the following objectives:

1.7.1 Structural basis of mechanism of action of SaSQS1:

SaSQS1 is one of the two isoforms of sesquisabinene synthase from Indian sandalwood (*Santalum album*), which catalyzes cyclization of FPP to form sesquisabinene and other minor products, which are significant components of sandalwood oil. Similar to other sesquiterpene synthases, SaSQS1 cyclizes FPP to varied product. However, the structural basis of this

conversion is not known. Hence, aim here is to decipher the architecture and mechanism of action of SaSQS1 by elucidating its structure and performing comparative sequence and structural analysis with other sesquiterpene synthases.

1.7.2 Mutational studies of SaSQS1 to identify the product specificity determinants:

A number of studies have been undertaken to define the product specificity of sesquiterpene synthases. These studies involve comparative analysis of the homologous structures and sequences which delineate the divergent residues in the vicinity of the ligand binding site and these residues are validated by mutagenesis[49,55,59]. We aim to identify the product specificity determinants of SaSQS1 by employing this traditional approach and validating the role of product specificity attributed by these residues through mutagenesis studies. The structure based mutants provide insights on how minor changes in the essential similar catalytic pocket alter the type and quantum of the product.

1.7.3 Determining a novel approach to study product specificity of SaSQS1 and other plant sesquiterpene synthases:

Given the huge size of the catalytic pocket and indeterminate *modus operandi* of sesquiterpene synthases, the traditional approach to identify the product specificity poses serious challenges. This is primarily due to the involvement of testing large number of mutants of divergent residues at the catalytic site. Furthermore, the inconsistency in reported methodologies to determine these product modulating residues has confounded field. Thus, a generalized approach that could provide putative set of product modulating residues of sesquiterpene synthases is lacking. Here, the objective is to design a generalized approach, using statistical coupling analysis and spectral decomposition of dynamical content of catalytic site residues for identifying the product modulating residues in few plant sesquiterpene synthases. The model thus developed would be validated with available biochemical data on this family of enzymes.

Chapter 2

Materials and Methods

This chapter outlines the methodology used for biochemical characterization and structure determination of proteins reported in the thesis. X-ray crystallography has been employed to elucidate the structures in the current studies. Other techniques, including molecular, biochemical, analytical and computational methods, which were used in addressing the research question, have also been discussed here.

2.1 Materials:

The reagents and materials used in the present work are listed in Table 2.1.

Chemicals	Company
20x SYBR Green	Invitrogen, USA
2β-mercaptoethanol	HiMedia, India
Anti-6xHis tag primary antibody	ThermoFisher, USA
Acrylamide	Sigma-Aldrich, USA
Agarose	SRL, India
Ampicillin	MP Biomedicals, USA
Bromophenol blue	HiMedia, India
Calcium chloride	MP Biomedicals
Column chromatography columns and resins	Cytiva, USA
cOmplete EDTA-free protease	Roche, Switzerland
Coomassie Brilliant Blue R-250	SRL, India
Commercial Crystallization suits	Qiagen, Germany & Molecular Dimensions, UK
Dithiothreitol	MP Biomedicals, USA
Dimethylsulfoxide (DMSO)	MP Biomedicals, USA
Ethanol	HiMedia, India
Ethylenediamine Tetraacetic acid (EDTA)	MP Biomedicals, USA
Gel extraction and PCR purification kit	Promega, USA
Glutathione reduced	MP Biomedicals, USA
Glycerol	SRL, India
HEPES	Gibco, ThermoFisher, USA

Hydrochloric acid	Merck, Germany
HRP-conjugated secondary antibody	ThermoFisher, USA
His60 Ni gravity column	Roche, USA
Hexylene glycol	Sigma-Aldrich, USA
iBlot Mini Western Blot Stacks	Invitrogen, USA
Isopropanol	MP Biomedicals, USA
Isopropyl-β-D-thiogalactopyranoside	MP Biomedicals, USA
Kanamycin	Gibco, ThermoFisher, USA
LB media	HiMedia, India
LB agar	HiMedia, India
Magnesium sulfate	Sigma-Aldrich, USA
Magnesium chloride	Sigma-Aldrich, USA
n-Hexane	Thomas Baker Chemicals, India
Noble agar	Sigma-Aldrich, USA
Oligonucleotide primers	Eurofins, India
pET vector system	Novagen, USA
pOPINSS vector system	Novagen, USA
Plasticware	Tarsons, India
Poly Ethylene glycol	Sigma-Aldrich, USA
QuickChange lightning kit	Agilent
Restriction endonucleases	NEB, USA
Tris-HCl	MP Biomedicals, USA
Sodium chloride	MP Biomedicals, USA
Sodium dodecyl sulfate	Sigma-Aldrich, USA
Sodium hydroxide	Sigma-Aldrich, USA
Sodium sulfate	HiMedia, India
Sucrose	Sigma-Aldrich, USA
Taq DNA polymerase	NEB, UK
T4-DNA ligase	Promega, USA
Vivaspin, protein concentrators	Cytiva, USA

Table 2.1. *List of reagents and chemicals used*

2.2 Molecular Biology Methods:

2.2.1 Primers:

For designing the primers, all the desired nucleotide sequences of genes were obtained from the NCBI Database. Short length sequences that match the beginning and end of the gene were

identified. 22bp-30bp was considered to be the optimal primer length. Online tool, Oligonucleotide Properties Calculator (OligoCalc)[65], was used to determine the other parameters such as Tm and GC content of primer. The designed primer sequences were commercially synthesized from M/s Sigma and M/s Eurofins, Bengaluru.

2.2.2 Polymerase chain reaction (PCR):

Polymerase chain reaction refers to the enzymatic amplification of the DNA template using a pair of primers and DNA polymerase, which results in a new DNA strand complementary to the template strand[66]. PCR was carried out in 30 µL reaction mixture consisting of 10X manufacturer's polymerase buffer, 30-50 ng of DNA template, 1 U of DNA polymerase enzyme, 2 mM dNTPs, 0 .1 µmoles of primers, and RNase free water. These reactions were carried out in a thermal cycler (Veriti, Applied Biosystems) with a particular cycling profile as mentioned in the Table 2.2.

Step 1	**Denaturation**	**95**	**1 min**	**Cycle 1**
Step 2	Denaturation	95	10 sec	Cycle 10
	Annealing	61.5	10 sec	
	Extension	68	As per gene length (1 kb/min)	
	Extension	72	10 sec	
Step 3	Denaturation	95	10 sec	Cycle 15
	Annealing	58.5	10 sec	
	Extension	68	As per gene length (1 kb/min)	
	Extension	72	10 sec	
Step 4	Denaturation	95	10 sec	Cycle 30
	Annealing	55.5	10 sec	
	Extension	68	As per gene length (1 kb/min)	
	Extension	72	10 sec	
Step 5	Extension	72	10 min	Cycle 1
Step 6	Hold	4	Infinite	

Table 2.2. *PCR condition with different temperatures and their respective duration and number of cycles.*

Post PCR amplification, agarose gel electrophoresis was performed to assess the amplification and also to extract the amplified PCR products. The amplified gene segments were further subjected to restriction digestion and then cloning.

2.2.3 Purification of PCR Products:

Usually, several primer dimers and non-specific amplicons are also synthesized together with the desired amplicons in PCR. The amplicons which were of the expected size were purified by resolving them on agarose gel, and MiniGel extraction Kit (Qiagen, Germany) was used for extraction.

2.2.4 Ligation:

The amplified PCR products and desired vectors are subjected to the restriction digestion test with a set of restriction endonuclease (RE) enzymes before ligation. For ligation, the reaction mixture of RE digested vector and PCR products, manufacture's ligation buffer and ligase enzyme (Promega) with the final reaction volume of 10 µL is set up. The reaction was incubated overnight at 4 °C or for 2 hours at room temperature. Post incubation, the reaction mixture was used for transformation in chemically competent DH5α cells.

2.2.5 USER Cloning:

USER (Uracil-Specific Excision Reagent)[67] cloning is a ligase-free uracil DNA glycosylase (UDG) mediated cloning. Here, the single standard flanking ends of the primer, which contains the deoxyuridine nucleotide (dU), permits ligase-free directional cloning. The DNA amplification reactions should be carried with polymerase, such as PfuTurbo Cx Hotstart (Agilent), that incorporates a deoxyadenosine complimentary to dU present in the primers. The USER™ enzyme (a mixture of endonuclease VIII and UDG) excise the dU residues to produce 3' single-strand overhangs on amplified DNA fragments that can incorporate in a linear vector with complementary overhangs (Fig 2.1)[68]. The ligation approach based on USER has numerous advantages, which includes (a) production of long single-stranded extensions, (b) unique single-stranded extensions enables the linear vector that cannot ligate to form circular DNA, (c) distinctive single-stranded extension allows the rapid assembly of multiple DNA fragments into the vector[67].

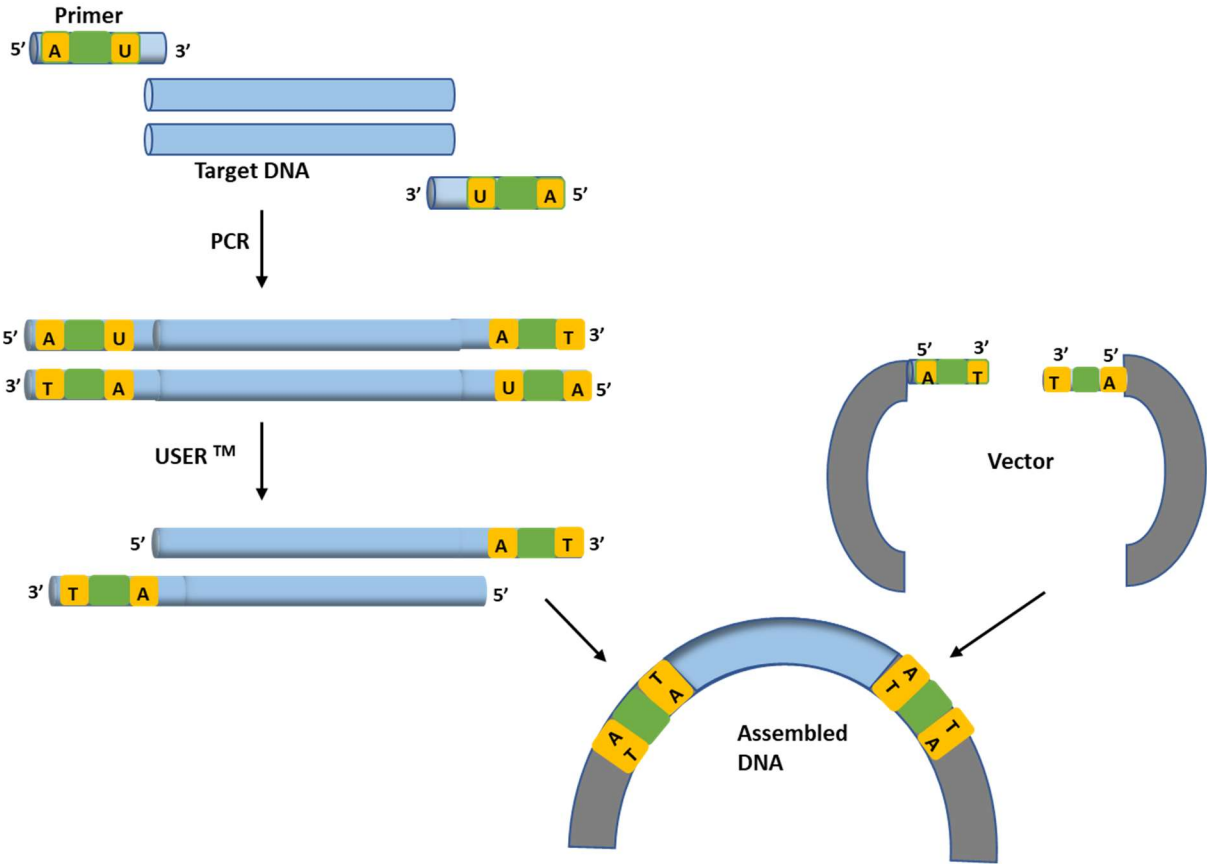


Fig.2.1.*USER cloning. Assembling of gene of interest with linear vector using USER cloning method*

2.2.6 Transformation:

The ultra-competent cells of *E.coli* strains like C41 and DH5α were made for transformation using the Inoue method[69]. Competent cells were thawed on ice for 5 mins, followed by addition of ligation reaction mixture. The cells and DNA were mildly flicked and were kept on ice for 30 mins. For the next 60 secs, the cells were heat-shocked at 42 °C in a water bath, which was followed by incubation on ice for 2 mins. Afterward, 1 mL of LB Broth was added to the transformation reaction, followed by incubation for an hour on a shaker at 37 °C. Further, the transformation reaction was plated on agar plates with the desired antibiotic for selection. The plates were then incubated overnight at 37 °C.

2.2.7 Plasmid DNA Purification:

The purification of plasmid DNA was based on alkaline lysis, followed by binding of plasmid DNA to the manufacturer's (Qiagen) anion exchange matrix columns under low-salt and pH conditions. Furthermore, the matrix was washed with ethanol, which contains a medium salt buffer to eliminate RNA, protein, and other impurities. Lastly, elution was carried out in warm nuclease-free water.

2.2.8 Quantification of DNA:

The quantification of DNA was done by a Nano-drop spectrophotometer (NanoVue GE Health Care). The concentration was identified by measuring the optimum absorption of double-stranded DNA at 260 nm. Protein has an absorption peak at 280 nm; therefore, the ratio of 260:280 aids in determining the purity of the isolated DNA. Pure DNA can be indicated with a ratio between 1.8 and 2.0[70] .

2.2.9 Site-directed mutagenesis:

With the QuikChange Lightning site-directed mutagenesis kit (Agilent) and mutation containing primer pair, specific mutants were amplified by PCR. The primers were synthesized by M/s Eurofins Bengaluru. The plasmid with the *wild* type gene was used as a template. PCR was carried out in a thermal cycler by using the following program: one cycle of 5 min at 95 °C denaturation followed by 18 cycles of 30 sec at 95 °C denaturation; 30 sec at 54 °C annealing; 7 min 30 sec at 68 °C extension and a single cycle of final extension at 68 °C for 10 min followed by hold at 4 °C. Amplified products were subjected to Dpn1 enzyme digestion at 37 °C for 2 hours and was followed by transformation in DH5α competent cells.

2.2.10 Confirmation of the clone and plasmid sequencing:

Prior sequencing, the purified plasmid samples were validated for the presence of genes of interest with restriction digestion by using particular restriction enzymes. The digestion reaction mixture consists of 0.5-1 µg of plasmid, 1 U of restriction enzyme, and 1X manufacturer's buffer. The mixture was incubated for 2-3 hours at 37 °C. The digestion pattern was analyzed using 0.7% agarose gel. Clones with the accurate RE digestion pattern were given for sequencing to M/s Eurofins Bengaluru. The sequencing chromatograms were analyzed with the help of chromas 2.1 software.

2.3 Analytical Methods:

2.3.1 Electrophoresis:

Electrophoresis refers to the migration of charged particles (ions), under the influence of electric field, through the matrix of agarose or polyacrylamide and their separation[71].

2.3.2 Agarose gel Electrophoresis:

Agarose gels were made by heating agarose in TAE buffer (1 mM EDTA, 40 mM Tris-acetate, pH 8.0). Bio-Rad Mini electrophoresis cells (Bio-Rad, USA) were used to cast and run the gel. The concentration of agarose depends on the size of the DNA which needs to be resolved. The

gel was placed in the electrophoresis cell. TAE buffer was poured over the gel up to 5 mm precisely. After mixing with the loading dye (0.25% Bromophenol blue, 0.25% Xylene cyanol FF, and 40% Sucrose), DNA samples were added the wells. 200 ng ladder (New England Biolabs, UK) was used as the molecular weight markers. Further, DNA samples were electrophoresed for 40 mins at 5 V/cm through the gel. The DNA was analyzed in a Gel Doc XR System (Bio-Rad, USA) under ultraviolet light.

2.3.3 SDS-poly acrylamide Gel Electrophoresis:

Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis (SDS PAGE) was performed to resolve denatured protein molecules according to their mass. All gels utilized in this analysis were made with either 12% or 10% resolving gel and 5% stacking gel in 1.0 mm mini gel cassettes (Bio-Rad, USA). To carry out SDS PAGE, gels were placed in a cassette at first, with 1X SDS Running buffer filled in the anode as well as cathode chamber. Prior to loading, 1X SDS sample buffer was mixed to the protein samples, and the sample was denatured for 5 mins at 95ºC. Electrophoresis was performed for 2 hours at 150 V. Then, to analyze the protein bands, gels were stained with InstantBlue Protein Stain prepared in the lab.

2.4 Biochemical Methods:

2.4.1 Protein Expression and Purification:

2.4.1.1 Expression Methodology:

The positive clones validated by plasmid sequencing were transformed in *E. coli* C41 strain. The colonies obtained were inoculated in 5 ml of LB broth with the desired antibiotic, followed by overnight incubation at 37 °C at 180 rpm on a rotary shaker. 1 ml of this starter culture was inoculated into 100 ml LB broth with the antibiotic and incubated on a shaker at 37 °C till OD_{600} reached 0.6-0.8. 1 ml culture was kept on ice before induction as uninduced control. To carry out induction, isopropyl 1-thio-D- galactopyranoside (IPTG) at the final concentration of 0.5 mM was added to the culture and was then incubated at 20 °C for 14 hours along with shaking. The samples were centrifuged at 4000 rpm for 15 mins at 4 °C. The cell pellet was stored at -80 °C until further use, and supernatant consisting of media was discarded.

2.4.1.2 Expression Analysis:

The stored cell pellet was resuspended in Tris-HCl buffer and sonicated for 30 sec. The lysate was centrifuged at the 12000 rpm, and the supernatant was loaded on 12% SDS-PAGE to confirm the protein expression.

2.4.1.3 Analysis of solubility:

The pellet was resuspended into a 5 ml buffer to analyze solubility of the over-expressed protein, and 0.75 mg/ml of lysozyme was added to it. The lysate was sonicated and then centrifuged at 12000 rpm for 20 mins at 4°C. The settled pellet and the clear supernatant were analyzed to determine the existence of the over-expressed protein. The protein is considered soluble if it is observed in the clear supernatant.

2.4.1.4 Large scale Expression and Purification:

The optimum protein expression was observed in *E. coli* strain, C41. For large-scale expression, cells were grown in 4L LB media to an OD_{600} of 0.5 at 37 °C, induced with 0.5 to 1 mM IPTG, and incubated at 20 °C for 14 hours. The culture was harvested by centrifugation. The cell pellet was suspended in chilled lysis buffer, containing lysozyme, sucrose and a tablet of cOmplete EDTA-free protease inhibitor, followed by lysing of cells using sonication. For affinity binding, the lysate was purified and loaded to the Ni-NTA column (GE Healthcare Biosciences, USA). The column was washed with wash buffer, and the protein was eluted with a particular concentration of imidazole prepared in the wash buffer. The elution fractions were dialyzed overnight at 10 °C together with the TEV protease, which was purified in our lab. Desalting of the dialyzed fraction was done, followed by a second affinity column to remove the TEV protease and the 6X-His tag. Size exclusion chromatography was carried out with Superdex or Sephacryl columns (GE Healthcare Biosciences, USA) to enhance the protein's purity. The peak fractions under a particular molecular weight were concentrated to perform crystallization and biochemical experiments.

2.4.2 Gas Chromatography-Mass Spectrometry (GC-MS) assay:

GC-MS based assay was performed to analyze enzyme activity. The reaction mixture having 600 µg of enzyme, 200 µM substrate in buffer containing 25 mM HEPES (pH 7.5), 10%glycerol, 5 mM DTT and 10 mM Magnesium Chloride was incubated at 25-30 ^{0}C for 2hrs. Compounds were extracted in n-Hexane. The extracted compound was concentrated with a stream of dry nitrogen and injected into GC-MS for analysis using the following program: 70 °C to 170 °C at 5 °C/min followed by 170 °C to 210 °C at 15 °C/min with a hold time of 5

minutes. Using a 30 m x 0.32 mm x 0.30 µm capillary column (HP-5 MS, J&W Scientific) [19] the separation was performed with helium gas as a carrier at a flow rate of 1 ml/min.

2.5 X-ray crystallography: Deciphering protein structures:

X-ray crystallography has been the most popular and powerful method to obtain protein structure at the atomic resolution. Recent advances including tunable X-ray sources of synchrotrons, robotics handling the samples, cryoprotection and CCD detectors have made this procedure easier and faster. The structure and function of biomolecules are correlated, thus, the elucidation of structural information of these biomolecules aids in understanding the detailed mechanism of the macromolecules. In order to study their structural aspects using X-ray diffraction, proteins are subjected to crystallization. Protein crystals consists of molecules that are arranged three-dimensionally as arrays. Stacking of unit cells periodically, which are the smallest building block of crystals, constitute the whole crystal. In-house systems or the synchrotrons can be used next to perform X-ray diffraction measurements of the crystals, thus obtained. The intensity and position of diffraction spots can be used to determine the structure. It is further followed by data processing involving phase determination, structure refinement and validation (Fig. 2.2).

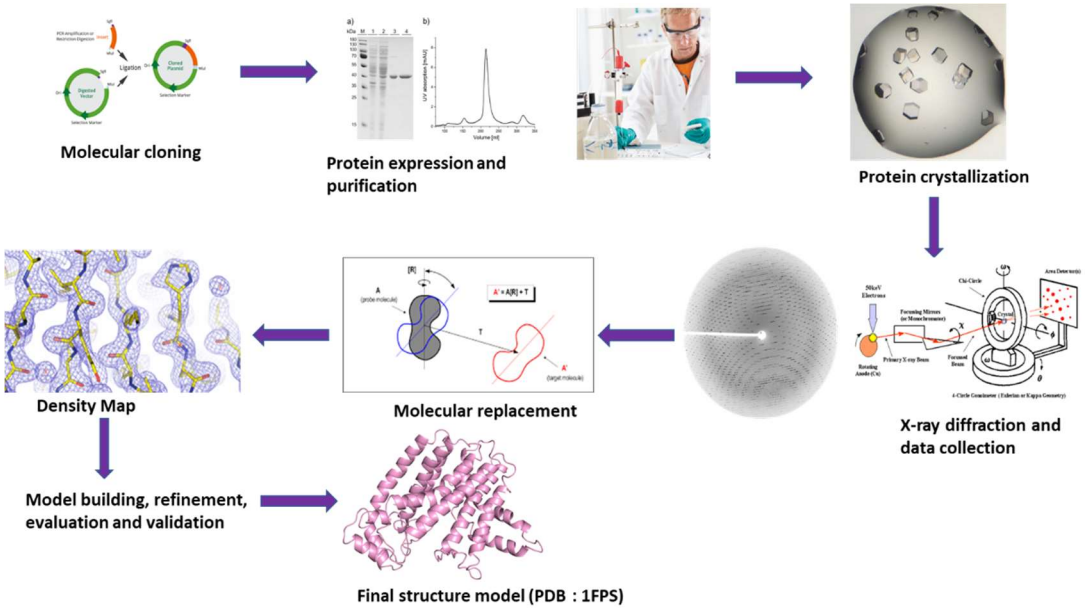


Fig.2.2. *Protein Structure determination. Major steps involved in three-dimensional structure determination by X-ray crystallography.*

Though due to advancements in the field of X-ray crystallography, the technique is used extensively to study the three-dimensional structure of macromolecules, there are a number of

limitations that still persist. These limitations include crystallizability of proteins due to the solubility issues, indeterminism of protein dynamics and indirect measurement of phases. Despite these limitations, X-ray crystallography is one of the predominant methods for determination of protein and nucleic acids structures.

2.5.1 Crystallization:

Obtaining well-ordered crystals is an important prerequisite for protein structure determination. Despite substantial progress in the development of completely automated robots and the types of crystallization screens, protein crystallization is a bottleneck in X-ray crystallography. An asymmetric unit is the initial entity of a crystal, which builds up the entire unit cell by undergoing symmetry operations like rotation, reflection, screw axis rotation, inversion and glide translation. Periodic repetition of a unit cell forms a crystal. The crystallization process is regulated by both thermodynamics and kinetics that cannot be controlled easily[72]. The protein of interest in a solution slowly concentrates to reach a metastable supersaturated solution followed by nucleation and crystal growth (Fig. 2.3A). As the mixture of reagents that aids in crystallization are not known priorly and neither can be deduced, testing of a number of parameters is required to identify the suitable conditions for crystallization. The macromolecular crystals are highly fragile and crystal mosaicity arises due to imperfections in the crystal lattice[73]. Thus, crystal growth is an important limiting step of structure determination.

There are different methods of crystallization. Earlier, batch method was employed mostly in which large volume of protein (> 1mL) was used with the precipitating agent to form a supersaturated solution. As the method requires huge volume of protein, it has been replaced mostly by another technique, vapor diffusion[74]. Vapor diffusion involves mixing a little volume (0.5-10 µL) of protein solution with the precipitant, which is equilibrated with a reservoir containing higher concentration of precipitant. Thus, the technique enables screening of a number of conditions for crystal optimization. There are a number of variations of vapor diffusion, out of which, sitting and hanging drops are widely used (Fig. 2.3 B,C)[75]. The sitting drop employs using robots for automated drop setting. In the present work, these two techniques have been mainly employed for crystallization. Commercial matrix screens are used to identify the initial conditions of crystallization of macromolecules[76]. These screens consist of cocktails of a wide range of precipitants, salts and pH that have been tested successful in the crystallization of other proteins[77]. Crystallization with these screens is mainly tested at room temperature or at lower temperatures (4 or 10 °C). A number of optimizations may be required

in order to obtain good quality crystals, for example, changing the protein concentration or buffer ionic strength, testing different concentrations of precipitants or salt, incubation temperature and screening various additives[78,79].

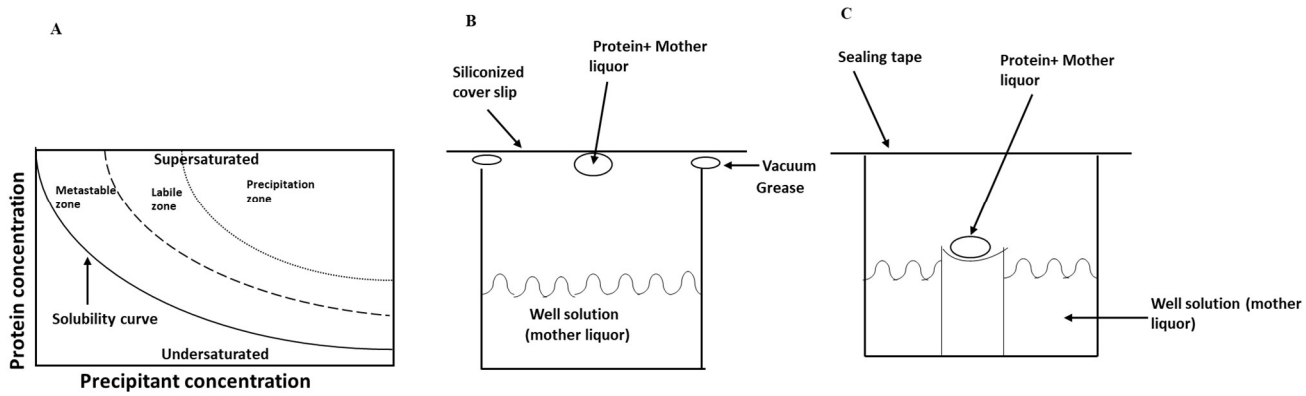


Fig.2.3. *Protein crystallization. (A) Phase diagram showing protein solubility as a function of the precipitant concentration. (B) Hanging drop vapor diffusion in which the drop is placed on a siliconized cover slip. (C) Sitting drop vapor diffusion in which the drop is placed in a well.*

2.5.2 Data collection and Processing:

Once good quality crystals are obtained, X-ray diffraction of the crystals is performed either using in-house system or synchrotron. The constituent atoms of a crystal scatter X-rays in various directions, however, the waves with the constructive interference forms a diffracted beam which are recorded on a detector as reflections or diffraction spots. All the atoms diffract the X-rays at the same diffraction angle which interfere to form a single reflection. Thus, each reflection constitutes details of all the atoms in the protein structure. This information from the diffraction pattern can be reconstructed to form the molecule image by employing numerous calculations.

There are two main types of X-ray radiation sources used for diffraction; laboratory X-rays or the in-house source and the synchrotrons. The laboratory X-ray sources have a rotating copper anode which generates monochromatic X-rays. Therefore, these sources have a limitation as they only produce X-rays of single wavelength. On the other hand, synchrotron generators can be tuned to different wavelengths and possess higher flux. Thus, they are used for data collection of high resolution. In a standard diffraction experiment, the crystal is mounted at the midpoint of the goniometer head of the diffractometer which positions the crystal precisely in the path of the X-ray beam. Monochromatic X-rays with a fine focused beam are implunged on the crystal, which scatter these X-rays. Different types of sources are used to record the

scattered X-rays. Initially, photographic film or an area detector, charge-coupled device (CCD), were used as detectors. However, now most of the in-house sources and synchrotrons employ the use of hybrid photon counting (HPC) detectors. These detectors are more sensitive and have a fast read out time ensuring faster data collection. The appropriate strategy of data collection depends on the crystal. Data acquisition depends on a number of factors including; the unit cell parameters, crystal symmetry and mosaicity, crystal orientation on the diffractometer, resolution limit and radiation damage. In the process of data collection, initially few frames are collected to delineate the optimal strategy by evaluating the initial orientation and rotation range. This helps in maximizing the completeness of the data.

Data collection is followed by the processing of the diffraction images. The diffraction data collected for a broad range of images is integrated providing information about miller indices (*hkl*) of the crystal planes and intensities of the waves scattered by these planes. Initially, a few frames are auto-indexed to predict few parameters of the crystal, like, crystal symmetry, unit cell parameters and crystal orientation relative to beam. This is followed by refinement of various parameters including, detector parameters (orientation, position and distortion), crystal parameters (orientation, unit cell and mosaic spread) and beam parameters (divergence and orientation). After refinement, integration of all the images is done. The integration provides an estimation of the intensity of a given reflection and its position on the image. The intensities of the spots detected are affected by various physical factors including changes in the incident radiation intensity or radiation damage of the crystal. This gives rise to differences in the scaling amongst the images. Data reduction, thus, place all the redundant symmetry-related reflections on a common scale. The comprehensive data quality can be evaluated by the parameters like signal-to-noise ratio, resolution, multiplicity, completeness and R_{merge}. Multiplicity factor includes average number of measurements for the equivalent reflections. The completeness of the data indicates recording of all the reflections. The R_{merge} factor measures the internal consistency of the data, which evaluates the agreement among various measurements of similar reflections (Equation 1). Different frames account for different measurements.

$$R_{merge} = \frac{\sum_{hkl}\sum_{i=1}^{n}|I_i(hkl)-\bar{I}(hkl)|}{\sum_{hkl}\sum_{i=1}^{n} I_i\,(hk\,)} \qquad (1)$$

Where I_i(hkl) is the intensity of a reflection, $\bar{I}$(hkl) is the average of the intensities of the reflections and n are the independent measurements of the intensity. As a rule of thumb, the Rmerge value for a good data should be less than 10% [80]. In the present work, XDS[81] is used

for indexing and integration and AIMLESS[82] is used for scaling the data. XDS aids in reduction of the diffraction data by using a set of programs which includes, XDS, XSCALE and XDSCONV. XDS predicts the crystal symmetry and draws a list of reflections with their integrated intensities. XDS employs eight software which work in succession. These programs include; XYCORR, INIT, COLSPOT, IDXREF, DEFPIX, XPLAN, INTEGRATE and CORRECT. XSCALE obtains the data from XDS and place them on a common scale, merge them and also gives information about the data quality and completeness. XDSCONV converts the data files of XDS or XSCALE to different formats required for structure determination by other software. It can produce test reflections which can be used to calculate R_{free}. AIMLESS is the scaling program. It scales all the measurements on a common level, calculates the average number of measurements for symmetry-related reflections and provides information about the parameters that can be used to assess the data quality.

2.5.3 Structure determination and the Phase problem:

The X-rays are scattered by the crystal in distinct directions according to the crystal lattice. The crystal structure determines both the phase and amplitude of each scattered ray. The intensity or amplitude of a reflection can be determined by the magnitude of the complex structure factor, F_{hkl}. Each scattered wave can be mathematically delineated as a Fourier series by the equation of structure factor (F_{hkl}). Thus, F_{hkl} represents the amplitude and phase of the wave diffracted by a crystal as a reflection hkl (Equation 2). It is the resultant of contributions of all the scattering atoms present in a unit cell (Equation 3).

$$F_{hkl} = |F_{hkl}|\, e^{\emptyset i} \qquad (2)$$

$$F_{hkl} = \sum_{j=1}^{n} f_j\, e^{2\pi i[hx+ky+lz]} \qquad (3)$$

Where n is the number of scattering atoms, f_j is the atomic scattering factor which is the measure of the X-ray scattering potential of each atom and depends on the direction of diffraction & the type of atom and x,y,z are the fractional atomic coordinates.

The scattered wave is characterized by both phase and amplitude. However, the measured intensities give information about the reflection amplitudes only and the information about the phases is lost in the diffraction experiment. Both the amplitudes and the phases of the structure factors are required to determine the electron density map by Fourier transformation (Equation 4). This loss of phases leads to the 'phase problem'. Three principal methods can be used to

determine the phases; direct method, molecular replacement and different experimental methods[83].

$$\rho(x,y,z) = \left({}^{1}/_{V}\right) \sum_{hkl} F_{hkl}\, e^{-2\pi i(hx+ky+lz)} \quad (4)$$

Where V is the volume of cell, F_{hkl} are the structure factors and h,k,l are the miller indices.

2.5.3.1 Molecular Replacement:

Molecular replacement is a computational method for solving the 'phase problem'. It calculates the phases by placing the structure (model) of a homologous protein in the unit cell. Thus, the homologous structure can be used as a search model to calculate initial set of phases for the target molecule with the unknown structure. These phases can be refined iteratively. The molecular replacement is carried out by orienting and positioning the search molecule in the unit cell of the target molecule in such a way that there is maximum overlay between the calculated diffraction pattern and the observed diffraction pattern. Therefore, the phase information is obtained from the generated model after positioning the model in the unit cell of the target protein. As the extent of similarity between the two protein structures correlates with their sequence identity, there should be approximately 40% sequence identity between the starting model and the protein with unknown structure.

As discussed above, the molecular replacement includes identification of the position and orientation of the starting model or the known protein structure with respect to the crystallographic axes of the target model or the unknown structure. This leads to solving of the problem in six dimensions (Equation 5).

$$X' = [R]X + T \quad (5)$$

Where X is the group of vectors indicating the positions of atoms in the target model and X' representing the transformed set, R is the rotation matrix and T is the translation matrix.

The classical molecular replacement method involves comparison of the Patterson function of the model and the experimental data. There is a good superimposition of the Patterson maps obtained from Patterson function (Equation 6) if the model is appropriately placed and oriented in the correct position in the unit cell. The Patterson map is represented as a vector map where the peaks correspond to the vectors between the atoms in the unit cell. The Patterson map can be a self-vector or an intramolecular vector which depends solely on the molecule orientation,

or it can be a cross-vector or an intermolecular vector that depends on both the position as well as the orientation of the molecule in the unit cell.

$$P\,(uvw) = \left({}^{1}/_{V}\right) \sum_{hkl} |F(hkl)|^2 \,.\, cos2\pi\,[hu + kv + lw] \quad (6)$$

Where u, v and w are the generic coordinates and are distances between a pair of atoms located at (x_1, y_1, z_1) and (x_2, y_2, z_2). Thus,

$$u = x_1 - x_2 \qquad v = y_1 - y_2 \qquad w = z_1 - z_2$$

The Patterson function splits the problem in determining the rotation matrix and the translation vector. This helps in calculation of the six variables that is otherwise a computationally extensive problem requiring huge calculations. As the self-vectors are only dependent on the molecule orientation, the rotation matrix can be determined by rotating and overlaying the model Patterson (which is self-convolution of the electron density) and the observed Patterson, (obtained from the experimental intensity). Thus, the rotation function can be calculated as a sum of the product of the two Patterson functions at each point (Equation 7).

$$F\,(R) = \int_r P_{cryst}\,(u)\, P_{self}\,(Ru)\, du \quad (7)$$

Where R is the rotation matrix, r is the integration radius and P_{self} and P_{cryst} are the calculated and experimental Patterson function, respectively.

The translation function involves identification of absolute position of the molecule by determining the translation vector (Equation 8). Assuming that the molecule has been correctly oriented in the cell, the intermolecular vectors of the molecule change upon translation. Thus, there will be a good agreement between the calculated cross-vectors Patterson functions from the model and the observed data if the molecules are in the correct position in the crystal.

$$T(t) = \int_v P_{cryst}\,(u)\, P_{cross}\,(ut)du \quad (8)$$

Where P_{cryst} is the experimental Patterson function, P_{cross} is the Patterson function obtained from the probe oriented in the experimental crystal, u is the cross-vector between two symmetry related molecules and t is the translation vector.

There are different programs used for molecular replacement; AMoRe[84], MOLREP[85] and Phaser[86]. All the structures in the present work were solved with the CCP4 Phaser module[87]. Though both AMoRe and MOLREP execute automation strategies, they lack scoring functions based on likelihood. Phaser has improved algorithms which use multivariate statistics and

maximum likelihood methods. For rotation and translation function in molecular replacement, Phaser has novel maximum likelihood algorithms. It consists of pruned tree search for multiple molecules and likelihood-based correction for anisotropy to solve complex molecular replacement problems.

2.5.4 Model Building, refinement and density improvement:

Phase estimation is done by identifying the coordinates of atoms in the structure in real space. The experimentally determined amplitudes are combined with the estimated phases to generate an electron density map. If there is close similarity between the initial estimated phases and the real phases of the model, and there is high quality experimental data, an electron density map can be interpreted for the complete macromolecule or significant structural elements. Model building improves the precision of the phases and thus leads to better and simply interpretable maps by determination of the atomic positions in the density maps.

The phases and thus quality of the electron density map are improved by employing solvent boundaries[88] and averaging over non-crystallographic (NCS) elements which are the characteristics of the molecule in the asymmetric unit[89]. These protocols help in improving reflection-to-parameter ratio. For modeling, two electron density maps are employed, that differ in the weighting parameters for the calculated structure factor (F_c) and the observed structure factor (F_o). For navigating model construction the 2 F_o - F_c map is used, however, to locate errors in the map the F_o - F_c map is used [90]. Programs like Coot [91] are used for viewing and building the model. The model refinement is done using programs like Refmac and Phenix[87], which have different algorithms to minimize the differences in the model and the observed data, however, retaining the geometries. Following a cycle of computational refinement, the model is improved by visually inspecting the map and fitting the model, accordingly. This cycle is done iteratively to improve the maps and thus the optimization of the model. A number of parameters like R-factors, atomic displacement parameters, electron density map correlation and model geometry are used to assess the quality of the optimized structure.

2.5.4.1 R-factor:

The R-factor is the measure of agreement between the model and the electron density map in real space. Corollary, it provides information on the agreement between the observed structure factors from the diffraction data (F_o) and the calculated ones from model (F_c) [92] (Equation 9).

$$R = \frac{\sum||F_o|-|F_c||}{\sum|F_o|} \quad (9)$$

The R-factor is not an independent parameter to assess the quality of the model as the model optimisation involves minimizing the difference between F_o and F_c and thus the R-factor. Thus, it is prone to biasness due to over fitting of the model to the data[93]. Hence, to overcome this problem, another parameter R_{free} was introduced[92]. R_{free} is the R-factor calculated for a set of 'free' randomly chosen reflections which are excluded from data used for the model refinement procedure. A default value of 5% of reflections are chosen for this purpose. The set of R_{free} should be maintained the same over the whole refinement process in order to avoid corrupting the "free" data set from model bias. The R_{free} value reflects the structure quality and if the model is getting overfitted, then there is an increase in the R_{free} value during the refinement cycles[94]. The R value for the working set of reflections is called R_{work}. The acceptable values of R_{work} and R_{free} are dependent on the resolution and usually R_{free} is higher by about 20% than R_{work}. Additionally, there is a high correlation between the R_{free} value and the phase accuracy of the model[95]. A large difference between R_{work} and R_{free} indicates over fitting of the model.

2.5.4.2 B-factors or atomic displacement parameters (ADP):

B-factors or the temperature factors are the thermal motion and displacement of atoms in the crystal, modelled either in the form of a displacement sphere (isotropic model) or ellipsoid (anisotropic model)[96]. This is known as the atomic displacement parameter (ADP) and defined as the sum total of thermal motion of individual atoms.

At the atomic resolution, complete anisotropic model is used in which ADP is defined by six parameters. Therefore, the complete description of every atom requires nine parameters, including three position coordinates and six B-factor parameters[97]. However, at low resolution, the data-to-parameters ratio may be not sufficient for the refinement of the macromolecule. Thus, ADP refinement is augmented with TLS (Translation, Libration, Screw) parametrization for refinement at low resolution [98].

2.5.4.3 Electron density map:

During structure refinement, different versions of density maps are calculated and used to interpret the structure. Amongst these versions, the maximum likelihood weighted maps are the most informative and used ones with least biasness[99]. These maps are with model bias correction ($2mF_o$-DF_c) and also a difference map (mF_o-DF_c) to point errors in the model, where D is the Luzzati coefficient and m is the figure of merit. They are generally used during model

building done manually and provide evidence for structural details, for example, ligands, in a model. The maps exhibit isosurfaces which contour at defined electron density levels generally expressed in rmsd units. Conventionally, for $2mF_o$-DF_c, values more than +1.0 rmsd are considered good indicators of presence of a structural detail. For mF_o-DF_c, rmsd values above +3.0 and below -3.0 indicate fragments of map not explained by the model or does not define a structural detail, respectively. These rmsd values are used generally while model inspection in Coot. However, the maps with varied rmsd values as the suggested levels, may also possess useful information as the map levels are not necessarily comparable between datasets or even distinct regions[100].

2.5.4.4 Geometrical restraints and constraints in refinement:

Prior information about the geometry of building blocks (amino acids and nucleotides), non-covalent interatomic distances like van der Waals interactions and hydrogen bonds, and ligands is required for the refinement of a correct macromolecule model. Thus, 'restrained refinement' is used for modelling and refinement of majority of the macromolecular structures. Restrained refinement allows optimization of atomic positions using the information from the experimental data, allowed geometry and non-bonding interactions[101]. Stereochemical restraints are the principal restraints that are used in the refinement of macromolecular structure.

During refinement, applying constraints and restraints to the atoms is required to improve the data-to-parameter ratio. Constraints are mathematical equations that are related to two or more parameters. It reduces the number of independent parameters to be refined by assigning fixed numerical values to certain parameters.

2.5.4.5 Bulk solvent correction:

Protein crystals consist of approximately 30% to 50% solvent of total crystal volume, majority of which is disordered in the solvent channels present between the protein molecules in the crystal lattice[102]. Therefore, the electron density of the protein molecules with an average value of 0.43 e/Å^3, are enclosed by a continuous solvent with electron density ranging from 0.33 e/Å^3 [103]. At low resolution, the contribution of solvent to structure factors is high, thus, bulk solvent correction is required for an improved electron density map. It can be corrected by the volume filled by the solvent that is identified by defining a solvent-accessible volume outside the van der Waals exclusion zone of protein molecule. The optimal value for the average solvent density may be estimated by calculating the minimum value of,

$$\sum(|F_o| - |F_c(total)|)^2 \quad (10)$$

$$\text{Where,} \qquad F_c(total) = F_c(protein) + K_{sol}\, F_c(solvent)\, exp\left(\frac{B_{sol}}{4d^2}\right) \quad (11)$$

Where, K_{sol} is a scale factor, F_c (solvent) is a Fourier transform of a binary function M (solvent mask), value of which is 1 in the solvent and 0 outside and is summed over low resolution reflections[103], B_{sol} is an artificial large temperature factor applied to a flat solvent density.

2.5.5 Validation:

An important part of structure determination is validation of macromolecular structure[104]. It is also an important step during structure refinement to assess the progress in refinement. Structure refinement should be performed till all the structural parameters are refined to their optimal value, in commensurate with the data quality. The objective of the validation is to identify the errors related present in the structure. There are three levels of model validation including agreement with the known parameters, quality of the data fitting and evaluation of non-bonded contacts. R and Rfree values are good indicators of the quality of data fitting. The other parameters discussed previously like B-factors and average positional error can also be used to assess the quality of the data in detail. The most crucial parameters that can be used to evaluate the protein model include deviations from the ideal bond lengths and bond angles, Ramachandran plot, close contact analysis, and computing the side-chain torsion angles. There should be analysis of the solvent network making sure that the water molecules form hydrogen bonds with the amino acids and maintain an appropriate distance with the surrounding atoms. Either web servers like Molprobity[105] or the standalone programs can be used to perform these tests that provide comprehensive information to validate the correctness of the model.

2.6 Computational methods:

2.6.1 Molecular Docking:

The docking techniques aid in prediction of the best binding mode of a ligand and its macromolecular partner, in the absence of experimentally determined structure of protein-ligand complex. The docking technique involves generation of a number of possible conformations of the ligand in the binding site of the protein. Each of these conformers are scored based on the interactions they make with the protein[106]. The ligand conformational sampling process efficiently explores the conformational space defined by the free energy landscape. This energy is estimated by the scoring function. The scoring function must connect

the native-bound conformation to the overall minima of the energy hypersurface. In the present work, Glide module of Schrödinger employing extra precision (XP) was used for docking[107].

The scoring function generally helps in selecting the most probable binding modes from the conformations generated by the sampling technique. There are mainly three kinds of scoring functions. The force-field based scoring function, which estimates the potential energy of a system with both bonded (intramolecular) and nonbonded (intermolecular) components. Next is the empirical scoring function, which calculates the summation of different energies like electrostatic, van der Waals, hydrogen bond, hydrophobicity & entropy and gives binding affinity data by least square fit[108]. The third type includes the knowledge-based scoring function that involves calculating the frequencies of protein-ligand atom pair contacts which are converted in a component of energy. The final score for particular docked ligand conformer is represented as the summation of these energy components.

2.6.2 Molecular dynamics simulation:

Molecular dynamics simulation is a computational technique used to simulate the dynamical nature of a molecule as a function of time in which all the entities in the simulation box like, protein, ligand, and explicit water molecules, are treated as flexible[109]. Molecular dynamics simulation includes calculation of positions and velocities of atoms in an N-particle system by using the classical Newton's equation of motion,

$$F_i(t) = m_i \frac{d^2 r_i}{dt^2} \quad (12)$$

Where, F_i is the force which acts on the i^{th} atom at time t, m_i is the atomic mass and r_i is the position of the atom at time t. The force on each atom can be calculated using potential energy based on the N-particles interactions. The force can be calculated as,

$$F_i(t) = -\frac{\partial V_{(r_i)}}{\partial r_i} \quad (13)$$

Where, $V_{(ri)}$ is the potential energy of the system. Therefore, this potential energy is a function of the atomic coordinates. This force can be used to compute the position, velocity and acceleration of each atom by numerically solving Newton's motion equation using an integration algorithm. Iteration of this process after discrete time intervals gives the progression of the individual particle motions as a function of time. Thus, molecular dynamics simulation produces an ensemble of consecutive conformations in order to identify the energy landscape of a system. In this thesis, simulation was carried out using Desmond program of Maestro-

Desmond Interoperability tools from Schrödinger[110]. The first step in any simulation study involves identification of appropriate force field, which is used to calculate the interatomic forces of the system. A force field determines a set of parameters and potential functions to define the interatomic interactions and the system energy as a function of atomic coordinates. There exist a number of force fields which depend on the energy function and the type of strategy for parametrization. For the present study, inbuilt force field of Schrödinger, Optimized Potentials for Liquid Simulations (OPLS3e)[111], is used. Following steps were performed to conduct the simulation runs:

2.6.2.1 System setup:

To initiate a simulation, the preliminary conformation of the system is generated by placing the macromolecule or the particles inside a box randomly which represents the system volume. The bonded and non-bonded parameters are assigned to the atoms or the particles by using the appropriate force field. This is used at a later stage for calculating the potential energy. The system is solvated by addition of water and ions in the box in order to mimic the biological conditions.

2.6.2.2 Energy minimization:

The geometrical defects of the structure such as steric clashes between the atoms, incorrect bond angles and bond distances, are rectified through minimizing the energy of the system. There could be difference between the initial state of the system and the thermodynamic equilibrium conditions of the simulation. Thus, a NVT (normal volume temperature) or NPT (normal pressure temperature) equilibration is performed by coupling system to thermostat and barostat, respectively, in order to bring the system to equilibrium.

2.6.2.3 Production Run:

After equilibration, the system evolves in the production run through the conformational space. For computing the position and velocity of the atoms, the time steps must be adequate to report the time-scale of the event of interest. Furthermore, the duration of the production run must be sufficient to warrant sufficient sampling of the phase space.

2.6.3 Statistical Coupling Analysis:

Structural and functional constraints of proteins, reflected in their amino acid sequence, can be unravelled from their evolutionary history[112]. As there is substantial enlargement of the

sequence databases, statistical analysis of this evolutionary record can be done to decode the sequence information [113–116]. One such technique is statistical coupling analysis (SCA) which makes use of multiple sequence alignment of proteins of a family to identify group of residues that coevolve. These groups of coevolving residues are known as "sectors". Thus, it can be hypothesized that co-evolving amino acids of different "sectors" are crucial in regulating varied functional and structural aspects of a protein. Sectors represent the structural basis of functions like allostery[117,118], signal transmission[113,115,119,120], cumulative dynamics of catalytic reactions[118] and protein adaptability[121].

2.6.3.1 Workflow of SCA:

The first step involves generating an alignment of homologous sequences. As SCA focuses on the conserved characteristics of protein sequences, it is robust to changes in the quality of the alignment, however, depends on the diversity and depth of homologous sequences. The sequence diversity within the alignment should be moderate as identical sequences will not provide information about the amino acid covariance. The alignment is pre-processed in which the sequences and positions with a number of gaps are removed.

Assuming a multiple sequence alignment represented as an N X L matrix A where a_{ki} is the amino acid in the k^{th} sequence at position i, a numeric matrix $\bar{E}$ is constructed as,

$$\bar{E}_{ki} = \frac{\varphi_i(a_{ki}) f_i(a_{ki})}{\sqrt{\sum_{b \neq gap} \varphi_i^2(b) f_i^2(b)}} \qquad \text{if, } a_{ki} \neq \text{gap} \qquad (14)$$

Where, φ_i(a) is the positional weight and f_i (a) is the frequency of amino acid a in the column i of the alignment. This positional weight is calculated as,

$$\varphi(a) = \log\left[\frac{f_i(a)}{1-f_i(a)}\frac{1-q(a)}{q(a)}\right] \qquad (15)$$

Where, q(a) is the background frequency of amino acid a in a huge protein database.

Then a covariance matrix, C_{ij}, is built with correlations between residues at i^{th} and j^{th} position. Each element of this covariance matrix is multiplied by positional weights to yield SCA matrix (Equation 16).

$$\bar{C}_{ij} = \varphi_i \varphi_j C_{ij} \quad (16)$$

The SCA matrix is the covariance matrix associated with $\bar{E}$ (Equation 17).

$$\bar{C}_{ij} = \left|\frac{1}{N}\sum_k \bar{\mathrm{E}}_{ki}\bar{\mathrm{E}}_{kj} - \frac{1}{N^2}\sum_{k,l} \bar{\mathrm{E}}_{ki}\bar{\mathrm{E}}_{lj}\right| \qquad (17)$$

Eventually, sectors are identified by searching the positions where the elements of the top eigenvectors of the SCA matrix reach beyond a given threshold.

The correlation between the amino acids raises questions on the representation of amino acid positions as the fundamental unit of protein. Thus, a transformation should be done that re-parameterizes the protein into sets of correlated positions that are mainly independent from each other. The first and the most important step in this procedure is spectral or eigenvalue decomposition. According to this decomposition, the $\bar{C}_{ij}$ matrix is a product of three matrices as,

$$\bar{C} = \bar{V}\bar{\Delta}\bar{V}^T \qquad (18)$$

Where, $\bar{\Delta}$ is an L X L diagonal matrix with eigenvalues and $\bar{V}$ is an L X L matrix in which the columns have the associated eigenvectors. The quantity of information or the variance captured in $\bar{C}_{ij}$ is given by each eigenvalue and the associated eigenvector $\bar{V}$ indicates the weights for integrating sequence positions into eigenmodes or transformed variables.

Thus, evaluation of $\bar{C}_{ij}$ includes spectral decomposition of $\bar{C}_{ij}$, determining significant eigenvalues, independent component analysis (ICA) that involves transformation of the top eigenvectors and studying the arrangement of residue contributions across the independent components (ICs).

Chapter 3

Structural studies on SaSQS1 to identify the conformational dynamics at the active site

3.1 Background:

As discussed in *Chapter 1*, in Indian sandalwood (*Santalum album*) two isoforms of sesquisabinene synthases, sesquisabinene synthase 1(SaSQS1) and sesquisabinene synthase 2(SaSQS2) have been reported[19]. Comparative studies of the expression and kinetic parameters of both isoforms showed that SaSQS1 is kinetically more active than its isoform. GC-MS analysis of SaSQS1 indicated that the enzyme converts FPP to sesquisabinene (>93%), β-sesquiphellandrene (~5%) and unidentified metabolite (~2%).

SaSQS1 follows the canonical catalytic mechanism with the formation of coordination of FPP diphosphate moiety with Mg^{2+} ions. The Mg^{2+} triad also coordinates with the DDXXD and NSE motif of the protein to position the FPP in the hydrophobic active site for further catalysis. FPP ionizes to form farnesyl cation followed by formation of a neutral intermediate nerolidyl diphosphate (Fig.3.1). This intermediate undergoes further ionization and rearrangements (1,6 closure) to form bisabolyl cation that forms a half-chair conformer, homobisabolyl cation by undergoing 1,2- hydride shift. Finally, after transannular carbon-carbon bond formation and proton quenching by pyrophosphate ion, bicyclic sesquisabinene is formed. However, the structural basis of this conversion and product specificity of SaSQ1 is not known.

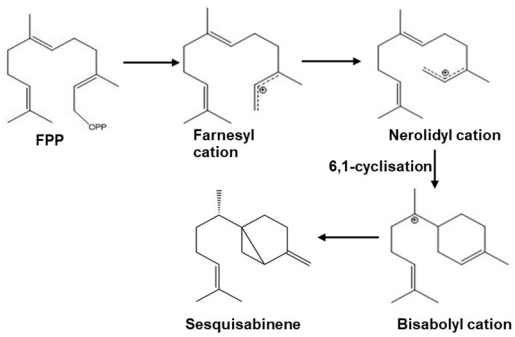


Fig. 3.1. *Proposed cyclisation mechanism for conversion of FPP to sesquisabinene involving generation of carbocation intermediates.*

Furthermore, comparative analysis of *apo* and ligand bound form of few sesquiterpene synthases have shown conformational deviations upon ligand binding[122,123]. These deviations were stated to be responsible for the open and closed state of the enzyme. However, mode of conformational changes and the "states" of the enzymes were not consistent across the reported structures. Therefore, we set out to study the architecture and structural basis of mechanism of action SaSQS1 by elucidating its structure. Also, comparative analysis of structures of SaSQS1 with other sesquiterpene synthases were performed to understand the role of open and closed state of SaSQS1 on its function.

3.2 Methodology:

3.2.1 Cloning of SaSQS1:

SaSQS1 construct was designed based on the disorder prediction from PHYRE[2] [124]. The full-length gene from the interface of heartwood and sapwood of Indian sandalwood was cloned in the vector pET28a. The gene (GenBank accession number: KJ665776.1) encoding 17-566 amino acids was cloned in pOPINss vector using USER (Uracil-Specific Excision Reagent) cloning[67] (Fig. 3.2A,B). The vector contains N-terminal 6x His and SUMO tag followed by TEV protease cleavage site.

A **B**

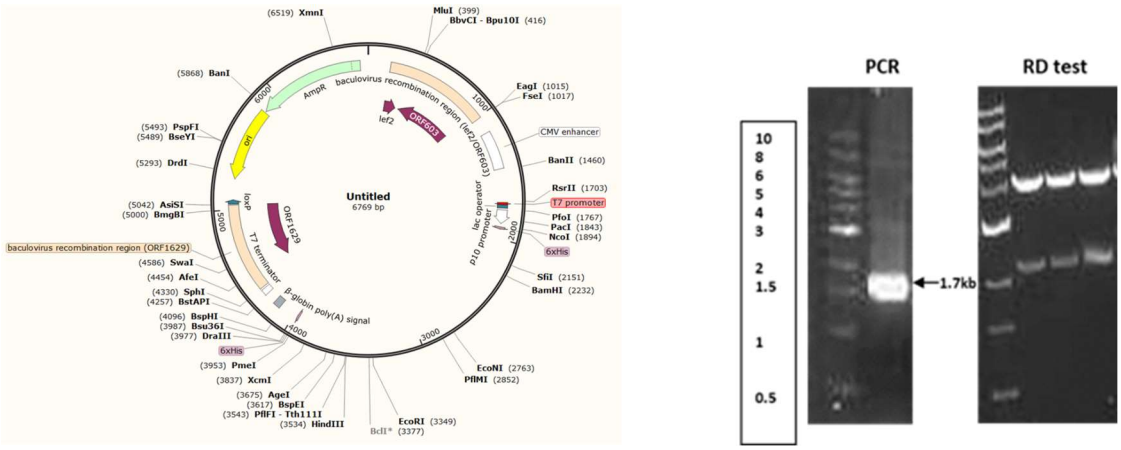


Fig. 3.2. *(A)* ***Sequence map of SaSQS1 in vector pOPINss.*** *The gene is cloned under T7 promoter highlighted in red color. The image is generated using the software (GSL Biotech). (B)* ***Agarose gel image of PCR and restriction digestion (RD) test of SaSQS1.*** *The gene size is 1.7kb and the RD test matches with the predicted pattern.*

3.2.2 Expression and Purification:

The protein expression was tested in different *Escherichia coli* strains and the optimal expression was observed in C41 strain, transformed with the plasmid. The cells were grown at

310.15 K till the OD_{600} value of 0.5 and gene expression was induced for 16 hours at 293.15 K using 0.5 mM Isopropyl β-D-thiogalactopyranoside (IPTG). The bacterial cell pellet was lysed by sonication in buffer containing 50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 10 mM imidazole, 5% glycerol, 2 mM β-Merkaptoethanol, 2 mM phenylmethylsulfonyl fluoride (PMSF), 2 mM benzamidine, 20 mM sucrose and lysozyme. The lysate was centrifuged at 20,000 rpm for 30 minutes and loaded on the Ni-NTA resin (GE Healthcare Biosciences, USA) for binding. The bound protein was eluted with 100 mM imidazole (Fig. 3.3A). The protein was found to be in the soluble fraction. The N-terminal 6xHis-SUMO tag was removed by incubating the protein, overnight, with TEV (Tobacco Etch Virus) protease under dialyzing conditions at 10 ^{0}C. To remove the TEV protease and cleaved 6xHis-SUMO tag, the protein was further subjected to inverse Ni-NTA affinity purification. The untagged protein was further purified by gel filtration chromatography using Hiprep26/60 sephacryl S-300 with buffer containing 10 mM Tris-HCl (pH 8.0), 500 mM NaCl, 5% glycerol and 2 mM DTT. The purity of the protein was assessed with SDS-PAGE. The protein was eluted in expected fractions (Fig 3.3B). Purified protein was used for crystallization and other biochemical assays.

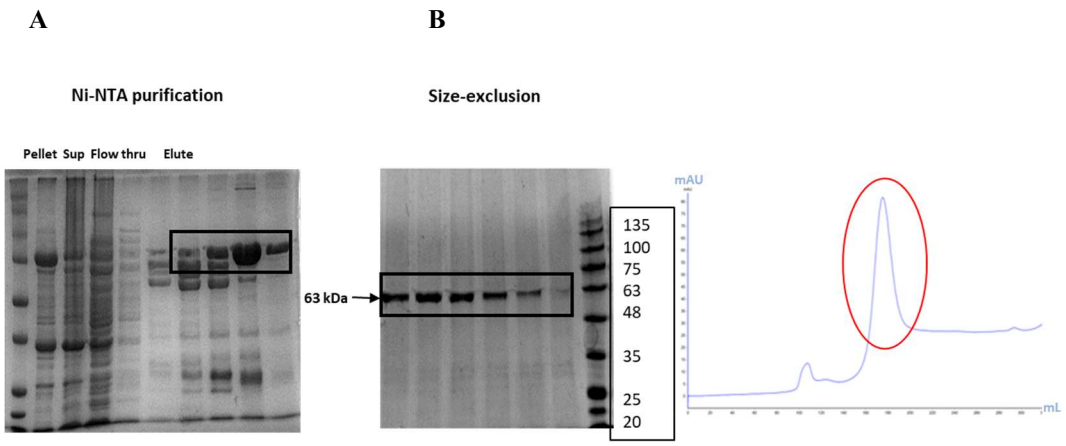


Fig. 3.3. *(A) SDS gel image of different fractions of affinity chromatography. (B) Elution fractions of SaSQS1 after gel filtration with the size-exclusion chromatogram.*

3.2.3 Crystallization and Data collection:

The protein was dialyzed overnight with buffer comprising of 20 mM Tris-HCl (pH 8), 150 mM NaCl and 2 mM DTT and was concentrated to 15 mg/ml for crystallization trials. Crystals were obtained with sitting drop method[74] at 20 ^{0}C using 100 nL of both protein and precipitating solution consisting of 0.2 M Magnesium Formate equilibrated against 50 µL of reservoir solution (Fig. 3.4). Crystals were observed within 4 days. For diffraction, the crystals were soaked in cryoprotectant containing reservoir buffer with 28% glycerol and were flash cooled

with nitrogen stream at 195 K. The data set was collected at XRD1 beamline facility of ELETTRA synchrotron. Co-crystallization trials of SaSQS1 with varying concentrations of FPP analog, Thio FPP, were carried out and the crystals thus obtained were diffracted at XRD1 beamline facility of ELETTRA synchrotron and PX-BL21 Indus-2 beamline of RRCAT synchrotron.

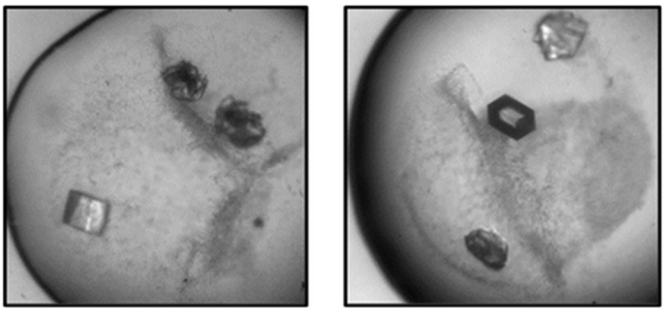


Fig. 3.4. *Crystals of SaSQS1.*

3.2.4 Data analysis and Structure determination:

The data was integrated with XDS[81] and scaled with Aimless[82] of CCP4 suite. The structure was solved using molecular replacement technique with PHASER[125] program by employing isoprene synthase (PDB ID: 3N0F) as a search model. Iterative rounds of manual model building was performed with COOT [91] and the structure was refined using PHENIX [126]. The R and R_{free} values are 0.197 and 0.228, respectively. Diffraction data collection and refinement statistics are summarized in Table 3.1.

Data Collection	
Space group	C121
Resolution (Å)	45.16-2.2
Cell parameters	
a, b, c (Å)	141.059, 85.821, 53.399
α, β, γ (°)	90, 97.56, 90
Observed reflections	120143 (10529)
Unique reflections	32042 (2760)
I/σI	9.3 (2)
R_{merge}	0.07 (0.57)
Completeness (%)	99.8 (99.8)
Multiplicity	3.7 (3.8)
$CC_{1/2}$	0.99 (0.78)
Refinement	
Resolution (Å)	36.57-2.2
Number of reflections	
Working set	32010
Test set	2890
R_{work}/R_{free}	0.197/0.228
Number of atoms	
Protein	4092
Water	48
Other	2
Mean B-factors (Å^2)	
Protein atoms	56.70
Water atoms	49.63
RMSD from Ideal values	
Bond length (Å)	0.021
Bond angles (°)	1.626
Ramachandran plot Statistics	
Preferred (%)	99
Allowed (%)	1

Table 3.1. *Data collection and refinement statistics of SaSQS1 apo form. Values in parenthesis are for the highest resolution shell.*

3.3 Results and Discussion:

3.3.1 Conformational deviations in Sesquiterpene Synthases structures:

To determine the structural basis of SaSQS1 catalyzed sesquisabinene formation, we elucidated the crystal structure of SaSQS1 at 2.2Å resolution (PDB ID: 6K16). Simultaneously other groups also reported the SaSQS1, determined under various conditions. The structure

elucidated by us is by and large identical to the other reported SaSQS1 structures[127](PDB ID: 6A1I, 6A3X, 6A1E and 6A1D). SaSQS1 exhibits classical αβ helical domain architecture, characteristic of plant class I terpenoid cyclase, comprising of 27 helices (Fig. 3.5A). The first 13 helices (α1- α13) form the N-terminal helical β domain whereas the rest of the 14 helices (α'1- α'14) form the catalytic C-terminal α domain (Fig. 3.5A). Amongst these 14 helices, α'1, α'2, α'7, α'8, α'10, α'11, α'13 form core of the substrate binding pocket. The conserved metal binding motifs, DDXXD and NSE/DTE, are located at α'2 and α'10, respectively. These motifs coordinate to Mg^{2+} ions, which are suggested to bind in presence of the substrate or diphosphate containing ligands. However, in our structure we do observe the density for one of the Mg^{2+} ions close to the DDXXD motif, despite the absence of bound ligand. The NSE/DTE motif is partially disordered. The conserved catalytic mechanism of sesquiterpene synthases, proposed on the basis of available structures and biochemical data, seems to be applicable for SaSQS1. In this cascade of reactions, Mg^{2+} ions play critical role by forming an enzyme - Mg^{2+} - ligand ternary complex. In SaSQS1, D316 and D320 residues of DDXXD motif coordinate with the first and last Mg^{2+} ions while the second Mg^{2+} ion supposedly binds to the NSE/DTE motif. However, in the reported structures, this ion is not observed in the proximity of N460 and T464 of the NSE/DTE motif [127] (Fig.3.5B). Coordination of ligand with Mg^{2+} ions is followed by its ionization leading to production of allylic cation which undergoes cyclization and hydride transfer. Finally, the reaction terminates with quenching of the positive charge which can be either due to deprotonation or by capture of a water molecule.

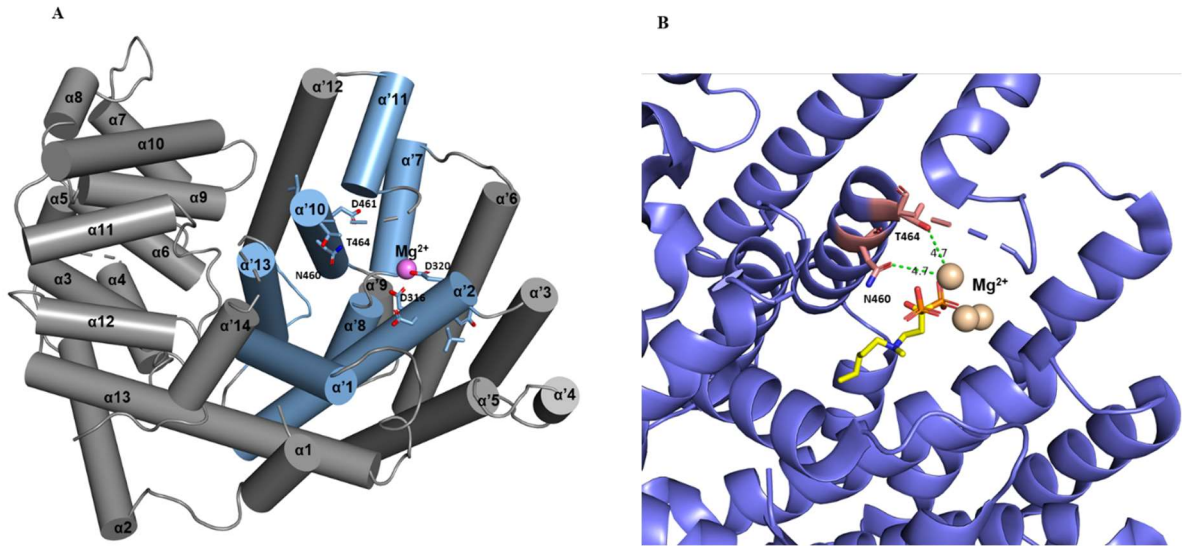


Fig. 3.5. Structure of SaSQS1. *(A)Overall structure of SaSQS1 (PDB ID:6K16). The core of the binding pocket colored in blue and the Mg^{2+} ion is shown as sphere. The conserved motifs, DDXXD and NSE/DTE, are highlighted at helices α'2 and α'10, respectively. (B) SaSQS1 bound with ibandronate represented with sticks*

(PDB ID: 6O9P). The residues of the NSE motif which coordinate with the Mg^{2+} ion is not found to be in close proximity with the ion.

Although the reported structures of SaSQS1 are by and large identical to our structure (PDB ID: 6K16) (Fig. 3.6A), for the apo (PDB ID: 6O9Q) and ibandronate bound (PDB ID: 6O9P) structures considerable conformational deviations are seen at the regions between the residues 270-285 and 315-330 (Fig. 3.6C). Interestingly, these regions harbor the DDXXD and the RXR(G) motifs. However, the reason behind these conformational deviations is not clear from the structures.

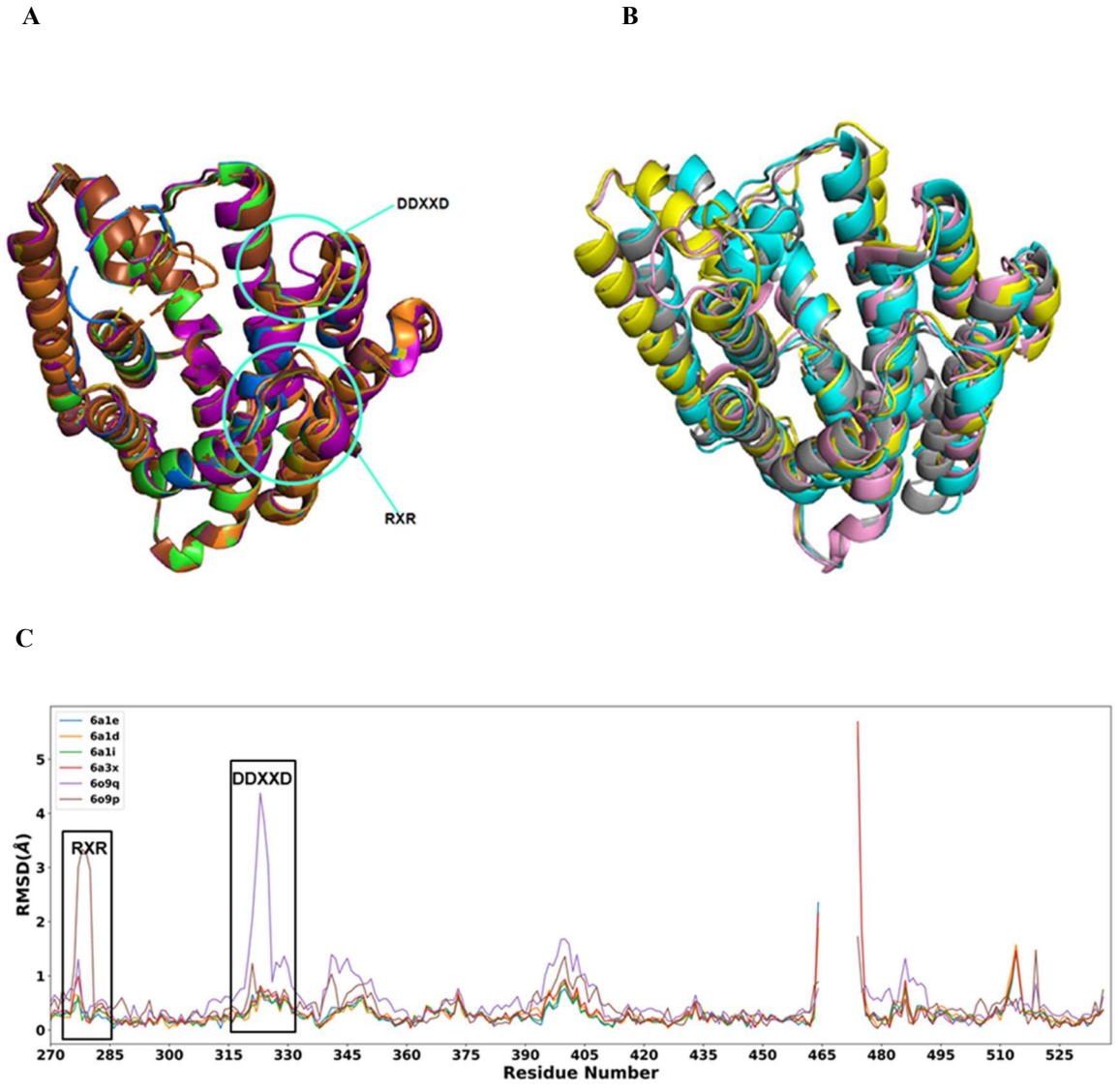


Fig. 3.6. Comparison of available structures of SaSQS1*. (A) Superposition of α domain of available SaSQS1 structures following coloring scheme: PDB ID: 6K16- brown; PDB ID: 6A1E- green; PDB ID: 6A1I- blue; PDB ID: 6A1D- red; PDB ID: 6A3X- olive; PDB ID: 6O9P- orange; PDB ID: 6O9Q- purple. (B) Structural comparison of α domain of SaSQS1 (grey) (PDB ID: 6K16) with other sesquiterpene synthases like 5-epi-aristolochene synthase (cyan) (PDB ID: 5IK0), AaBOS (PDB ID: 4FJQ) (pink) and $AaBOS^{M2}$ (yellow) (PDB ID:*

4GAX). (C) Overlay of plots of Cα RMSDs between PDB ID:6K16 and other SaSQS1 structures. The conserved RXR and DDXXD motifs are highlighted with boxes.

3.3.2 Co-crystallization trials of SaSQS1 with substrate analog:

Sesquiterpene synthase structures have been elucidated with their substrate analogs in order to study the interaction of the ligand with the enzyme and thus its mechanism of action[128–131]. These methods are based on incubation of protein with the analog. Thus, co-crystallization was tried for SaSQS1. The protein with the final concentration of 15 mg/ml and different concentrations of substrate analog of FPP, Thio FPP, were incubated for 1-2 hours and then subjected to crystallization trials. The different concentrations of Thio FPP ranged from 4-20 mM. Crystals were observed in the conditions highlighted in Table 3.2.

Crystallization condition	Crystals observed
0.2M Cacl2, 0.1M Sodium Acetate (pH-4.6), 20% Isopropanol	
0.2 M Sodium Acetate, 0.1M Sodium Citrate (pH-5.5), 10%PEG 4000	
0.1M MgCl2, 0.1M HEPES (pH-7.5),10%PEG 4000	
0.1M Mgcl2, 0.1M MES (pH-6), 8% PEG 6000	
0.1M Sodium Citrate (pH-5.5), 15% PEG 6000	
0.1M Sodium dihydrogen phosphate (pH-6.5), 12% PEG 8000	
0.1M Magnesium Acetate, 0.1M MES(pH-6.5), 10% PEG 10000	
0.1M Sodium Acetate (pH- 5), 1M Ammonium sulphate	
0.2M MgCl2, 0.1M HEPES(pH 7.5),30%PEG 400	

Table 3.2. *Crystals of SaSQS1 with Thio FPP observed in different crystallization conditions.*

The crystals with good quality were diffracted and data was collected. However, no density was observed for the ligand in the binding pocket.

3.3.3 Definition of Open and closed states of SaSQS1 structure are not absolute:

Based on the structures of *apo* and ligand bound forms of SaSQS1, 5-epi-aristolochene synthase, Bornyl diphosphate synthase, Isoprene synthase, it was proposed that these synthases

exhibit open and closed states upon ligand binding[122,123,127,132]. These states are attributed to the conformational changes in the RXR loop leading to the interaction between the first Arg and second Asp of RXR and DDXXD motifs, respectively. Therefore, with reference to this particular interaction it was hypothesized that sesquiterpene synthases exhibit open and closed conformation depending upon their ligand bound state. In SaSQS1, R279 of RXR motif changes the conformation to interact with D317 of DDXXD motif to adopt what is known as the closed state of the enzyme [127] (Fig. 3.7A,B). This conformational change was suggested to be essential for protecting the reactive carbocation intermediates from water mediated nucleophilic attack.

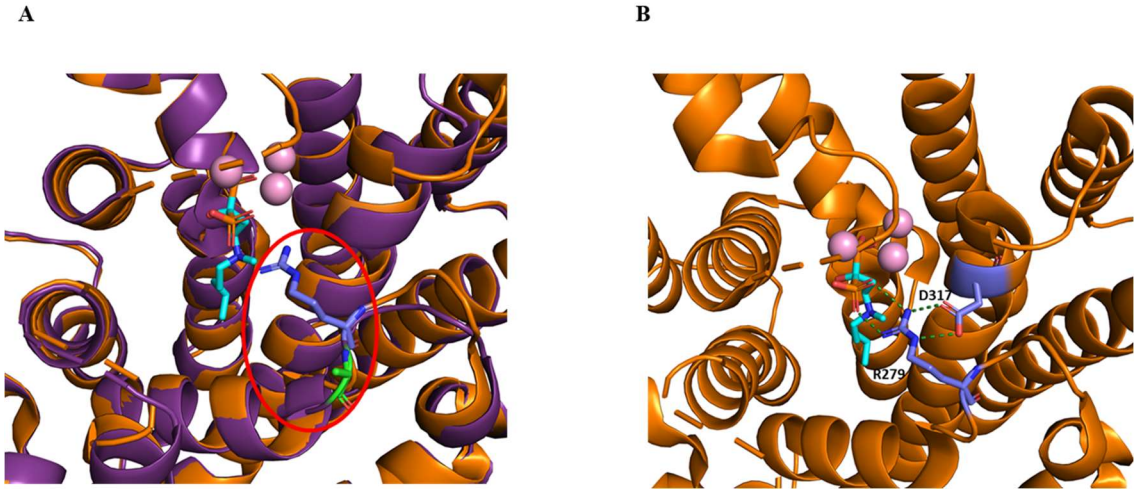


Fig. 3.7. Open and closed state of SaSQS1. *(A)Superposition of apo (PDB ID: 6O9Q) and ligand bound (PDB ID: 6O9P) forms of SaSQS1 colored purple and orange, respectively, showing change in R279 conformation. (B) Biphosphate-R279-D317 H-bond network in ligand bound SaSQS1 (PDB ID: 6O9P), hypothesized to be responsible for active site closure. The ligand & Mg^{2+} ions are shown in ball and stick.*

However, comparative analysis of structures of α-bisabolol synthase (AaBOS), *wild* type and *M2* mutant ($AaBOS^{M2}$) [133], show closed conformations in their ligand free forms (Fig. 3.8A). On the other hand, in α-bisabolene synthase, the arginine of the RXR motif is observed in the closed state irrespective of the presence of ligand in the binding pocket (Fig. 3.8B)[129]. Interestingly, the corresponding arginine in the ligand free form of Abietadiene synthase, which is a diterpene synthase, also exhibits the closed conformation (Fig. 3.8C). However, mutational analysis on Abietadiene synthase suggests that this arginine has modest effect on the catalytic efficiency of the enzyme[134]. Furthermore, other reported structures of SaSQS1 also suggest that this transition may have little influence on the catalytic efficiency of the enzyme (Fig. 3.8D). Thus, from comparative analysis of all the available structures of SaSQS1 with structures of other sesquiterpene synthase, we can conclude that the open and closed states of the enzyme,

as defined with reference to the interaction of the arginine of RXR motif, need not be necessarily dependent on the presence of the ligand in the catalytic site. Also, the earlier proposed open and closed conformational states might have minimal influence on the catalytic function of the enzyme. However, the comparative structural analysis of SaSQS1 with its other forms (Fig. 3.6A) and also with its homologs (Fig. 3.6B), suggest that the catalytic pocket is highly dynamic and could play some role in other aspects of the enzyme such as product specificity.

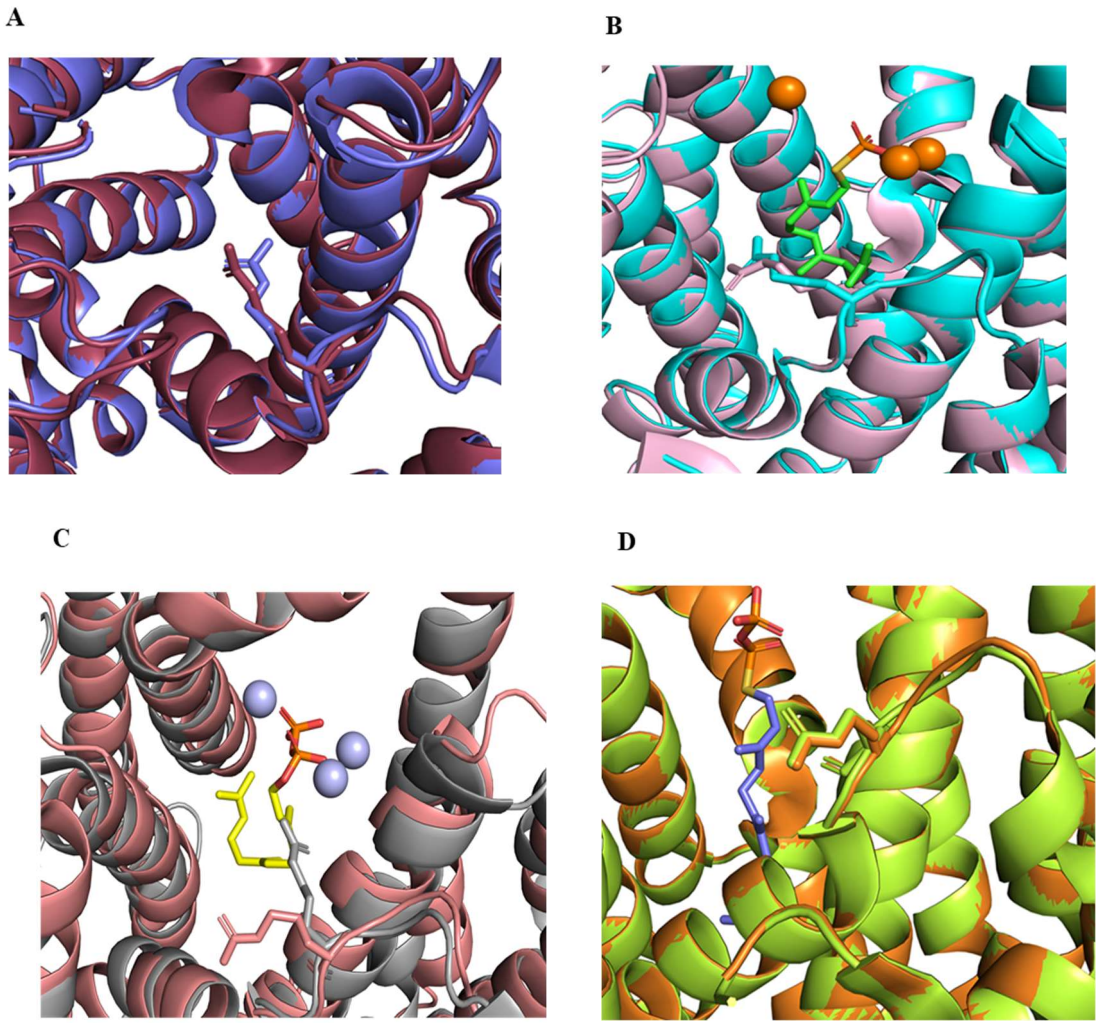


Fig. 3.8. Comparison of conformational changes in the loop harboring RXR motif of *(A) the wild type AaBOS (PDB ID: 4FJQ) and AaBOS^{M2} (PDB ID: 4GAX) shown in violet and red color, respectively. (B) apo (PDB ID: 3SDQ) and ligand bound (PDB ID: 3SAE) forms of α-bisabolene synthase represented with cyan and pink color, respectively. (C) apo Abietadiene synthase (PDB ID: 3S9V) and ligand bound 5-epi-aristlochene synthase (PDB ID: 5IK0) shown in raspberry and grey color, respectively (D) apo (PDB ID: 6A1I) and ligand bound (PDB ID: 6A1E) forms of SaSQS1 represented in green and orange color, respectively. Ligand and Mg^{2+} ions are represented as ball and stick.*

3.4 Conclusion:

SaSQS1 is a key member of sesquiterpene synthases family that catalyzes the cyclization of FPP to sesquisabinene, an important component of sandalwood oil. The structure of SaSQS1 was elucidated at 2.2Å resolution. It provides insights to the architecture of the class I terpene synthases. The conserved active site motifs, DDXXD and NSE/DTE, can be mapped on the structure of SaSQS1. Structural comparison of our structure with the other reported structures of SaSQS1 shows conformational deviations at the conserved regions, DDXXD and RXR(G). Furthermore, various trials of co-crystallization of SaSQS1 with the substrate analog were done to obtain the bound form structure of the enzyme. However, no density for the ligand was observed in electron density maps calculated from different data sets collected.

The structural comparison of *apo* and ligand bound forms of few sesquiterpene synthases have shown conformational deviations in these enzymes upon ligand binding. However, the definition and mode of occurrences of open and closed state of the enzymes are found to be inconsistent across different members of terpene synthases. In the context of SaSQS1, these structural deviations appear to have minimal influence on the mechanistic of the enzyme.

Chapter 4

Mutational studies of SaSQS1 to identify the product modulating residues

4.1 Background:

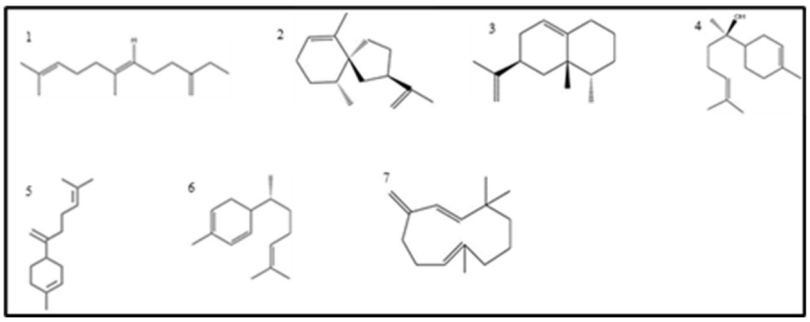
The most diverse group of terpenes is sesquiterpenes, which comprise of more than 300 stereochemical discrete hydrocarbons[18]. This diversity can be attributed to different types of sesquiterpene synthases which can stabilize and fold the acyclic FPP into various sesquiterpene hydrocarbon skeletons. The product formation catalyzed by sesquiterpene synthases significantly depends on the carbocation intermediates formed. There a number of carbocation intermediates formed during the cyclisation of FPP to sesquiterpenes. Generation of varied types of intermediates during the cyclization process leads to formation of different types of products and thus, product specificity. Structural and biochemical studies of a number of sesquiterpene synthases have provided insights to the residues that play role in stabilizing these intermediates and also in defining the product specificity of sesquiterpene synthases[49,52,59,135].

Despite exhibiting significant sequence and structural homology, especially in the catalytic pocket nestling α domain, sesquiterpene synthases form diverse products. However, the factors that determine the product specificity of this family of enzymes have still remained unclear. Previous efforts in determining product modulating residues (PMRs) of few sesquiterpene synthases were focused on identifying the divergent residues in the binding pocket, by performing structure based comparisons of the two close homologs. The residues in close vicinity of the active site, especially those which lie within a radius of 6Å from the ligand, were checked for their role in product specificity. The role of these residues was further validated by mutagenesis and biochemical studies (Table 4.1). For example, structural comparison of 5-epi-aristolochene synthase (TEAS) and premnaspirodiene synthase led to identification of nine residues in 5-epi-aristolochene synthase (A274, V291, V372, T402, Y406, S436, I438, I439, V516), which when mutated together, switch the major product formation from 5-epi-aristolochene to premnaspirodiene[135,136]. Conversely, mutation of corresponding residues in

vetispiradiene synthase led to the production of 5-epi-aristolochene, which is a cognate product of 5-epi-aristolochene synthase[137]. In a similar exercise, on comparison of structural models of β -farnesene synthase and amorpha-4,11-diene synthase, a combination of two mutations in β -farnesene synthase, Y402L and Y430A, was observed to switch the product from β-farnesene to cyclic products[59]. Similarly, domain swapping experiments between α-bisabolol synthase and amorpha-4,11-diene synthase led to the identification of different combinations of a set of residues in both the proteins, which when mutated generate an additional product, γ-humulene[133].

SSQ	Mutation	Product formed		Reference
		Major	Minor	
TEAS	W273E, W273F, W273C	β-farnesene[1]		138
	M9 (A274T, V291A, V372I, T402S, Y406L, S436N, I438T, I439L, V516I)	Premnaspirodiene[2]	5-epi-aristolochene, 4-epi-eremophilene, Germacrene A	135
Vetispiradiene Synthase	M9 (T281A, A298V, I379V, S409T, L413Y, N443S, T445I, L446I, I523V)	5-epi-aristolochene[3]	Premnaspirodiene, 4-epi-eremophilene	137
Amorpha-4,11-diene synthase	G439C	β-farnesene, α-bisabolol[4]	amorpha-4,11-diene	139
β-farnesene synthase	Y402L, Y430A	β-bisabolene[5], zingiberene[6], β-farnesene,	putative amorphadiene isomer, γ-curcumene, ar-curcumene,, cis-α-bergamotene, β-sesquiphelleandrene, α-bisabolene, α-bisabolol	59
AaBOS	CH9(I350F, K356R, N358D, I359L, D369E, V373, Q379K, L381A, L392H, S394P, I395V, A396V, V397I, N398I, L399T) CH12-M1 (I350F, K356R, N358D, I359L, D369E, V373N, Q379K, L381A, I395V, V3971, N398I, L399T) CH12-M2 (I350F, K356R, N358D, I359L, D369E, V373N, Q379K, L381A, I395V, L399T) CH12-M3 (I350F, K356R, N358D, I359L, D369E, V373N, Q379K, L381A, N398I, L399T) CH12-M6 (I350F, K356R, N358D, I359L, D369E, V373N, Q379K, L381A, I395V, V397I, L399T) CH12-M4 (I350F, K356R, N358D, I359L, D369E, V373N, Q379K, L381A, I395V, N398I, L399T) CH12-M7 (I350F, K356R, N358D, I359L, D369E, V373N, Q379K, L381A, V397I, N398I, L399T) CH13-M1 (V373N, L392H, S394P, I395V, A396V, V397I, N398I, L399T) CH13-M2 (V373N, L381A, L392H, S394P, I395V, A396V, V397I, N398I, L399T) AaBOS-M1 (V373N, I395V, N398I, L399T) AaBOS-M2 (V373N, L381A, I395V, N398I, L399T) AaBOS-M11 (V373N, I395V, N398V, L399T) AaBOS-M12 (V373G, I395N, N398I, L399T	α-bisabolol ,γ-humulene[7]		133

Table 4.1. *Sesquiterpene synthases with their known product modulating residues (PMRs) and the products formed due to mutation of these PMRs.*



Scheme 7. *Chemical structures of major products formed by mutating PMRs of the sesquiterpene synthases listed in Table 4.1.*

Apart from the above discussed plant sesquiterpene synthases, mutagenesis studies of synthases from other organisms have also been performed to highlight the role of the divergent residues in close vicinity of the binding pocket and few distant residues. For example, mutations of residues like D81, D84, W308 and H309 in pentalenene synthase, a sesquiterpene synthase from *Streptomyces*, form an additional product. These residues lining the active site of this enzyme when mutated, form germacrene A in addition to the cognate product of this enzyme, pentalenene[140]. Mutational and structural studies of epi-isozizaene synthase from *Streptomyces coelicolor*, have underlined a number of aromatic and aliphatic residues at the active site contour like Y69, L72, F95, F96, F198, W203 and V329 that play an important role in the product specificity[52,53]. Thus, to identify the divergent residues of SaSQS1 that may influence the product formation, this classical approach was adopted. In the present work, site-directed mutagenesis studies and biochemical validation of these mutants were done in order to identify the product specificity determinants of SaSQS1. Structures of two key mutants, T313S and G418A, were elucidated by employing X-ray crystallography to study the structural basis of the product formation catalyzed by SaSQS1. Structural comparisons of the mutants and the *wild* type were also carried out to understand the conformational changes that affect the product yield.

4.2 Methodology:

4.2.1 Site-directed Mutagenesis:

Specific mutants were amplified by PCR using the QuikChange Lightning site-directed mutagenesis kit (Agilent) with the suitable primer pairs (Appendix B). All primers were synthesized by Eurofins, Bangalore. The *wild* type gene in pOPINss vector was used as a template with the N-terminal 6xHIS-SUMO tag. PCR was performed in thermal cycler with the following program: Denaturation at 95 °C for 5 minutes followed by 18 cycles of

denaturation for 30 sec at 95 °C; annealing for 30 sec at 54 °C; and extension of 7 min 30 sec at 68 °C followed by single cycle of final extension at 68 °C for 10 min. Amplified products were subjected to Dpn1 digestion for 2 hours at 37 °C followed by the transformation in DH5α competent cells. Colonies positive for mutation were confirmed by sequencing and the chromatograms were analyzed with the help of chromas 2.1 software.

4.2.2 Expression and Purification:

The optimal expression of all the mutants was observed in C41 strain, transformed with the plasmid. The cells were grown at 37 °C till the OD_{600} value of 0.5 and gene expression was induced for 16 hours at 20 °C using 0.5 mM Isopropyl β-D-thiogalactopyranoside (IPTG). The bacterial cell pellet was lysed by sonication in buffer containing 50 mM Tris-HCl (pH 8.0), 500 mM NaCl, 10 mM imidazole, 5% glycerol, 2 mM β-Merkaptoethanol, 2 mM phenylmethylsulfonyl fluoride (PMSF), 2 mM benzamidine, 20 mM sucrose and lysozyme. The lysate was centrifuged at 20,000 rpm for 30 minutes and loaded on the Ni-NTA resin (GE Healthcare Biosciences, USA) for binding. The bound protein was eluted with 100 mM imidazole. The N-terminal 6xHis-SUMO tag was removed by incubating the protein, overnight, with TEV (Tobacco Etch Virus) protease under dialyzing conditions at 10 °C. To remove the TEV protease and cleaved 6xHis-SUMO tag, the protein was further subjected to inverse Ni-NTA affinity purification. The untagged protein was further purified by gel filtration chromatography using Superdex 200 increase 10/300 GL with buffer containing 20 mM Tris-HCl (pH 8), 150 mM NaCl and 2 mM DTT. The purity of the protein was assessed with SDS-PAGE.

4.2.3 Gas Chromatography-Mass Spectrometry (GC-MS) assay:

Enzyme activity was confirmed by product ratio studies. The assay mixture comprising of 600 µg of enzyme and 200 µM Farnesyl Pyrophosphate in 500 µL of HEPES buffer (25 mM, pH 7.5) containing 10%glycerol, 5 mM DTT and 10 mM Magnesium Chloride was incubated at 25-30 °C for 2 hrs. After this incubation period, the assay was quenched by keeping in ice, and 0.2 g NaCl was added and extracted using n-Hexane (500 µL X 3). The pooled extract was dried over anhydrous sodium sulphate, concentrated using nitrogen jet and analyzed by GC-MS fitted with 30 m x 0.32 mm x 0.30 µm capillary column (HP-5, J&W Scientific) by using the following program: 70 °C to 170 °C at 5 °C/min followed by 170 °C to 210 °C at 15 °C/min with a hold time of 5 minutes. Helium gas was used as a carrier gas at a flow rate of 1 ml/min.

4.2.4 Crystallization and Data collection:

The two mutants, T313S and G418A, were concentrated to 15 mg/ml after size exclusion for crystallization trials using commercial crystallization suites. Good crystals for both the mutants were observed within two weeks with the following condition: 0.1M MES (pH 6.5), 0.1 M magnesium acetate, 10% PEG 10,000. Crystals were obtained with sitting drop method at 20 ^{0}C using 100 nL of both protein and precipitating solution equilibrated against 50 µL of reservoir solution. For diffraction, the crystals were soaked in cryoprotectant containing reservoir buffer with 28% glycerol and were flash cooled with nitrogen stream at 195 K. The data set was collected at I03 beamline facility of Diamond synchrotron and BL-21 beamline of Indus-2, RRCAT for G418A and T313S, respectively.

4.2.5 Data analysis and Structure determination:

Both the data sets were integrated with XDS[81] and scaled with Aimless[82] of CCP4 suite. The structure was solved using molecular replacement technique with PHASER[125] program by employing SaSQS1 (PDB ID: 6K16) as a search model. The structure was refined with PHENIX [126] suite of programs and COOT [91] was used for iterative rounds of refinement.

4.3 Results and Discussion:

4.3.1 Identification of divergent residues at the binding site of SaSQS1:

To identify the product modulating residues of SaSQS1, we followed the classical approach of mutating divergent residues, which lie within a radius of 6 Å from the ligand. Based on sequence and structure comparative studies of SaSQS1 with its closest homolog, 5-epi-aristolochene synthase (Fig. 4.1A), eleven mutants of SaSQS1 (T313S, T313V, Q387S, Q387L, Y391F, Y391V, S416A, I417F, G418A, F419A and R457K) were tested for product specificity (Fig.4.1B). Here, the rationale for the mutation was based on the earlier observation that both nature of the residue (hydrophobic or hydrophilic) and length affect the product formed.

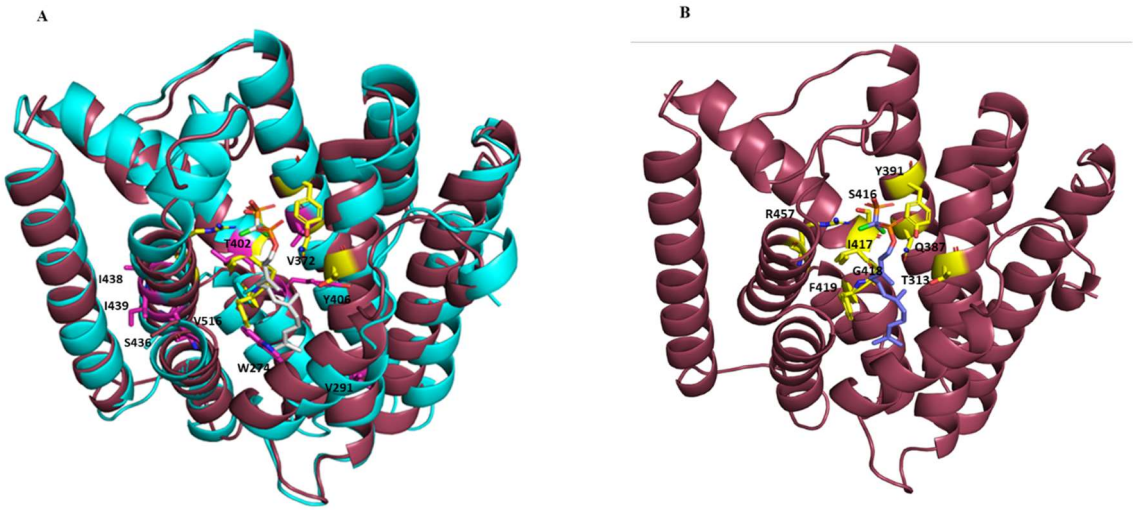


Fig. 4.1. Structural comparison of SaSQS1 and 5-epi-aristolochene synthase colored in deepsteal and cyan, respectively. *(A)The α domain comparison of both the structures (RMSD- 2.1 Å), with the product specificity determinants of 5-epi-aristolochene synthase highlighted with pink sticks and the residues of SaSQS1 selected for mutagenesis studies highlighted with yellow color. The labelled residues correspond to 5-epi-aristolochene synthase. (B) The residues of SaSQS1 in the active site vicinity and tested for product specificity. The ligand is represented with sticks.*

4.3.2 Expression and Purification of the mutants:

The mutants in pOPINss expression vector were confirmed by sequencing. It contains N-terminal 6x His and SUMO tag followed by TEV protease cleavage site. All the mutants were expressed in C41 strain of *Escherichia coli* and were found to be in the soluble fraction. The TEV protease treatment was followed by Ni-NTA affinity chromatography and gel filtration (Fig. 4.2A) for all the mutants. The mutants eluted in the expected fractions. However, in the case of the mutants I417F and R457K, degradation of the protein was observed after purification. Chromatogram after size exclusion of the mutant S416A, representative for all the mutants, is shown in Fig. 4.2B. Purified protein was used for activity assays and crystallization trials.

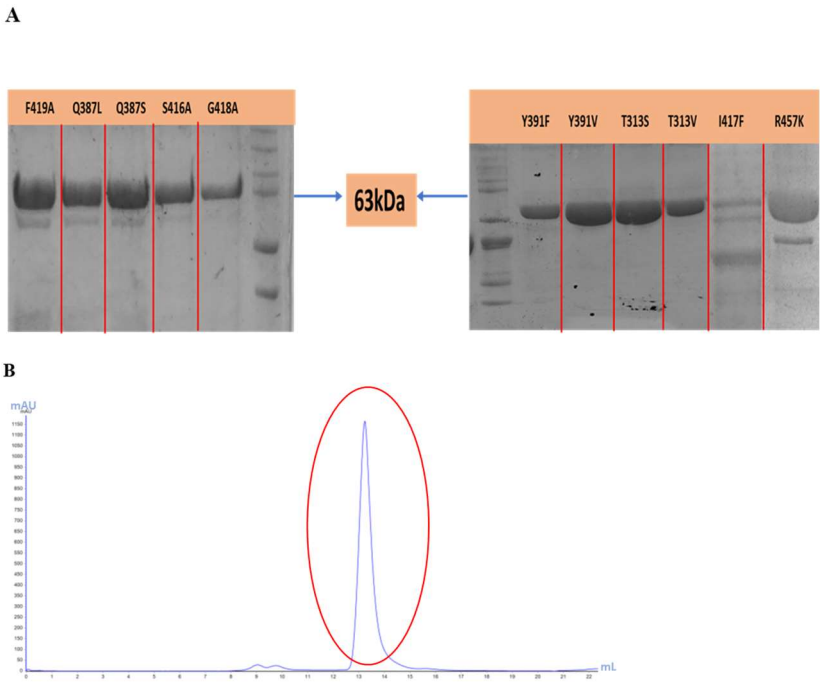


Fig. 4.2 Purification of mutants of SaSQS1. *(A)Purity of the mutants after size exclusion assessed by SDS PAGE. (B) Size exclusion chromatogram of S405A mutant.*

4.3.3 Individual divergent residues at the binding pocket of SaSQS1 do not modulate the product:

The purified mutants were tested for their activity using product ratio studies. The assay was carried out using the enzyme and FPP in HEPES buffer and the extracted product was concentrated to approximately 100 µL and analyzed by GC-MS. The product was confirmed by comparing the fragmentation pattern of *wild* type SaSQS1[141] and was characterized based on the NIST search. The mutants Q387S, Q387L, Y391F, Y391V, S416A and F419A, when incubated with FPP show the formation of sesquisabinene in GC-MS (Fig. 4.3), however, no change in the product type was observed for these mutants. Particularly, F419A and Y391V show increase in the product yield, whereas, all others, showed the reduction in product level when equal quantity assay mixtures were compared. Replacing T313 with serine and valine, I417 with phenylalanine, G418 with alanine and R457 with lysine knocked out the activity as no sesquisabinene formation was observed. Table 4.2 summarizes the product ratios of all the mutants.

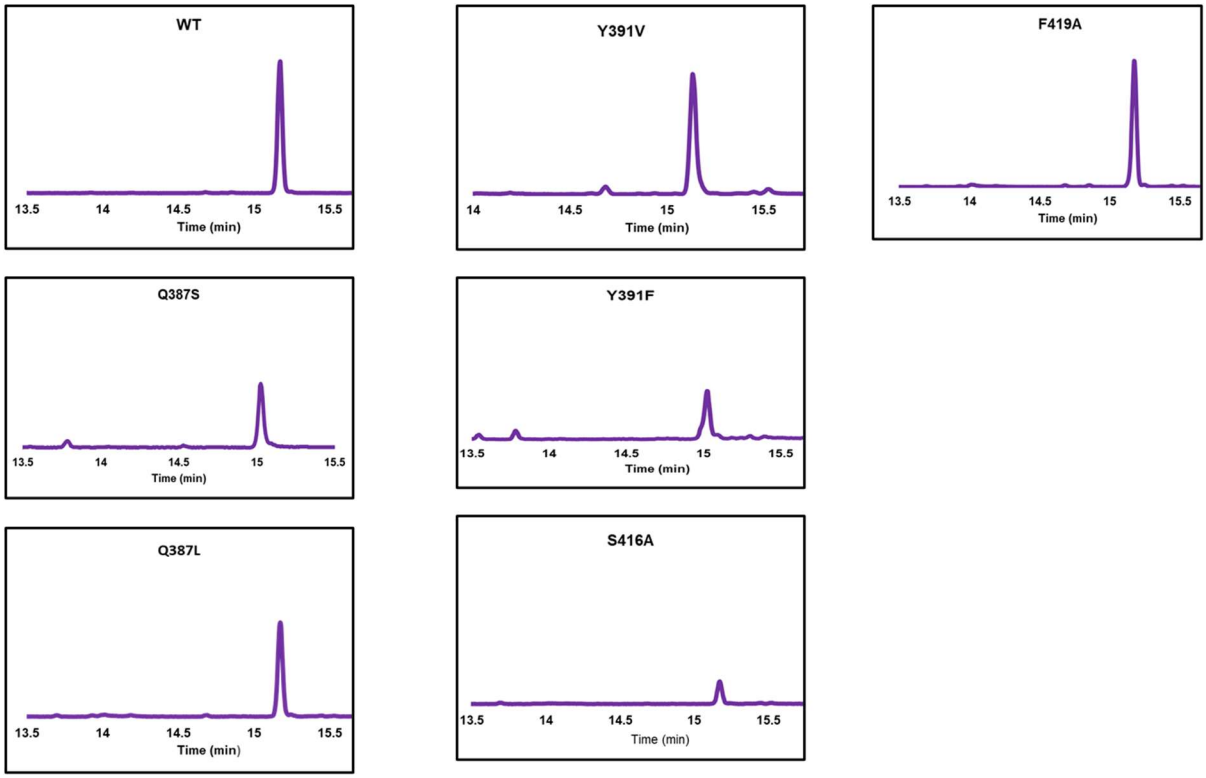


Fig. 4.3. *The GC-MS profiles of products from wild type and mutant forms of SaSQS1. Only those mutants, which exhibit the activity are represented here. In all of the chromatograms, the peak corresponds to sesquisabinene.*

Mutant	Expression	Product
F419A	✓	Sesquisabinene
Q387L	✓	Sesquisabinene
Q387S	✓	Sesquisabinene
S416A	✓	Sesquisabinene
Y391F	✓	Sesquisabinene
Y391V	✓	Sesquisabinene
G418A	✓	Not detected
I417F	Degradation and less expression	Not detected
R457K	Degradation and less expression	Not detected
T313S	✓	Not detected
T313V	✓	Not detected

Table 4.2. *All the active and inactive mutants with their expression details.*

4.3.4 Crystallization and Data collection of the mutants:

Structural studies of various mutants have provided information regarding the product specificity determinants of sesquiterpene synthases[52,53,142]. These studies also help in delineating the plausible mechanism of action of these enzymes[143,144]. Henceforth, in the present work, the key mutants were subjected to structural studies. Thus, the mutants which did not form sesquisabinene were subjected to crystallization in order to observe the structural basis of the conformational changes in the enzyme that renders it inactive. All the inactive mutants with a concentration of 15 mg/mL were screened for crystallization using commercial suits. Crystals for two mutants, T313S and G418A, were obtained with sitting drop in a condition comprising of 0.1 M MES (pH 6.5), 0.1 M magnesium acetate, 10 % PEG 10,000 (Fig. 4.4). The diffraction data for both the mutants was collected and the data sets were processed. Diffraction data collection and refinement statistics are summarized in Table 4.3.

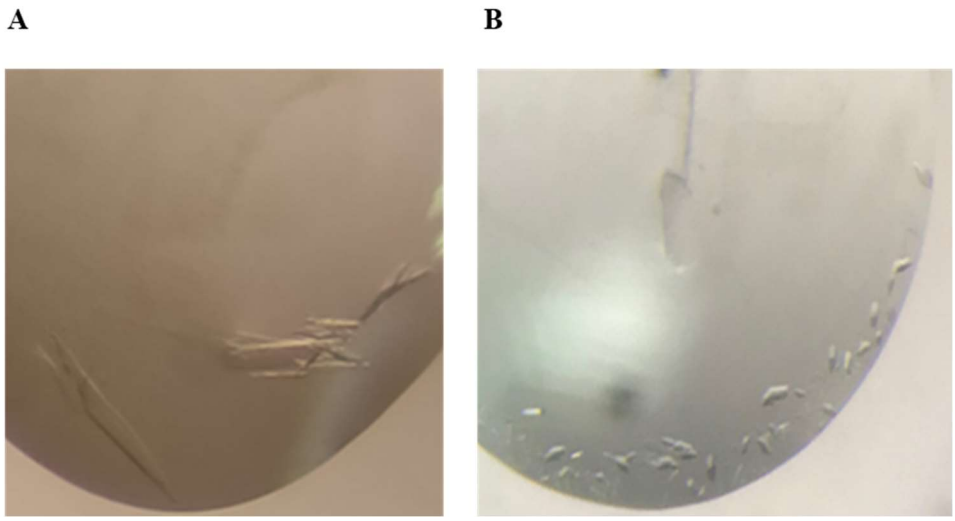


Fig. 4.4. Crystals of the mutants. *(A) T313S (B)G418A observed in similar condition.*

Data Collection	*G418A*	*T313S*
Space group	P1211	P1
Resolution (Å)	47.58-3.10	45.62-3.4
Cell parameters		
a, b, c (Å)	54.07, 61.03, 182.71	81.88, 82.06, 136.83
α, β, γ (°)	90, 93, 90	99.77, 91.55, 119.78
Observed reflections	135060 (24935)	88830 (9651)
Unique reflections	21943 (3923)	40630 (4530)
I/σI	11.5 (1.7)	3.4 (1.2)
R_{merge}	0.08 (0.86)	0.24 (0.75)
Completeness (%)	99.8 (99.8)	98.4 (96.6)
Multiplicity	6.2 (6.4)	2.2 (2.1)
$CC_{1/2}$	0.99 (0.89)	0.93 (0.52)
Refinement		
Resolution (Å)	45.61-3.10	45.61-3.4
Number of reflections		
Working set	21671	40621
Test set	1074	2163
R_{work}/R_{free}	0.237/0.274	0.219/0.282
Number of atoms		
Protein	8165	16424
Other	-	8
Mean B-factors (Å^2)		
Protein atoms	108.0	39.0
RMSD from Ideal values		
Bond length (Å)	0.011	0.011
Bond angles (°)	1.553	1.334
Ramachandran plot Statistics		
Preferred (%)	98.79	96.99
Allowed (%)	1.21	3.01

Table 4.3. *Data collection and refinement statistics of mutants, G418A and T313S. Values in parenthesis are for the highest resolution shell.*

4.3.5 Conformational studies of mutant T313S:

Structure of the mutant, T313S, was elucidated at a resolution 3.4 Å (PDB ID: 7E9R) to study the structural basis of the inactivity of this particular mutant. The residue T313 is in close vicinity of the binding pocket and on comparison with the close homolog, 5-epi-aristolochene synthase, was found to be divergent. Thus, the residue was mutated to study its effect on the activity of the enzyme and product specificity. The mutants, T313S and T313V, did not play role in the product specificity, however, were found to be inactive. Thus, the residue T313, has an important role in the catalysis leading to product formation.

Crystals of the mutant T313S were obtained and the data was solved up to the R and R_{free} values of 0.219 and 0.282, respectively. The mutant structure shows varied crystal packing when compared to the *wild* type as it possesses four molecules in the asymmetric unit (Fig. 4.5A), unlike the *wild* type which has one molecule in the asymmetric unit. However, the oligomeric state of the enzyme is a monomer validated by size exclusion chromatography (Fig. 4.5B). Furthermore, the mutant crystallized in a different space group, P1, as opposed to the *wild* type which has a C121 symmetry. Similar to the *wild* type, density for one of the Mg^{2+} ions close to the DDXXD motif was observed, despite the absence of bound ligand in the mutant structure.

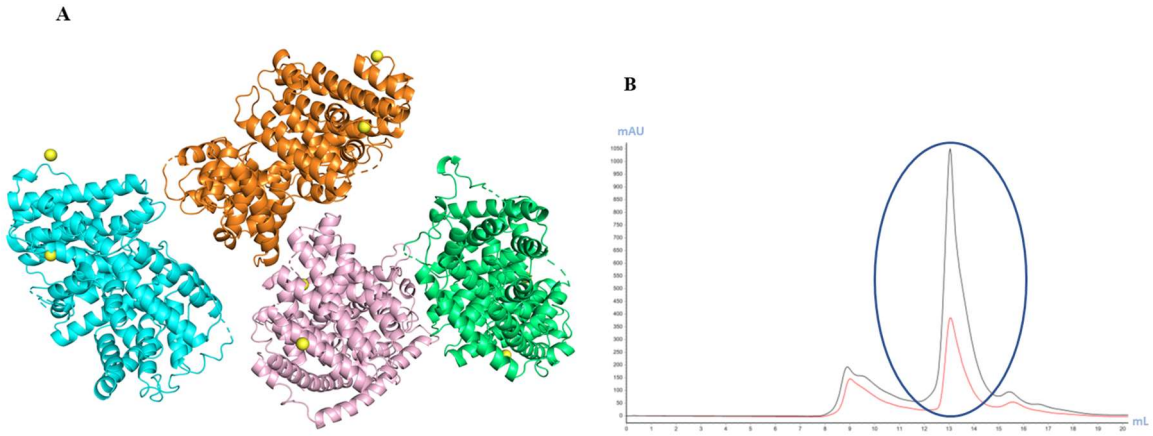


***Fig. 4.5. Crystal structure of T313S (PDB ID: 7E9R)**. (A) Structure of T313S with four molecules in the asymmetric unit. The Mg^{2+} ions are represented as yellow spheres. (B) Overlay of size exclusion chromatograms of SaSQS1 wild type, colored in red, and T313S mutant, colored in black, showing same elution volume for both.*

4.3.6 Structural studies of G418A mutant:

Another key residue G418, located in the binding pocket was mutated to another small residue, alanine. This mutation, however did not play role in product specificity, rendered the enzyme inactive reflecting the importance of the residue in the catalysis of the enzyme. This glycine

introduces a kink in the contiguous helix. The corresponding glycine in another sesquiterpene synthase, selinadiene synthase, is a part of an 'effector triad' which rearranges on ligand binding and triggers the cleavage and release of pyrophosphate group of the substrate, thus, initiating the catalysis[130]. Hence, to study the structural basis of the effect of this residue on the conformation and mechanism of SaSQS1, structure of G418A mutant was elucidated at a resolution of 3.1 Å (PDB ID: 7E6W).

The final structure model was refined till R and R_{free} values of 0.237 and 0.274, respectively. Similar to the mutant T313S, it exhibits distinct crystal packing as compared to *wild* type SaSQS1. It has two molecules in the asymmetric unit (Fig. 4.6A) with P1211 space group. However, in this case also the mutant was observed to be a monomer in the solution state (Fig. 4.6B). Furthermore, no density for any of the Mg^{2+} ions was observed in this case. This complies with structures of other *apo* sesquiterpene synthases where no bound Mg^{2+} ions were observed in the absence of ligand[133,145,146]. Interestingly, there was introduction of another mutation, F419N, observed in the structure next to the engineered mutation (Fig. 4.6C). Mutagenesis studies of this particular residue, F419, are also carried out in the present work. The mutant thus generated, F419A, was found to yield sesquisabinene as the product. However, it cannot be stated that the inactivity of the mutant G418A is due to substitution of the glycine alone or it is a compound effect of the double substitution.

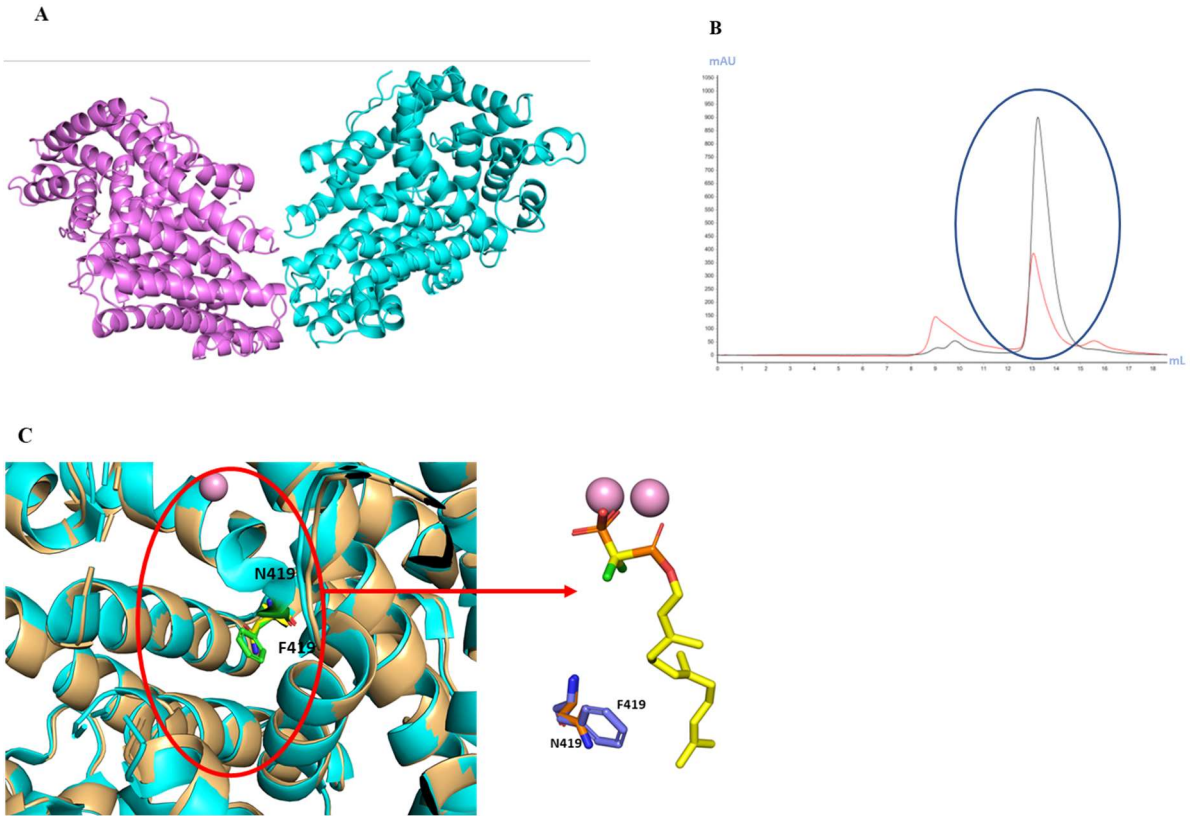


Fig. 4.6. Crystal structure of G418A (PDB ID: 7E6W). *(A) Structure of G418A mutant with two molecules in the asymmetric unit. (B) Overlay of size exclusion chromatograms of SaSQS1 wild type, colored in red, and G418A mutant, colored in black, showing same elution volume for both. (C) Superimposition of wild type SaSQS1, colored cyan, and mutants G418A, colored light orange, showing mutation F419N. The residues are represented as sticks and* Mg^{2+} *ion as sphere.*

4.3.7 Differential dynamics of the mutants:

Comparative analysis of the conformational deviation between the *wild* type and the mutant proteins was done to understand the role of the mutated residues in the enzyme activity. The structural superimposition of the wild type SaSQS1 with the mutants, G418A (RMSD- 0.96 Å) and T313S (RMSD- 0.53 Å), show conformational deviations in the binding pocket vicinity. (Fig. 4.7A, B).

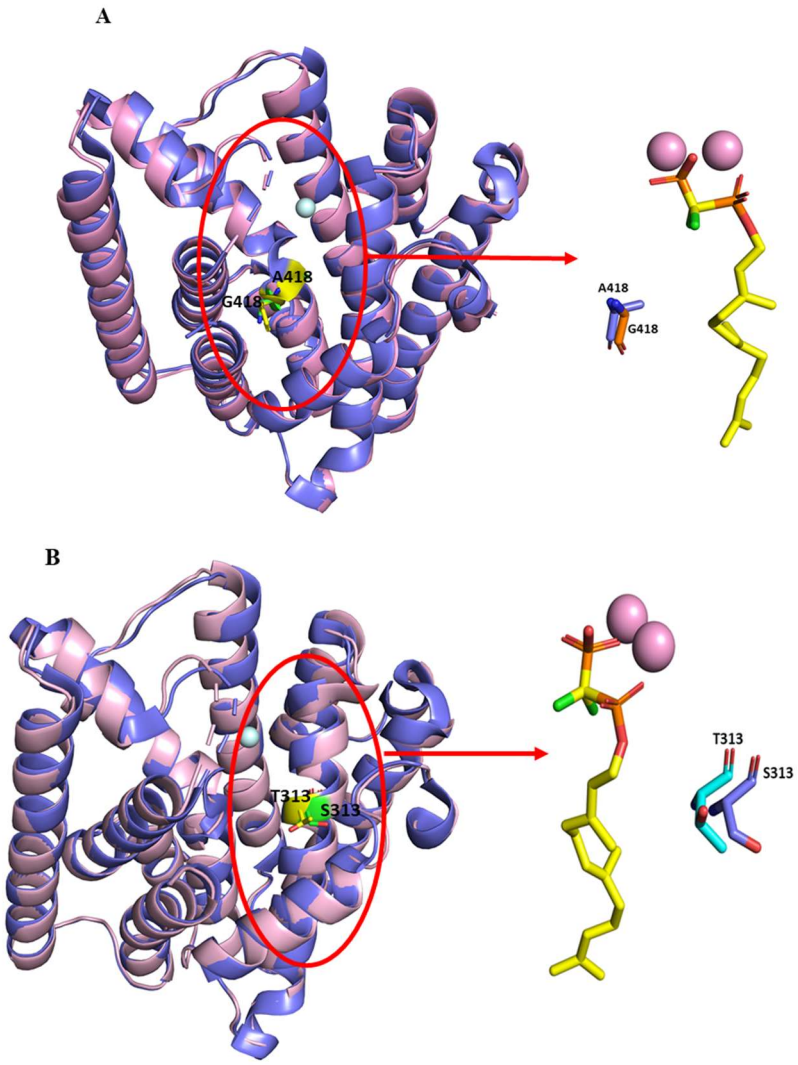


Fig. 4.7 Structural superimposition of core helices of wild type SaSQS1 and mutants *(A) Superimposition of the wild type SaSQS1 (PDB ID: 6K16) and mutant G418A (PDB ID: 7E6W) represented with violet and pink color, respectively. (B) Structural overlay of wild type SaSQS1 (PDB ID: 6K16) and T313S (PDB ID: 7E9R) colored violet and pink, respectively. The mutated residues are represented with yellow sticks.*

An overlay plot of the Cα RMSD of the *wild* type, both *apo* (PDB ID: 6K16) and ligand bound (PDB ID: 6A1E) and the mutants show deviation of these mutants at the specific site of mutation (Fig. 4.8). Also, conformational variations are observed in the regions 316-331, that harbor the DDXXD motif, and 391-406, region close to the residue G418. These conformational deviations may play role in differential dynamics of the mutants, rendering the mutants inactive.

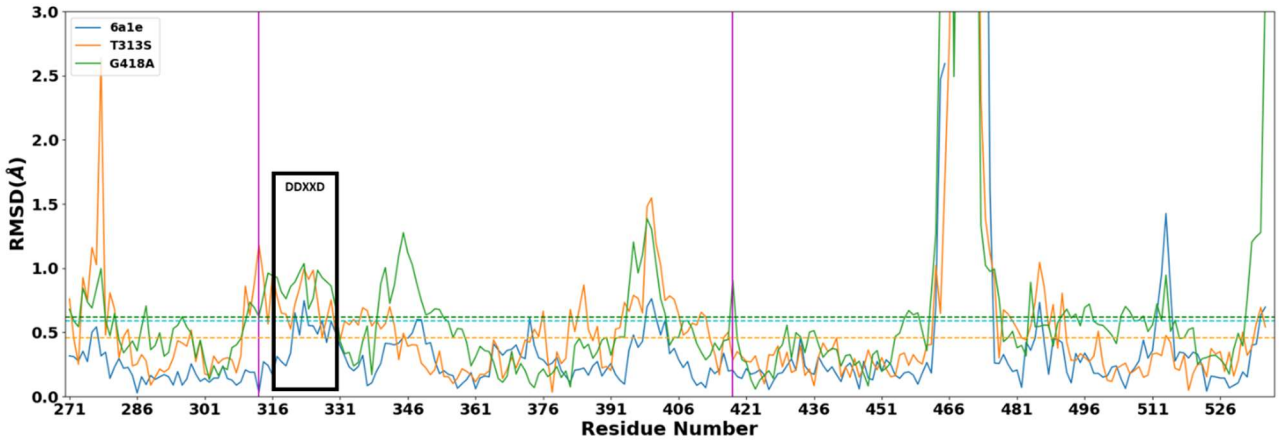


Fig. 4.8. *Overlay of plots of Cα RMSDs between apo SaSQS1 (PDB ID:6K16) & ligand bound SaSQS1 (PDB ID: 6A1E) and mutants T313S & G418A. The site of mutation is shown with a pink line. The conserved DDXXD motif is highlighted with box.*

4.4 Conclusion:

The sesquiterpene diversity can be attributed to the active site contour of the sesquiterpene synthases. A number of mutagenesis and structural studies of sesquiterpene synthases have provided an array of residues that play role in product specificity of these enzymes. For SaSQS1 the conventional method of comparative analysis with the homolog was done to identify the potential product specificity determinants. Based on this approach, eight residues were identified from which eleven mutants were designed. Product ratio studies of these mutants help in identification of residues in the binding site that did not yield sesquisabinene indicating their role in product formation. These residues include T313, I417, G418 and R457. Few mutants including F419A, Q387L, Q387S, S416A, Y391F and Y391V show change in the product yield but not in the product specificity. Amongst these, F419A and Y391V show increase in the level of product formation as compared to the *wild* type, indicating proper stabilization of the carbocation intermediates. Structural studies of two key inactive mutants, T313S and G418A, show conformational deviations that may play role in differential dynamics of the mutants. These deviations might play role in the enzyme catalysis, altering the product yield.

Chapter 5

Identification of a novel approach to determine product specificity of plant sesquiterpene synthases and SaSQS1

5.1 Background:

The conventional sequence and structure based approaches to identify product defining residues of sesquiterpene synthases (SSQs) are marred by a number of limitations. Earlier efforts were focused on identifying the divergent residues in the binding pocket, especially those which lie within 6Å radius with the ligand [59,136]. However, given the size of the binding pocket, the possible residue combinations that could affect the product are astronomical. For example, to identify product modulating residues (PMRs) of β-farnesene synthase, a library of 2^{24} mutants were identified. Using structure based combinatorial protein engineering, 282 mutants were selected and tested biochemically for their role in product specificity[59]. Although, this approach seems to have worked for few sesquiterpene synthases like 5-epi-aristolochene synthase (TEAS)[135], Vetispiradiene synthase[137], β-farnesene synthase[59] and α-bisabolol synthase (AaBOS) [133], the rationale underlying the identification of PMRs appear to be failing when applied across the other divergent members of the family (Fig. 5.1).

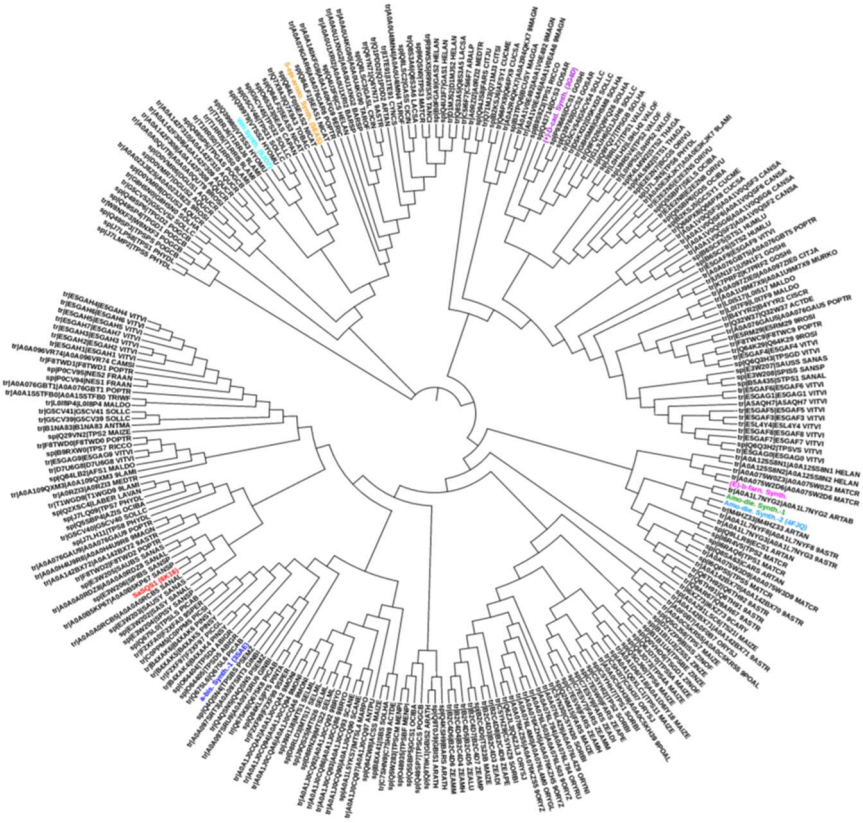


Fig. 5.1. Phylogenetic tree of plant sesquiterpene synthases from the curated database*[147] *designed using Interactive tree of life (iTOL) tool [148]***.*** *The synthases used for obtaining sequence based sectors are color coded with the scheme: SaSQS1- red; α-bisabolene synthase- blue; amorpha-4,11-diene synthase 1- green; α-bisabolol synthase or amorpha-4,11-diene synthase 2- light blue; β-farnesene synthase; Pink; Delta cadinene synthase- Purple; 5-epi-aristolochene synthase- orange; Vetispiradiene synthase- cyan.*

Furthermore, in all of the above cases, the methodology employed in devising the subset of the divergent residues markedly varies. Also, this conventional approach was tried for SaSQS1 to identify the product specificity of the enzyme, as explained in chapter 4 of this thesis. However, the residues identified by this approach did appear to influence the type of the product formed by the enzyme. Thus, a generalized approach that could provide putative set of PMRs was the need of the hour.

Therefore, in order to fill this lacuna, a novel approach employing statistical coupling analysis (SCA) was employed. Previously SCA has been applied to identify group of coevolving residues by analyzing multiple sequence alignment of proteins of a family[115,149]. These groups of coevolving residues, known as "sectors", when mapped on the structure of the protein being investigated showed to cluster the functionally important residues of the protein. Sectors could unravel the structural basis of protein functions like allostery[117,118], signal

transmission[113,115,119,120], cumulative dynamics of catalytic reactions[118] and protein adaptability[121] (Fig. 5.2).

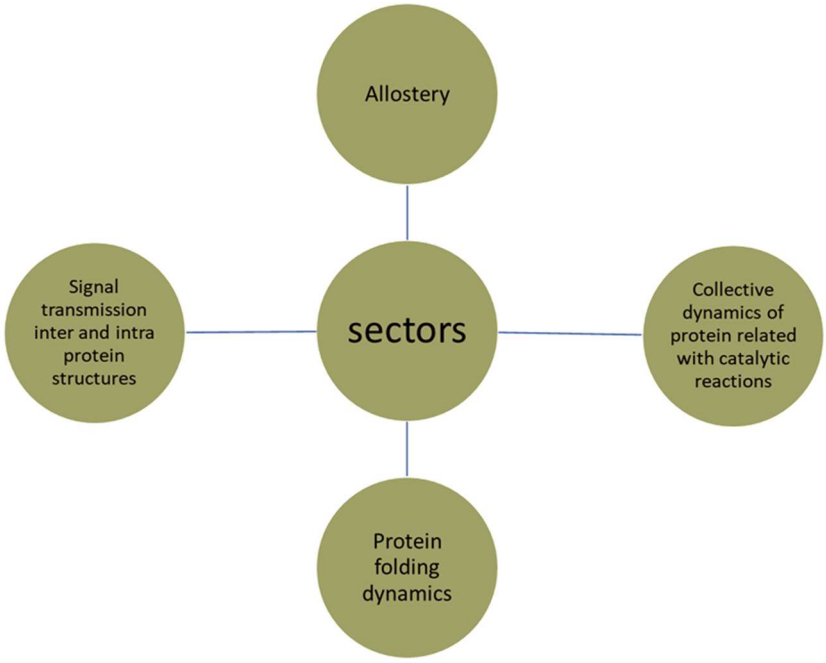


Fig. 5.2. *Various aspects of functions and structure of a protein represented by different sectors.*

Although, SCA could pin down the functionally important residues of a particular protein, the factors that cause subtle deviations in the functions arising from the sequence divergence within a sub-family or a sub-set of proteins cannot be easily deduced from this technique. For example, the system of our interest, sesquiterpene synthases. Sequence based SCA of sesquiterpene synthases clearly map catalytically important residues. Indeed, the analysis also provided residues that could be essential for evolving the structure of catalytic pocket. However, identifying the residues between two homologous members of the family, that are responsible for the product divergence, from this technique is not straightforward. This is primarily due to the statistical nature of the technique. In principle, SCA could have delineated the product modulating residues of sesquiterpene synthases, if the data set had large number of sequences of very close homologous proteins producing similar products. Furthermore, divergent residues at the catalytic pocket, identified from the simple structure comparison did not help in pinning down the product specificity factors (explained in chapter 4). However, structural comparison of various sesquiterpene synthases within the family and also in their ligand free and bound forms showed that the catalytic pocket is highly dynamic (Fig. 5.3A, B). From first principles of protein structure one can ascertain that the primary structure (amino acid sequence) of the protein manifests in the three-dimensional structure of the protein and also the dynamical

content of their structure. In other words, sequence, and hence the structural divergence, reflects in the dynamics of the protein. Since structural dynamics is a key functional element for all enzymes, we asked whether product modulating residues of the sesquiterpene synthases could be characterized by looking at the divergence of the dynamical content of these enzymes.

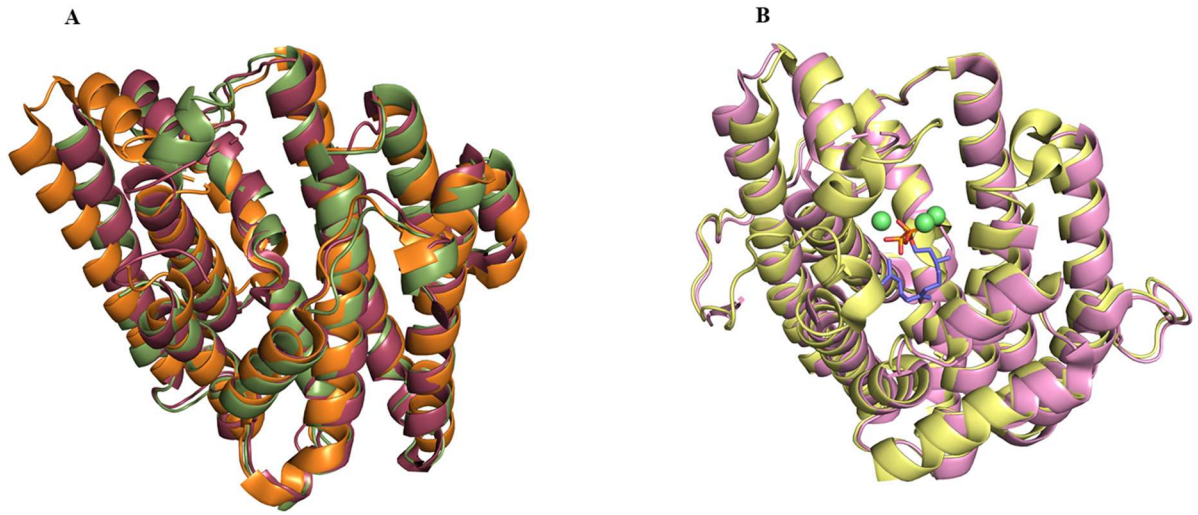


Fig. 5.3. *Structural comparison of α domain of: (A) sesquiterpene synthases including δ-cadinene synthase (PDB ID: 3G4D) represented with raspberry color, 5-epi-aristolochene synthase (PDB ID: 5EAS) represented with green color and α-bisabolol synthase (PDB ID: 4FJQ) represented with orange color. The superposition shows changes in the dynamics of the active site of different synthases. (B) ligand bound (PDB ID: 4OKZ) and apo form (PDB ID: 4OKM) of Selinadiene synthase represented with yellow and pink color, respectively showing changes in the active site upon ligand binding.*

In the present work, a generalized approach to identify putative PMRs of sesquiterpene synthases and SaSQS1 was developed employing SCA of their structural dynamics. This novel technique employs a combination of sequence, structure and dynamical features of sesquiterpene synthases. Using spectral decomposition technique, sequence and dynamics based sectors of few sesquiterpene synthases were identified. The dynamics based sectors were further refined by combining sequence based hydrophobicity index and structure based ligand vicinity index to identify most plausible product specificity determinants. It has been shown that the methodology developed here works well for all of the sesquiterpene synthases with known PMRs, despite their significant phylogenetic divergence. Furthermore, we applied this methodology to identify putative product defining residues of SaSQS1.

5.2 Methodology:

5.2.1 Docking and Molecular Dynamics (MD) Simulation:

Coordinates for the MD simulation were prepared using Maestro module of Schrödinger suite 2020-2. The initial coordinates of SaSQS1 (PDB ID: 6K16), TEAS (PDB ID: 5IK0), AaBOS (PDB ID: 4FJQ) and mutant *M2* of AaBOS (PDB ID: 4GAX), mutants G418A (PDB ID: 7E6W) and T313S (PDB ID: 7E9R) were obtained from Protein Data Bank and missing loops were built with Modeller[150]. All these structures, except TEAS, were in *apo* form. In the event of non-availability of ligand bound structures, FPP was docked on these structures using Glide (v8.7) module of Schrödinger, employing extra precision (XP)[107] docking mode. The conformers with highest glide score were used for MD simulation. Simulation system was prepared by placing coordinates in an orthogonal box with 10 Å buffer distance in all the directions after assigning bonds. The system was solvated with TIP3P (transferable intermolecular potential with 3 points)[151] solvent atoms and neutralized by addition of Na^+/Cl^- ions. The system was further subjected to 2000 steps of conjugate grade with 120 kcal/mole/Å under NPT. Finally, the system was subjected to MD simulation for 300 ns and the trajectories were sampled at every 50 ps.

5.2.2 Sequence based Statistical Coupling Analysis (sSCA):

The python version of SCA module, PySCA was used to generate sequence based sectors (sSCA)[149]. For sSCA calculation, sequences from curated plant SSQs database [147] were used. The database contains 248 sequences, out of which, 139 were identified as unique sequences by the program. For the current analysis, the multiple sequence alignment (MSA) was trimmed at both N and C terminus such that it contains the α domain starting from residue G271 to Q537 (residue numbers as per SaSQS1 sequence). The MSA was further processed as described by Rivoire *et.al* [149]. Briefly, this method involves generation of a positional weighted coevolution matrix (sC_{ij}) from the MSA. The group of coevolving residues, known as sectors (sICs) will be identified from the spectral decomposition (diagonalization of C_{ij}) of the positional coevolution matrix[149]. Each of the eigenmodes (K^*) obtained from decomposition represent a set of coevolving amino acids, which were used for calculating the sectors. The significant K^* would be identified from the histogram of eigenvalues of the C_{ij} with the average spectrum obtained from many trials of the randomized alignments.

5.2.3 Molecular dynamics based sector identification:

The dynamical cross-correlation matrix (DCCM) dC_{ij} , [152] was calculated as:

$$dC_{ij} = \langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle$$

Where, r_i and r_j are the position vectors of the i^{th} and j^{th} Cα atoms, respectively, at time *t*. To calculate this, R implementation of Bio3D program (version 2.4-1) was used [153]. The dynamical sectors were obtained from the spectral decomposition algorithm developed by Rivoire *et.al.* [149]. The MATLAB implementation of SCA was employed for this purpose. The dimensions of dC_{ij} were kept identical to that of sC_{ij}.

5.2.4 Calculation of Hydrophobicity and Vicinity indices:

The hydrophobicity and vicinity indices were calculated for these dynamical sectors. The hydrophobicity index (HI) is the average kd hydrophobicity scores of the sectors[154]. To calculate the vicinity index (VI), first a pseudo reference atom was fixed at the center of mass of the Cα atoms of residues A289, I312, G418 and C456 (the residue type and numbers are according to the SaSQS1 structure and for other structures, equivalent residues were taken) (Fig. 5.4). The vicinity index for a dynamical sector, dIC, is calculated as:

$$V_{dIC} = \frac{1}{N} \sum_{i \forall dIC} \delta_i$$

Where, N is total number of residues in sector dIC and δ_i= 1 if the Cα atom of the i^{th} residue lies within 12Å from the reference atom, else δ_i = 0.

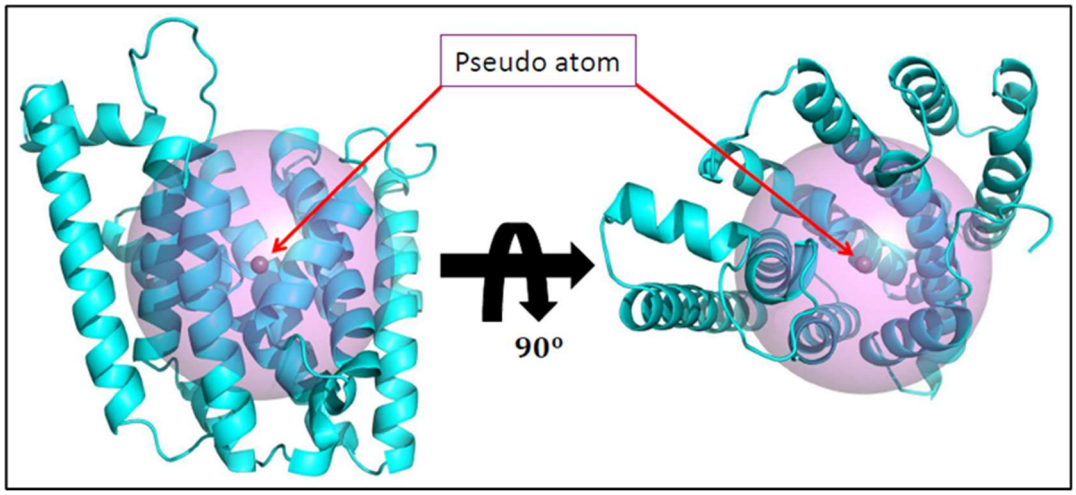


Fig. 5.4. *A pseudo reference atom, colored purple, shown at the center of the binding pocket of AaBOS (PDB ID: 4FJQ). The pink sphere shows the radius considered for calculating the vicinity index.*

5.3 Results and Discussion:

5.3.1 Role of evolutionary coupling of sequences in defining the PMRs of Plant sesquiterpene synthases:

Previous studies to define the product specificity of sesquiterpene synthases like TEAS, AaBOS, β-farnesene synthase, etc. showed that product modulation requires alterations in a set of residues rather than residues at one or two specific positions, sometimes including residues distant from the active site (Fig. 5.5). These observations suggest that an intrinsic coupling amongst the catalytic site residues alter the product formed. We probed the positional coupling in SSQs using sequence based statistical coupling analysis (sSCA). As explained in the methods section, sSCA computes covariance matrix (sCij) based on the sequence conservation and decomposes to identify sectors in the protein, which are shown to have functional correlation[149]. Briefly, this method involves generation of a positional weighted coevolution matrix (C_{ij}) from the multiple sequence alignment. The group of coevolving residues, known as "sectors" will be identified from the spectral decomposition (diagonalization of C_{ij}) of the positional coevolution matrix[149]. To begin with, approach was explored to identify PMRs in sesquiterpene synthases.

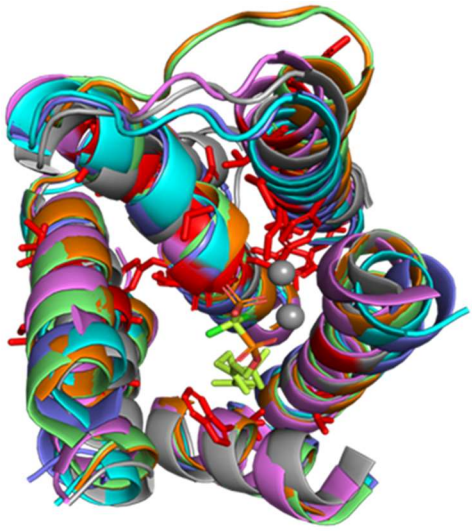


Fig. 5.5. *Catalytic site helices of SSQs color-coded with the scheme: SaSQS1 (PDB ID: 6K16)- grey; TEAS (PDB ID: 5IK0)- cyan and AaBOS (PDB ID: 4FJQ)- pink. PMRs of these proteins are represented as red colored sticks. Ligand is shown as sticks and* Mg^{2+} *as grey spheres.*

The sICs were mapped on MSA and respective structures of eight plant sesquiterpene synthases (α-bisabolene synthase, Delta cadinene synthase, Vetispiradiene synthase, TEAS, β-farnesene synthase, amorpha-4,11-diene synthase 1, amorpha-4,11-diene synthase 2 or AaBOS and

SaSQS1) (Fig. 5.6, 5.7). For all of the sesquiterpene synthases considered here, top eight eigenmodes were used for determining the sectors. Interestingly, first five sICs, except sIC3, can be mapped to the conserved residues, encompassing functionally important motifs such as DDXXD, RXR and NSE/DTE, across all of the eight SSQs (Fig. 5.6). Rest of the three sICs (sIC6 to sIC8) largely map on the divergent residues at the catalytic pocket of SSQs. Since different sICs are suggested to indicate divergent functional states of proteins, it was checked whether these sectors indicate PMRs [59,133,135,137–139]. Surprisingly, the sICs of SSQs with known PMRs do not show any clustering of their PMRs. Although a subset of PMRs appears to fall in few sICs, they do not show any particular pattern such as the hierarchy of sICs (Fig. 5.6). For example, PMRs of Vetispiradiene synthase (A298 & L446) and of TEAS (V291 & I439) can be mapped to sIC8. However, in both, just these two residues of the PMR set are not adequate enough to modulate their product[135–137]. Even the intra-sector divergent residues do not show any well-defined correlation between their respective sICs and their ability to modify the product. This can be seen for AaBOS, where members of sIC6, Q379 and A396, exhibit sequence divergence within the sector but do not change the cognate product of the enzyme, when mutated.

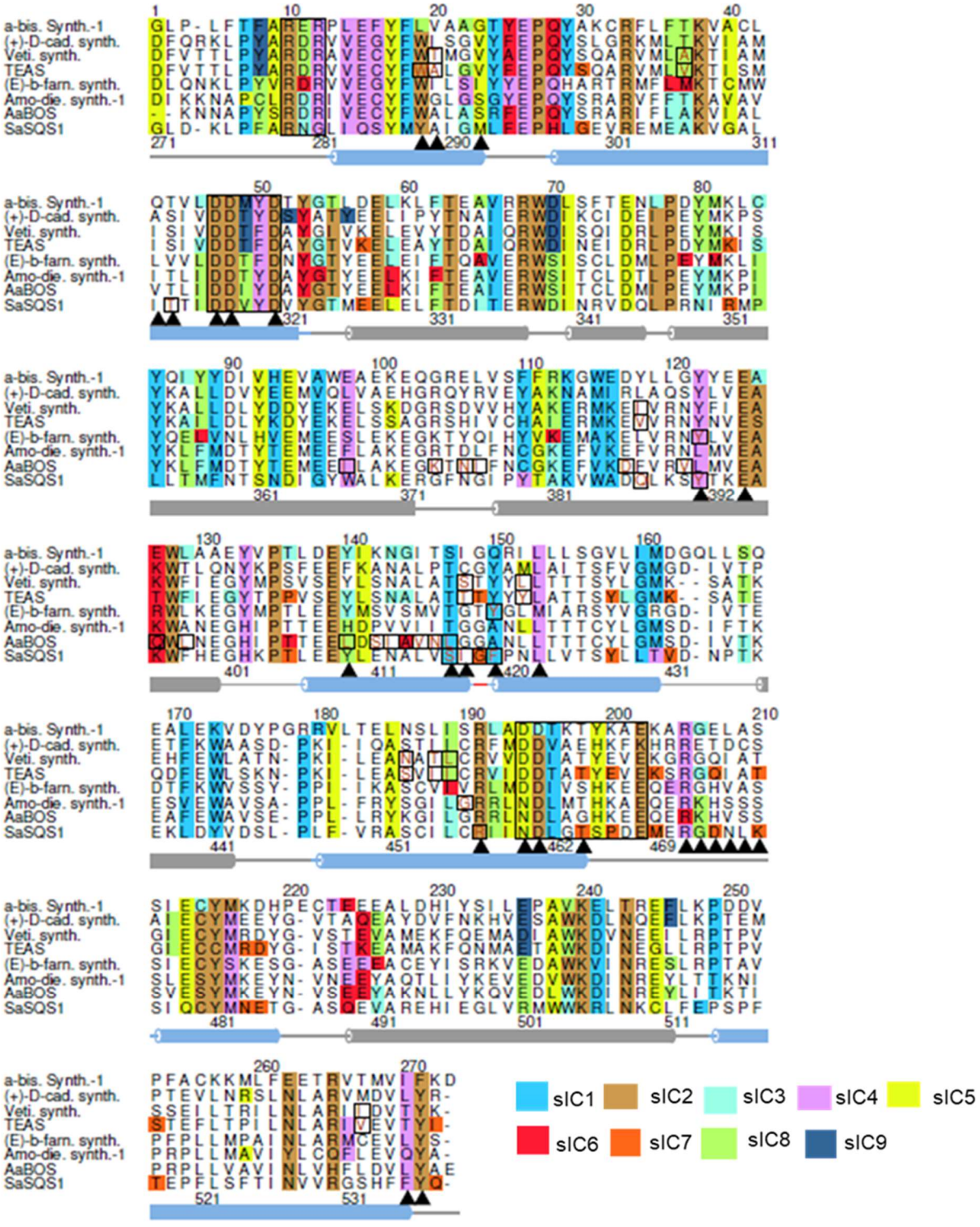


Fig.5.6. Sequence based sectors (sICs). *sICs mapped on multiple sequence alignment of plant sesquiterpene synthases (α domain). Conserved motifs like DDXXD, NSE/DTE and RXR are shown in boxes. The core and adjacent catalytic site helices, corresponding to the SaSQS1 structure, are shown in blue and grey colors, respectively. The residues corresponding to the sectors (sICs), from sIC1-sIC9, are highlighted as per the color-code given in the figure. The residues of SaSQS1 that lie within 6Å from the ligand are marked with arrows. The*

known PMRs of SSQs are highlighted in brown color with box. For SaSQS1 these boxes indicate the residues considered for mutagenesis study.

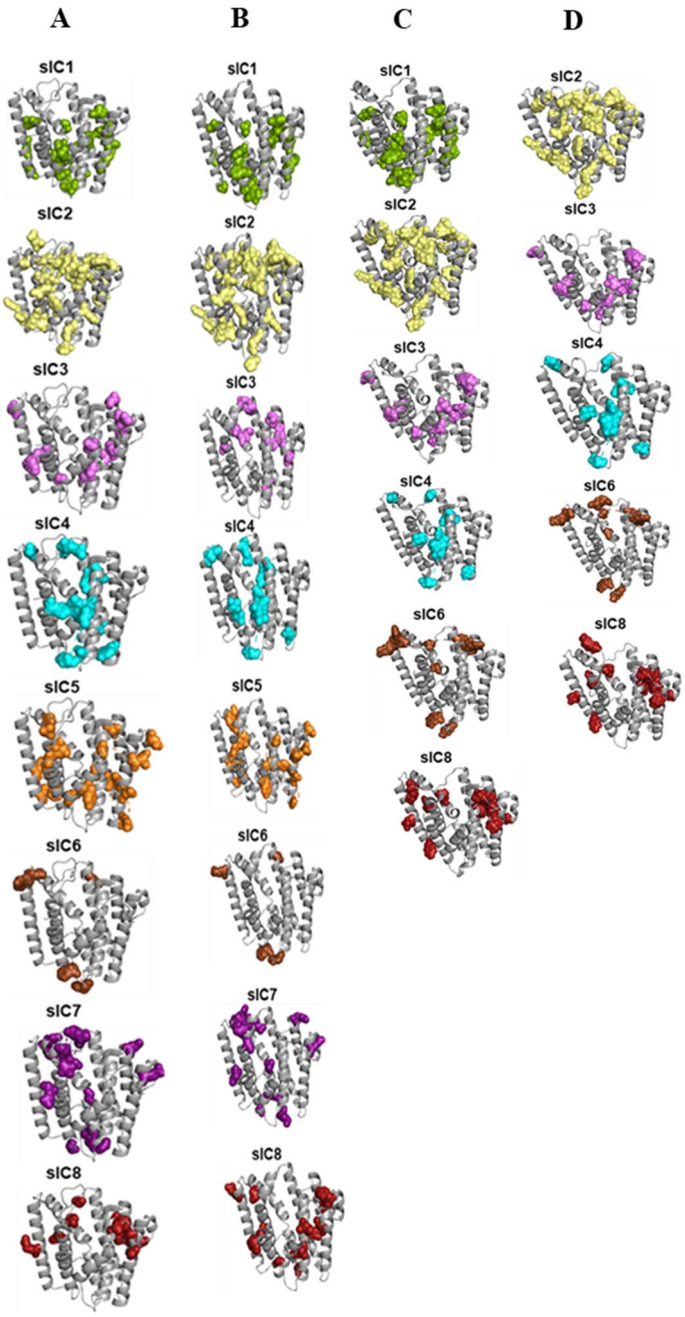


Fig. 5.7. *sICs mapped on the α domains of: (A) SaSQS1 (PDB ID: 6K16), TEAS (PDB ID: 5IK0), AaBOS (PDB ID: 4FJQ) and AaBOS^{M2} (PDB ID: 4GAX). Color scheme for different sICs is as follows: sIC1- green; sIC2- yellow; sIC3- pink; sIC4- cyan; sIC5- orange; sIC6- brown; sIC7- purple; sIC8- red.*

In summary, sICs help to identify functionally important residues of the SSQ family and also some of the PMRs occur in few sICs. However, the PMRs appear to be scattered amongst different sICs and do not follow any particular pattern. Thus, sequence based statistical coupling analysis unravels positional correlation between the residues, however neither

provides any systematic clues on PMRs nor provides rationale for the observed effect of PMRs in the SSQs analyzed here. Perhaps, non-availability of sufficient biochemical data on different SSQs could be the reason for this limitation of sSCA.

5.3.2 Dynamical sectors provide leads to identify product-defining residues:

The PMRs of TEAS[55,138] and AaBOS[133], when mapped on their respective structures, lie within a ring-shaped shell at the center of the binding pocket (Fig. 5.5). Interestingly, some of the PMRs are located beyond 6 Å from the center of the binding pocket and unlikely to interact with FPP (Fig. 5.5). Thus, as mentioned earlier, even a combination of structure-sequence comparison does not provide unified rules for identifying the PMRs for sesquiterpene synthases. Besides sequence and three-dimensional structure, the dynamical features of proteins are equally consequential in dictating their function. Furthermore, for few sesquiterpene synthases dynamical rearrangement of catalytic pocket helices is seen upon ligand binding. Therefore, the role of dynamical coupling was explored between the catalytic pocket residues of sesquiterpene synthases in defining PMRs. To achieve this, 300 ns of MD simulation on the FPP bound structures of SaSQS1, TEAS, AaBOS and mutant *M2* of AaBOS was performed. To enhance statistical significance, all the MD simulations were performed in triplicates and subsequently the trajectories were pooled for further analysis. Trajectories were sampled at every 50 ps to obtain 18000 frames for the analysis. For the statistical coupling analysis, matrix of covariance vectors, of residues belonging to the α domain were generated from the MD trajectories. The sectors in sequence based statistical coupling analysis, sICs, group the residues that are coevolving. However, in the current context, the dynamical sectors (dICs) indicate set of residues, which are calculated from the eigenmodes of DCCM (Fig. 5.8B). Although, there are about seven to ten dICs observed for different SSQs, the ones that contain only the subset of the α domain and not the entire domain are considered for further analysis. The hierarchy of the dICs for different SSQs are shown in Fig. 5.8C-F. Unlike for the PDZ domain[155], the sICs and dICs of SSQs hardly overlap (Fig. 5.7, Fig. 5.8). Furthermore, despite having homologous structures, TEAS and AaBOS (RMSD is 2.81) do not have identical dICs (Fig. 5.8D, E). This is contrary to the sICs, where the sectors are found to be largely common for all the analyzed SSQs. Since, sICs are obtained from the residue conservation in the sequence alignment, the sectors here reflect the frequency of positional conservation and thus expected not to vary largely across different proteins of any particular family.

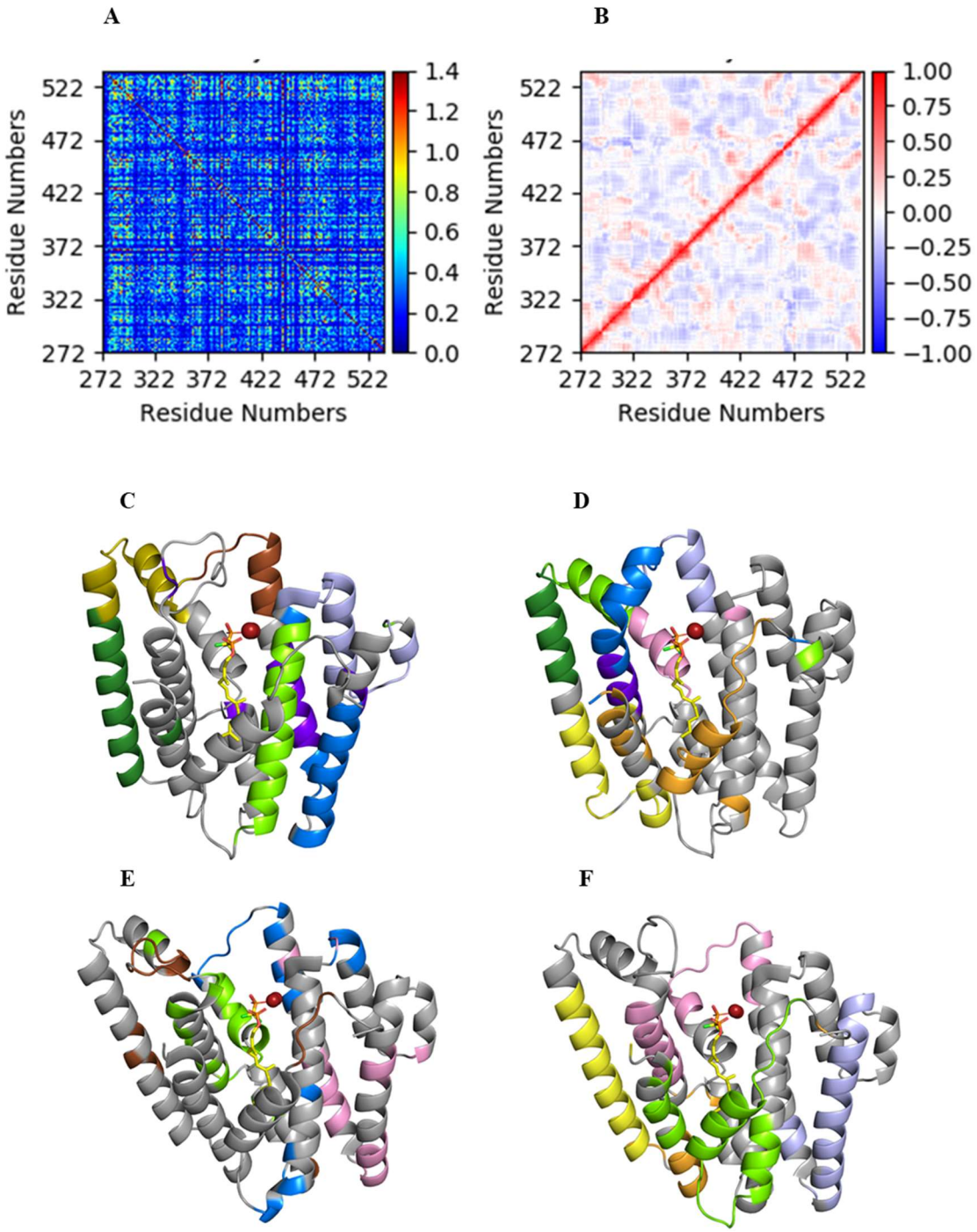


Fig. 5.8. Dynamical sectors of sesquiterpene synthases. *(A) Sequence and (B) MD simulation based covariance matrices for SaSQS1. Dynamical sectors mapped on the α domain structures (C) SaSQS1(PDB ID: 6K16) (D) TEAS (PDB ID: 5IK0) (E) AaBOS (PDB ID: 4FJQ) (F) AaBOS^{M2} (PDB ID: 4GAX). For all the structures following scheme for dICs color coding is used: dIC1- dark blue, dIC2- brown, dIC3- lime, dIC4- pink, dIC5- yellow, dIC6- violet, dIC7- orange, dIC8- green, dIC9- purple, dIC10- golden. Ligand & Mg^{2+} ions are shown in ball and stick.*

Interestingly, the PMRs of TEAS and AaBOS are found to be clustering within particular dICs (Table 5.1). However, the hierarchy of these dICs between the two structures is not unique. For example, the PMRs of AaBOS are found to be falling in dIC3 (L392, S394, I395, A396, V397, N398, L399) and dIC4 (I350, I359, D369). Similarly, the PMRs of TEAS belong to dIC7 (W273, A274, V516) and dIC9 (I438 & I439). Further, the observed pattern was validated by performing similar analysis on the *M2* mutant structure of AaBOS, which produces its non-cognate product, γ-humulene. For this, all the structural parameters, like position of Mg^{2+}, ligand, and simulation conditions were kept identical to that of *wild* type protein. Clearly, the composition of dICs of the mutant is significantly different from that of the *wild* type AaBOS (Table 5.1). This indicates that the dynamical coupling between the residues of the binding pocket largely influence the product. Since, mutant is still capable of producing terpenoids, although of different kind, we are tempted to argue that the mutant has been essentially transformed into a new type of terpene synthase. Further, the residues that can alter the products of this new enzyme could be identified from the dICs of the mutant enzyme. Another interesting observation emerged from our analysis is that the residues which are far from the binding pocket yet influence the product formed belong to the PMR defining dICs. Therefore, we can argue that the dynamical coupling also hints at structural allostery in regulating the type of products. Thus, a pattern for PMRs, in terms of their dynamical sectors, was observed for two highly sequentially divergent SSQs (Fig. 5.1), which was otherwise not clear from sequence and structure comparison studies.

	sICs Residues	dICs Residues	sICs Residues	dICs Residues	sICs Residues	dICs Residues	sICs Residues	dICs Residues
	SaSQS1		TEAS		AaBOS *(wt)*		AaBOS *(M2)*	
IC1	P276, M287 ,L293, L298, T335, L354, F358, N362, G365, K382, W384, A385, S416, F419, L429, L439, Y441, P446, L462, R506, E513	Y319, T356, M357, F358, N359, T360, S361, N362, D363, I364, G365, Y366, W36, A368, L369, K370, E371, N375, G376, I377	P261, F272, Y278, Y283, I320, Y339, L343, K347, E350, H364, E367, M369, K370, T401, Y404, G414, M415, F423, W425, P430, T446, D490, P498	L260, D444, D445, T446, A447, T448, Y449, E450, V451, E452, K453, S454, R455, G456, Q457, I521	P259, F270, R276, Y281, I318, Y337, M341, T345, E348, C362, E365, V367, K368, L399, A402, G412, M413, F422, W424, P429, L445, D489, T497	P279, Q280, S282, R285, V291, G306, Y308, E309, E310, L374, V376, E377, L381, E383, H385, I386, P387, T388, T389, I416	NU	NU
IC2	R279, Y288, P296, R302, K307, D316, D317,D320, Y322, G323, E327, T332, R337, W338, L346, P347, E394, A395, W397,P404, E408, R457, D461, E468, C480, Y481, K505, N508, N527, R530, Y536	T392, K393, E394, A395, K396, W397, F398, H399, E400, G401, H402, K403, P404	R264, W273, P281, R287, K292, D301, D302, D305, Y307, T309, E312, T317, R322, W323, L331, P332, E379, S380, W382, P389, E393, R441, D445, E452, C464, C465, K489, N492, N511, R514, Y520	NU	R262, W271, P279 , R285, K290, D299, D300, D303, G306, T307, E310, T315, R320, W321, I329, P330, E377, A378, W380, P387, E391, R440, D444, E451, S463, Y464, K488, N491, N510, H513, Y519	Y260, S261, R262, D263, S414, T418, E451, Q452, E453, K455, H456, V457, S458, S459, S460, Q482, E484, L486	R262, W271, P279 , R285, K290, D299, D300, D303, G306, T307, E310, T315, R320, W321, I329, P330, E377, A378, W380, P387, E391, R440, D444, E451, S463, Y464, K488, N491, N510, H513, Y519	NU
IC3	N280, G291, E305, A306, T324, I334, N349, I350, M357, A381, T435, L455, V491, W504	L298, G299, E300, V301, R302, E303, M304, E305, A306, K307, V308, G309, A310, L311, I312, T313, T314, I315, D316, D317, V318, L346	V300, G308, L313, Y316, S338, A341, F383, G386, Y394, N397, A398, L413, V442, G456	T258, V391, S392, E393, T460, G461, I462, E463, C464, C465, M466, R467, D468, Y469, G470	D263, A274, L288, A289, A317, Y332, M333, F340, K364, T418, L438, Y474, W487	E391, L392, S394, I395, A396, V397, N398, L399, G401, K419, F422, W424, P429, L432, K435, L438, L442, A446, H448, K449, E462	D263, A274, L288, A289, A317, Y332, M333, F340, K364, T418, L438, Y474, W487	Y260, S261, R262, D263, R264, I265, V266, V267, C268, Y269, F270, W271, A272, L273, A274, S275, R276, F277, E278, P279, Q280, Y281, S282, R283, A284, R285, I286, F287, L288, A289, N510, L511, V512, H513, F514, L515, L518
IC4	G281, I283, Q284, S285, E295, Y319, W367, Y379, Y391, H402, L422, T430, R471, M482, F535	NU	R266, V268, E269, C270, E280, F304, E352, Y376, Y387, L407, R455, M466, T473, I515, T519	F304, Y394, L395, S396, N397, A398, L399, A400, T401, T402, T403	R264, V266, E267, C268, E278, Y302, I350, L374, H385, L405, M465, L518	T307, I323, K338, L339, F340, M341, E349, I350, L351, A352, K353, E354, G355, I359, F360, N361, C362, G363, K364, E365, F366, V367, D369, A378	R264, V266, E267, C268, E278, Y302, I350, L374, H385, L405, M465, L518	M375, V376, Q379, W380, N382, E383, G384, H385, I386, P387, T388, T389, E390, E391, L392, D393, R433, Y434, K435, G436, I437, L438, G439, R440, R441, L442, N443, D444, L445, A446
IC5	G271, A278, Y286, M292, G309, L328, I340, D344, S361, D363, K370, T380, L410, F448, A451, I459, N460, S465, Q479, W503, C510	NU	D255, A263, Y271, I294, I325, D329, Y346, D348, S355, A365, L395, I432, A435, I443, D444, Y449, E463, A487, G494	E434, A435, T486, A487, W488, K489, D490, I491, N492, E493, G494, L495, L496, R497, P498, P500, V501, S502, T503, E504, F505, L506, T507	NU	NU	NU	K476, N477, L478, L479, Y480, K481, Q482, V483, E484, D485, L486, W487, K488, D489, I490, N491, R492, E493, Y494, L495, A521
IC6	F294, H297, K396, Q489, E490	V321, Y322, G323, T324, M325, E326, E327, L328, E329, L330, F331, T332, D333, I334, T335, E336, R337, W338, D339, I340, N341, R342, V343, D344, P347	F279, Q282, T381, K474	V378, E379, S380, T381, W382, F383, I384, E385, G386, Y387, T388, P389, P390, I458, A459	F277, Q280, Y305, L311, Q379, T388, A396, R454, E472, E473	NE	F277, Q280, Y305, L311, Q379, T388, A396, R454, E472, E473	E331, Y332, M333, K334, P335, I336, Y337, K338, L339, F340, M341, D342, T343, Y344, T345, E346, M347, E348, E349, I350, A352, K353, G355, I359, F360, N361, C362
IC7	G299, E326, R351, T405, G418, Y427, V431, T464, P466, M469, D473, K476, N483 , E484, T518, Q537	NU	S284, K311, K336, P390, T403, Y412, K416, T448, E450, K453, Q457, T460, R467, D468, S502, I521	Y262, A263, R264, D265, R266, V267, V268, E269, C270, Y271, F272, W273, A274, L275, G276, V277, Q285, A286, L512, A513, R514, V516, E517, Y520	NU	NE	NU	A258, S472, P429, I496, T497, T499, I500, P501, R502, P503, L504, L505, V506, A507, V508
IC8	F277, I315, V318, F331, D339, P353, T356, Y409, I458, G472, R501	E494, H495, I496, E497, G498, L499, V500, R501, M502, W503, W504, K505, R506, L507, N508, K509, C510, L511, F512, G531	D265, G276, L290, V291, A319, Y334, M335, I342, I366, T419, I439, I462, E475, W488	I471, S472, T473, K474, E475, A476, M477, A478, K479, F480, Q481, N482, M483	Y260, I298, I301, F314, S322 , I336, L339, L392, R441, K455, E484,Y494	NE	Y260, I298, I301, F314, S322 , I336, L339, L392, R441, K455, E484,Y494	NE
IC9	NE	L354, L355, P378, Y379, T380, A381, K382, V383, W384, A385, D386, Q387, L388, P420, P466, D467	NE	V437, I438, I439, C440, R441, V442, I443	NE	NE	NE	NE
IC10	NE	T405, L406, E407, I478, Q479, C480, Y481, M482, N483, E484 , T485, Q486, A487, S488, Q489, E490, V491, A492, R493	NE	NE	NE	NE	NE	NE

Table 5.1. *Residue composition of sequence (sICs) and dynamics (dICs) based sectors of SaSQS1, TEAS, AaBOS and AaBOS^{M2}. Known product modulating residues (PMRs) of these sesquiterpene synthases are highlighted in red. The residues of SaSQS1 which were mutated are shown in blue color. Sectors labelled as NU (Not Unique) were not found to be independent sectors and mapped the entire α domain of SSQs. The sectors which were not found in the SSQs are marked as NE (Not existing).*

Although, dICs cluster the PMRs, their hierarchy varies amongst structures. For instance, PMRs of TEAS are in dIC7 and dIC9 whereas for AaBOS they are found in dIC3 and dIC4 (Table 5.1). Therefore, to further narrow down on the PMR containing dICs, we augmented dynamical information with structural and sequence data. Ligand vicinity and hydrophobicity of the catalytic site residues are also known to play important role in the reaction cascade by

protecting the carbocation intermediates[12,140] and hence the product formed[140,156–158]. Therefore, we calculated average kd hydrophobicity values (hydrophobicity index, HI) and vicinity index (VI) for all dICs, as defined in the methods section. Clearly, dICs with positive HI and VI (>0) are likely to harbor the PMRs and interestingly, that is the case for both TEAS and AaBOS (Table 5.2). Thus, it appears that the group of catalytic site residues of plant SSQs, which are dynamically coupled and have positive HI and VI values modulate their product. In line with our observations for TEAS and AaBOS, the residues of SaSQS1, which were mutated, do not cluster in any of the dICs (Table 5.1). However, the residues, which abrogate the activity of SaSQS1 on substitution, such as T313, belong to dIC3. Interestingly, dIC3 has positive HI and VI value greater than zero, which is the criteria developed for identifying potential PMRs. Further, to test the technique developed here on SaSQS1, a combination of residues from dIC3 were tried. Unfortunately, these SaSQS1 mutants were found to be insoluble, forming inclusion bodies. Although, a direct validation of the technique was not possible with SaSQS1, mutation of individual members of dIC3 suggest that indeed they play significant role in influencing the product in terms of its quantity.

SSQ	dIC1	dIC2	dIC3	dIC4	dIC5	dIC6	dIC7	dIC8	dIC9	dIC10
SaSQS1	-0.43 (0.00)	-1.85 (0.00)	0.47 (0.54)	NU	NU	-0.54 (0.00)	NU	-0.29 (0.05)	-0.18 (0.12)	-0.47 (0.00)
TEAS	-0.97 (0.00)	NU	-0.25 (0.00)	0.57 (0.45)	-0.30 (0.00)	0.19 (0.00)	0.18 (0.46)	-0.39 (0.00)	2.84 (0.43)	NE
AaBOS (*wt*)	-0.36 (0.05)	-1.70 (0.00)	0.24 (0.33)	0.46 (0.04)	NU	NE	NE	NE	NE	NE
AaBOS (*M2*)	NU	NU	0.14 (0.40)	-0.76 (0.13)	-0.88 (0.00)	-0.26 (0.00)	1.14 (0.06)	NE	NE	NE

Table 5.2. *Hydrophobicity index and vicinity index (in the parenthesis) values of different dICs of sesquiterpene synthases. The sectors which contain PMRs are highlighted in orange. Putative PMRs containing dICs, obtained from the current analysis, are highlighted in violet. NU include the dICs which are not unique and mapped the entire α domain of SSQs, while NE are not existing sectors in a particular SSQ.*

5.3.3 Distinct dynamical sectors of the mutants:

Next, the sector analysis was done for the two important mutants that do not form sesquisabinene, T313S and G418A in order to observe the changes in the dynamics of the

mutants. Similar methodology was applied to study the sequence and dynamics based sectors of the mutants.

As anticipated, the sequence based sectors of both the mutants were found to be similar to the SaSQS1 *wild* type, except for IC6 (Table 5.3). Hence, the dynamical sectors of the mutants were identified by using the similar docking and simulation conditions as the *wild* type were used for studying the variations amongst them. Interestingly, the dICs of the *wild* type and the mutants are not identical and their residue composition is markedly different as compared to the *wild* type (Fig. 5.9, Table 5.3). The mutant T313S has six unique dICs (dIC1, dIC2, dIC3, dIC6, dIC7 and dIC8) and G418A constitute five unique dICs (dIC, dIC2, dIC3, dIC5 and dIC6). Although, the dICs of the mutants do not overlap with the *wild* type, dIC1 of both the mutants have few residues in common (Table 5.3). The difference in the dynamical sectors of the mutants and the *wild* type thus shows differential dynamics of the mutants as compared to the wild type SaSQS1, which could be responsible for making the enzyme inactive towards its cognate product. Thus, we can conclude that the residue dependent dynamics plays critical role in influencing the product formed. In other words, mutations in the binding pocket could have converted SaSQS1 into another type of sesquiterpene synthase, rather than rendering the enzyme inactive. However, this merits further investigations.

	sICs Residues	dICs Residues	sICs Residues	dICs Residues	sICs Residues	dICs Residues
	SaSQS1 (*wt*)		T313S		G418A	
IC1	P276, M287 ,L293, L298, T335, L354, F358, N362, G365, K382, W384, A385, **S416, F419**, L429, L439, Y441, P446, L462, R506, E513	Y319, T356, M357, F358, N359, T360, S361, N362, D363, I364, G365, Y366, W36, A368, L369, K370, E371, N375, G376, I377	P276, M287 ,L293, L298, T335, L354, F358, N362, G365, K382, W384, A385, **S416, F419**, L429, L439, Y441, P446, L462, R506, E513	N280, **G281**, **L282**, I283, **Q284**, **M287**, **I290**, **G291**, **L298**, G299, E300, V301, R302, E303, M304, E305, A306, K307, V308, G309, **F419**, P420, N421, L422, L423, V424,T425, S426	P276, M287 ,L293, L298, T335, L354, F358, N362, G365, K382, W384, A385, **S416, F419**, L429, L439, Y441, P446, L462, R506, E513	F277, R279, N280, **G281**, **L282**, **Q284**, S285, Y286, **M287**, Y288, A289, **I290**, **G291**, M292, L293, F294, E295, P296, H297, **L298**, V449
IC2	R279, Y288, P296, R302, K307, D316, D317,D320, Y322, G323, E327, T332, R337, W338, L346, P347, E394, A395, W397,P404, E408, **R457**, D461, E468, C480, Y481, K505, N508, N527, R530, Y536	T392, K393, E394, A395, K396, W397, F398, H399, E400, G401, H402, K403, P404	R279, Y288, P296, R302, K307, D316, D317,D320, Y322, G323, E327, T332, R337, W338, L346, P347, E394, A395, W397,P404, E408, **R457**, D461, E468, C480, Y481, K505, N508, N527, R530, Y536	K403, P404, T405, L406, E407, E408, Y409, L410, K476, S477, I478, Q479, C480, Y481, M482, N483, E484, T485, G486, A487, S488, A492	R279, Y288, P296, R302, K307, D316, D317,D320, Y322, G323, E327, T332, R337, W338, L346, P347, E394, A395, W397,P404, E408, **R457**, D461, E468, C480, Y481, K505, N508, N527, R530, Y536	G299, E300, E303, M304, E305, A306, K307, V308, G309, L311, I312, P347, R348, I350, K382
IC3	N280, G291, E305, A306, T324, I334, N349, I350, M357, A381, T435, L455, V491, W504	L298, G299, E300, V301, R302, E303, M304, E305, A306, K307, V308, G309, A310, L311, I312, **T313**, T314, I315, D316, D317, V318, L346	N280, G291, E305, A306, T324, I334, N349, I350, M357, A381, T435, L455, V491, W504	E295, D432, T464, S465, P466, D467, E468, M469, E470, R471, G472, D473, L475	N280, G291, E305, A306, T324, I334, N349, I350, M357, A381, T435, L455, V491, W504	T380, V383, W384, D386, **Q387**, L388, K389, S390, **Y391**, T392, K393, E394, A395, K396, W397, H399, E400, G401, H402, T405, L406, L410
IC4	G281, I283, Q284, S285, E295, Y319, W367, Y379, **Y391**, H402, L422, T430, R471, M482, F535	NU	G281, I283, Q284, S285, E295, Y319, W367, Y379, **Y391**, H402, L422, T430, R471, M482, F535	NU	G281, I283, Q284, S285, E295, Y319, W367, Y379, **Y391**, H402, L422, T430, R471, M482, F535	NU
IC5	G271, A278, Y286, M292, G309, L328, I340, D344, S361, D363, K370, T380, L410, F448, A451, I459, N460, S465, Q479, W503, C510	NU	G271, A278, Y286, M292, G309, L328, I340, D344, S361, D363, K370, T380, L410, F448, A451, I459, N460, S465, Q479, W503, C510	NU	G271, A278, Y286, M292, G309, L328, I340, D344, S361, D363, K370, T380, L410, F448, A451, I459, N460, S465, Q479, W503, C510	Q345, L346, N362, E497, L499, V500, M502, W503, K505, R506, L507, N508, K509, C510, L511, F512
IC6	F294, H297, K396, Q489, E490	V321, Y322, G323, T324, M325, E326, E327, L328, E329, L330, F331, T332, D333, I334, T335, E336, R337, W338, D339, I340, N341, R342, V343, D344, P347	F294, H297, L328, K396, A413, Q489, E490	Y288, A289, L293, K389, S390, **Y391**, T392, K393, E394, A395, K396, W397, F398, H399, E400, G401, H402	F294, H297, L328, K396, A413, Q489, E490	L414, V424, L428, D432, P434, T435, K436, E437, K438, L439, D440, Y441, D443, L447, N460, D461, L462, G463, T464
IC7	G299, E326, R351, T405, **G418**, Y427, V431, T464, P466, M469, D473, K476, N483 , E484, T518, Q537	NU	G299, E326, R351, T405, **G418**, Y427, V431, T464, P466, M469, D473, K476, N483 , E484, T518, Q537	M292, Y322, G323, T324, M325, E326, E327, L328, E329, L330, F331, T332, D333, I334, T335, E336, R337, W338, D339, I340, N341, R342, V343	G299, E326, R351, T405, **G418**, Y427, V431, T464, P466, M469, D473, K476, N483 , E484, T518, Q537	NU
IC8	F277, I315, V318, F331, D339, P353, T356, Y409, I458, G472, R501	E494, H495, I496, E497, G498, L499, V500, R501, M502, W503, W504, K505, R506, L507, N508, K509, C510, L511, F512, G531	F277, I315, V318, F331, D339, P353, T356, Y409, I458, G472, R501	**Q387**, E411, N412, A413, L414, V415, **S416**, **I417**, **G418**, S452, C453, I454, L455, C456, **R457**, I458, I459, N460, D461, L462, G463, R530, H533, F534, Y536, Q537	F277, I315, V318, F331, D339, P353, T356, Y409, I458, G472, R501	NE
IC9	NE	L354, L355, P378, Y379, T380, A381, K382, V383, W384, A385, D386, **Q387**, L388, P420, P466, D467	NE	NE	NE	NE
IC10	NE	T405, L406, E407, I478, Q479, C480, Y481, M482, N483, E484 , T485, G486, A487, S488, Q489, E490, V491, A492, R493	NE	NE	NE	NE

Table 5.3. *Residue composition of sequence (sICs) and dynamics (dICs) based sectors of wild type SaSQS1, T313S and G418A. The residues of SaSQS1 which were mutated are shown in blue color. The residues common in dIC1 of both the mutants, T313S and G418A, are shown in red color. Sectors labelled as NU (Not Unique) were not found to be independent sectors and mapped the entire α domain of SSs. The sectors which were not found in the SSQs are marked as NE (Not existing).*

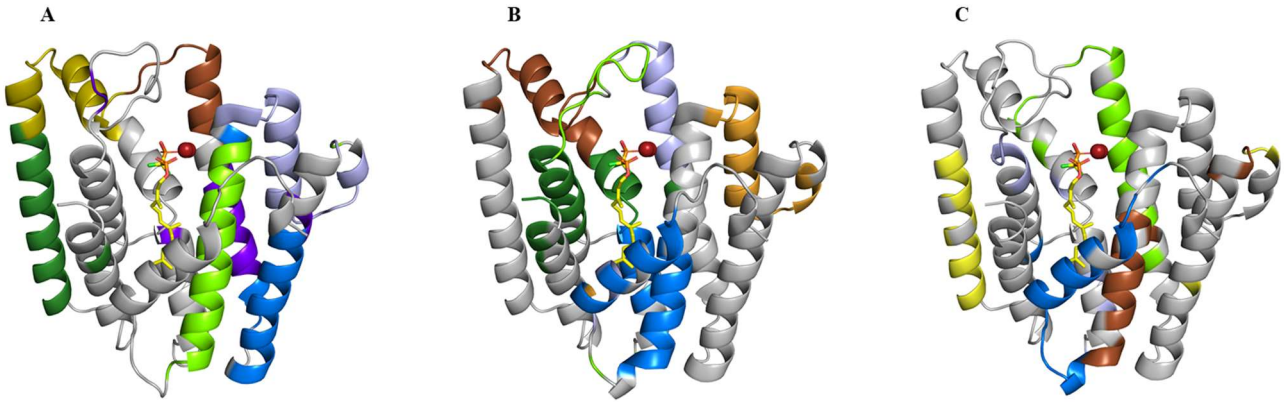


Fig. 5.9. Dynamical sectors of. *(A) Wild type SaSQS1(PDB ID: 6K16) (B) T313S (PDB ID: 7E9R) (C) G418A (PDB ID: 7E6W) (F). For all the structures following scheme for dICs color coding is used: dIC1- dark blue, dIC2- brown, dIC3- lime, dIC4- pink, dIC5- yellow, dIC6- violet, dIC7- orange, dIC8- green, dIC9- purple, dIC10- golden. Ligand & Mg^{2+} ions are shown in ball and stick.*

If introduction of mutations has converted SaSQS1 into a new type of sesquiterpene synthase with altered product, then there are has to be product specificity defining sectors in the mutant enzymes. Therefore, dynamical sectors of the mutants were further checked for their hydrophobicity and vicinity indices in order to determine the putative sector of these mutants that can influence the product formation. As described in the above section, the dIC with the positive HI and VI values greater than zero could be the potential sectors that define the product specificity. Interestingly, it can be seen that in the case of T313S, dIC1 and dIC8 fulfill this criterion while in the case of G418A, dIC1 is the sector with positive HI and VI values (Table 5.4). In contrast to the wild type, wherein the mutants were found to be scattered amongst different dICs, the mutant enzymes have dICs that group the mutant residues of T313S and G418A. For example, dIC1, dIC6 & dIC8 of T313S and dIC3 of G418A group the mutations introduced in the enzyme. From these observations we can infer that the putative PMR containing sectors of the mutants are different as compared to the *wild* type, where dIC3 was observed to be the key dynamical sector. However, in the case of both the mutants, it can be seen that dIC1 is the common putative sector with positive HI and VI values. Additionally, this sector is also observed to map few common residues of both the mutants (Table 5.3). Thus, dIC1 of both the mutants might play role in influencing the product formed by the mutant enzymes.

	dIC1	dIC2	dIC3	dIC4	dIC5	dIC6	dIC7	dIC8	dIC9	dIC10
SaSQS1 (*wt*)	-0.43 (0.00)	-1.85 (0.00)	0.47 (0.54)	NU	NU	-0.54 (0.00)	NU	-0.29 (0.05)	-0.18 (0.12)	-0.47 (0.00)
T313S	0.40 (0.68)	-0.58 (0.00)	-1.79 (0.00)	NU	NU	-1.36 (0.12)	-0.47 (0.04)	0.24 (0.54)	NE	NE
G418A	0.35 (0.48)	-0.30 (0.47)	-0.96 (0.14)	NU	-0.04 (0.00)	-0.54 (0.16)	NU	NE	NE	NE

Table 5.4. *Hydrophobicity index and vicinity index (in the parenthesis) values of wild type SaSQS1 and the mutants, T313S & G418A. dICs, that contain the putative residues affecting the product formation obtained from the current analysis, are highlighted in violet. NU include the dICs which are not unique and mapped the entire α domain of SSQs, while NE are not existing sectors.*

5.4 Conclusion:

A number of studies have been done to identify the PMRs of sesquiterpene synthases. However, these sequence and structure based methods were less generic and involved bottom up approach of testing extensive combination of residues. Hence, we employed statistical coupling analysis to validate the biochemically known PMRs of sesquiterpene synthases and identify the product specificity determinants of SaSQS1. The sequence based sectors, obtained from SCA, aided in identification of coevolving residues and also mapped the conserved motifs of sesquiterpene synthases under study. However, they did not show any clustering of the previously identified PMRs. Therefore, we analyzed the dynamically coupled residues of these synthases and SaSQS1 using SCA, in pursuit to unravel the patterns that provide leads in identifying the PMRs of the enzymes. Clearly, the dynamical sectors formed cluster the PMRs for the enzymes with known product influencing residues. However, the hierarchy or the unique occurrence of these sectors was not uniform amongst the sesquiterpene synthases. Therefore, to pin-down the most putative dynamical sector that harbors the most putative PMRs, the dynamical information was combined with the sequence and structure data by calculation of hydrophobicity and vicinity indices of the dynamical sectors. This novel approach correctly predicts the biochemically validated PMRs of sesquiterpene synthases and also provides leads on identification of PMRs for SaSQS1. We also extended this approach to identify the residues of the mutants, T313S and G418A that could define the product formed by the mutant enzymes.

Chapter 6

Conclusion

Secondary metabolites are a crucial group of natural products that play an important role in numerous biological activities of plants. These compounds function as defense system against biotic and abiotic stresses and attractants for pollinators and signaling molecules. They are also commercially important as they are largely used as food and aromatic additives and ingredient for medicines. One such important class of secondary metabolites are terpenes or isoprenoids. Terpenes are a diverse group of natural products found in plants, fungi and bacteria. In plants, these compounds can function as primary and secondary metabolites. A number of terpenes are used in the fields of medicine, agriculture, nutraceuticals and are also used for flavors and fragrances. Depending on the number of fundamental isoprene units (C_5), terpenes can be classified as monoterpenes (C_{10}), sesquiterpenes (C_{15}), diterpenes (C_{20}), triterpenes (C_{30}) and tetraterpenes (C_{40}). Till date, more than 7000 chemically diverse forms of sesquiterpenes are known, with greater than 300 different carbon skeletons. Sesquiterpenes find myriad biotechnological applications as pharmaceuticals, highly valued essential oils, flavoring agents and fragrances. Sesquiterpene synthases are a class of enzymes that are responsible for this huge chemical diversity of sesquiterpenes. These synthases catalyze the cyclization of farnesyl pyrophosphate (FPP) to different types of sesquiterpenes. Despite utilising the same substrate and exhibiting significant sequence and structural homology, these enzymes form different products. The enzyme binds to the linear flexible isoprenoid substrate to produce highly reactive carbocation intermediates, which subsequently leads to formation of particular skeletal forms. The contour of the active site serves as a template for the catalysis and controls the conformations attained by various carbocation intermediates, providing product diversity. The members of this class of enzymes share a conserved catalytic mechanism, however, there is a lacuna in understanding the structural basis for the conversion of the same substrate to produce varied products. Structural studies of these enzymes have provided insights into the residues lining the active site which play role in stabilizing the carbocation intermediates formed during catalysis. This thesis focuses on structural studies of sesquisabinene synthase 1 (SaSQS1), a sesquiterpene synthase found in Indian sandalwood, which catalyzes the cyclization of FPP to form sesquisabinene, a key component of sandalwood oil. Sandalwood oil, primarily

comprising of sequiterpenes, has great commercial value due to its pleasant and woody fragrance. The oil has applications in industries like cosmetics, perfumery and aromatherapy and has also shown to act as an anti-inflammatory, antidepressant, antifungal and antiseptic agent. Due to huge commercial importance of this oil, understanding the mechanism of action and product specificity of SaSQS1 is a crucial aspect to limit the direct use of plant sources for the extraction and purification of these metabolites and enhance the product yield.

To determine the structural basis of SaSQS1 catalyzed sesquisabinene formation and its product specificity, we elucidated the crystal structure of the enzyme. The structure provides insights to the domain architecture of the enzyme which maps the conserved domains of this class of enzymes, DDXXD and NSE/DTE motifs. Despite number of co-crystallization trials of the enzyme with the substrate analog, ligand bound structure of the enzyme was not obtained. The structure of SaSQS1 highlights the conformational changes of the protein when compared to other homologous sesquiterpene synthases. Also, conformational deviations are observed in these enzymes upon ligand binding upon structural comparison of *apo* and ligand bound forms of few sesquiterpene synthases. However, these structural deviations appear to have minimal influence on the mechanistic aspect of SaSQS1.

Furthermore, the classical approach based on sequence and structure comparative studies was used to identify the residues in the binding pocket that may regulate the product specificity of SaSQS1. However, the site-directed mutagenesis studies show that these residues affect the product yield and not the specificity. Structural studies of two important inactive mutants provide insights to the differential dynamics of these mutants with reference to the *wild* type protein that may affect the product formation.

Earlier methods to determine product modulating residues and hence product specificity of sesquiterpene synthases were primarily dependent on the sequence and limited structure information. Furthermore, the combination of residues as suggested by these techniques are colossal in number as the enzymes possess large catalytic site. That is why, perhaps, these techniques fail to provide generalized rationale for the residue dependent product specificity. In order to develop a generalised method to determine product modulating residues, sequence based and dynamics based statistical coupling analysis of SaSQS1 and other sesquiterpene synthases was performed. Putative dynamic “sectors” were identified that might contain the product defining residues of SaSQS1 and other sesquiterpene synthases. These putative sectors were further combined with the sequence and structural aspects of the enzyme including

hydrophobicity and vicinity indices, respectively, to pin down the residues that may define the product specificity of the enzymes. The approach, hence developed, correctly predicts the product modulating residues of sesquiterpene synthases for which such information has been biochemically obtained and thus validates the rationale of the methodology. Furthermore, it also provides leads on residues that could influence the product of SaSQS1 and the two mutants, T313S and G418A.

Thus, the structural and biochemical investigations conducted here on SaSQS1 provided valuable insights to its architecture and product specificity. In the present work, a novel approach combining sequence, structure and dynamical information of few sesquiterpene synthases and SaSQS1 is developed that aids in identifying the product modulating residues of these enzymes. However, there is a limitation of the biochemical data available for sesquiterpene synthases. This approach can be employed further by performing biochemical and structural studies of member of each clade of the family and extending the application of the approach for these members.

References

1. Walton, N. J., Brown, D. E. & HARBORNE, J. B. Classes and functions of secondary products from plants. *Chemicals from Plants* 1–25 (World Scientific / Imperial College Press, 1999). doi:10.1142/9789812817273_0001.

2. Bourgaud, F., Gravot, A., Milesi, S. & Gontier, E. Production of plant secondary metabolites: A historical perspective. *Plant Science* vol. 161 839–851 (2001).

3. Koes, R. E., Quattrocchio, F. & Mol, J. N. M. The flavonoid biosynthetic pathway in plants: Function and evolution. *BioEssays* **16**, 123–132 (1994).

4. Helena Caporale, L. Chemical ecology: A view from the pharmaceutical industry (molecular evolution/natural product screens/scaffold/genetic code/combinatorial chemistry*)*. vol. 92 (1995).

5. Cox-Georgian, D., Ramadoss, N., Dona, C. & Basu, C. Therapeutic and medicinal uses of terpenes. in *Medicinal Plants: From Farm to Pharmacy* 333–359 (Springer International Publishing, 2019). doi:10.1007/978-3-030-31269-5_15.

6. Liu, B., Guo, F., Chang, Y., Jiang, H. & Wang, Q. Optimization of extraction of evodiamine and rutaecarpine from fruit of Evodia rutaecarpa using modified supercritical CO2. *J. Chromatogr. A* **1217**, 7833–7839 (2010).

7. Antognoni, F. *et al.* Induction of flavonoid production by UV-B radiation in Passiflora quadrangularis callus cultures. *Fitoterapia* **78**, 345–352 (2007).

8. Zhang, L. *et al.* Metabolic profiling of chinese tobacco leaf of different geographical origins by GC-MS. *J. Agric. Food Chem.* **61**, 2597–2605 (2013).

9. Lange, B. M., Rujan, T., Martin, W. & Croteau, R. Isoprenoid biosynthesis: The evolution of two ancient and distinct pathways across genomes. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 13172–13177 (2000).

10. Jennewein, S. & Croteau, R. Taxol: Biosynthesis, molecular genetics, and biotechnological applications. *Applied Microbiology and Biotechnology* vol. 57 13–19 (2001).

11. Recent Developments on Secondary Metabolite Biosynthesis in Artemisia annua L.

https://www.researchgate.net/publication/257262950_Recent_Developments_on_Secondary_Metabolite_Biosynthesis_in_Artemisia_annua_L.

12. Christianson, D. W. Structural and Chemical Biology of Terpenoid Cyclases. (2017) doi:10.1021/acs.chemrev.7b00287.

13. Zhang, F., Rodriguez, S. & Keasling, J. D. Metabolic engineering of microbial pathways for advanced biofuels production. *Current Opinion in Biotechnology* vol. 22 775–783 (2011).

14. Sharkey, T. D. & Yeh, S. Isoprene emission from plants. *Annu. Rev. Plant Biol.* **52**, 407–436 (2001).

15. Silver, G. M. & Fall, R. Enzymatic synthesis of isoprene from dimethylallyl diphosphate in aspen leaf extracts. *Plant Physiol.* **97**, 1588–1591 (1991).

16. Sharkey, T. D., Wiberley, A. E. & Donohue, A. R. Isoprene emission from plants: Why and how. *Annals of Botany* vol. 101 5–18 (2008).

17. Singh, P. *et al.* Chemical profile, antifungal, antiaflatoxigenic and antioxidant activity of Citrus maxima Burm. and Citrus sinensis (L.) Osbeck essential oils and their cyclic monoterpene, DL-limonene. *Food Chem. Toxicol.* **48**, 1734–1740 (2010).

18. Tippmann, S., Chen, Y., Siewers, V. & Nielsen, J. From flavors and pharmaceuticals to advanced biofuels: Production of isoprenoids in Saccharomyces cerevisiae. *Biotechnology Journal* vol. 8 1435–1444 (2013).

19. Srivastava, P. L. *et al.* Functional characterization of novel sesquiterpene synthases from Indian Sandalwood, Santalum album. *Sci. Rep.* **5**, 1–12 (2015).

20. Bommareddy, A., Rule, B., Vanwert, A. L., Santha, S. & Dwivedi, C. α-Santalol, a derivative of sandalwood oil, induces apoptosis in human prostate cancer cells by causing caspase-3 activation. *Phytomedicine* **19**, 804–811 (2012).

21. Klayman, D. L. Qinghaosu (artemisinin): An antimalarial drug from China. *Science (80-.).* **228**, 1049–1055 (1985).

22. Flasiński, M., Ha¸c-Wydro, K. & Broniatowski, M. Incorporation of Pentacyclic Triterpenes into Mitochondrial Membrane - Studies on the Interactions in Model 2D Lipid Systems. *J. Phys. Chem. B* **118**, 12927–12937 (2014).

23. Yamamoto, R. *et al.* Brassinosteroid levels increase drastically prior to morphogenesis of tracheary elements. *Plant Physiol.* **125**, 556–563 (2001).

24. Lichtenthaler, H. K., Rohmer, M. & Schwender, J. Two independent biochemical pathways for isopentenyl diphosphate and isoprenoid biosynthesis in higher plants. *Physiol. Plant.* **101**, 643–652 (1997).

25. Demmig-Adams, B., Gilmore, A. M. & Iii, W. W. A. In vivo functions of carotenoids in higher plants. *FASEB J.* **10**, 403–412 (1996).

26. Rao, A. V. & Rao, L. G. Carotenoids and human health. *Pharmacological Research* vol. 55 207–216 (2007).

27. Men, X., Wang, F., Chen, G. Q., Zhang, H. B. & Xian, M. Biosynthesis of natural rubber: Current state and perspectives. *International Journal of Molecular Sciences* vol. 20 (2019).

28. Rynkiewicz, M. J., Cane, D. E. & Christianson, D. W. Structure of trichodiene synthase from Fusarium sporotrichioides provides mechanistic inferences on the terpene cyclization cascade. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 13543–13548 (2001).

29. Whittington, D. A. *et al.* Bornyl diphosphate synthase: Structure and strategy for carbocation manipulation by a terpenoid cyclase. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 15375–15380 (2002).

30. Dickschat, J. S. Bacterial Diterpene Biosynthesis. *Angew. Chemie Int. Ed.* **58**, 15964–15976 (2019).

31. Aaron, J. A. & Christianson, D. W. Trinuclear metal clusters in catalysis by terpenoid synthases. *Pure Appl. Chem.* **82**, 1585–1597 (2010).

32. Prisic, S. & Peters, R. J. Synergistic substrate inhibition of ent-copalyl diphosphate synthase: A potential feed-forward inhibition mechanism limiting gibberellin metabolism. *Plant Physiol.* **144**, 445–454 (2007).

33. Gutta, P. & Tantillo, D. J. Theoretical studies on farnesyl cation cyclization: Pathways to pentalenene. *J. Am. Chem. Soc.* **128**, 6172–6179 (2006).

34. Hong, Y. J. & Tantillo, D. J. A potential energy surface bifurcation in terpene biosynthesis. *Nat. Chem.* **1**, 384–389 (2009).

35. Tantillo, D. J. The carbocation continuum in terpene biosynthesis—where are the secondary cations? *Chem. Soc. Rev.* **39**, 2847–2854 (2010).

36. Hare, S. R. & Tantillo, D. J. Dynamic behavior of rearranging carbocations - Implications for terpene biosynthesis. *Beilstein Journal of Organic Chemistry* vol. 12 377–390 (2016).

37. Shishova, E. Y., Di Costanzo, L., Cane, D. E. & Christianson, D. W. X-ray crystal structure of aristolochene syrithase from Aspergillus terreus and evolution of templates for the cyclization of farnesyl biphosphate. *Biochemistry* **46**, 1941–1951 (2007).

38. Baer, P. *et al.* Induced-fit mechanism in class i terpene cyclases. *Angew. Chemie - Int. Ed.* **53**, 7652–7656 (2014).

39. Chen, M. *et al.* Probing the Role of Active Site Water in the Sesquiterpene Cyclization Reaction Catalyzed by Aristolochene Synthase. *Biochemistry* **55**, 2864–2874 (2016).

40. Yamada, Y. *et al.* Terpene synthases are widely distributed in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 857–862 (2015).

41. Hindra *et al.* Strain prioritization for natural product discovery by a high-throughput real-time PCR Method. *J. Nat. Prod.* **77**, 2296–2303 (2014).

42. Tarshis, L. C., Sacchettini, J. C., Yan, M. & Poulter, C. D. Crystal Structure of Recombinant Farnesyl Diphosphate Synthase at 2.6-Å Resolution. *Biochemistry* **33**, 10871–10877 (1994).

43. Mann, F. M. *et al.* Characterization and inhibition of a class II diterpene cyclase from Mycobacterium tuberculosis. Implications for tuberculosis. *J. Biol. Chem.* **284**, 23574–23579 (2009).

44. Wendt, K. U., Poralla, K. & Schulz, G. E. Structure and function of a squalene cyclase. *Science (80-.).* **277**, 1811–1815 (1997).

45. Feil, C., Süssmuth, R., Jung, G. & Poralla, K. Site-directed mutagenesis of putative active-site residues in squalene-hopene cyclase. *Eur. J. Biochem.* **242**, 51–55 (1996).

46. Rudolf, J. D. & Chang, C. Y. Terpene synthases in disguise: Enzymology, structure, and opportunities of non-canonical terpene synthases. *Natural Product Reports* vol. 37 425–463 (2020).

47. KÖksal, 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.* **7**, 431–433 (2011).

48. Köksal, M., Jin, Y., Coates, R. M., Croteau, R. & Christianson, D. W. Taxadiene synthase structure and evolution of modular architecture in terpene biosynthesis. *Nature* **469**, 116–122 (2011).

49. Li, J. X. *et al.* Rational engineering of plasticity residues of sesquiterpene synthases from Artemisia annua: Product specificity and catalytic efficiency. *Biochem. J.* **451**, 417–426 (2013).

50. Abdallah, I. I., Czepnik, M., Van Merkerk, R. & Quax, W. J. Insights into the three-dimensional structure of amorpha-4,11-diene synthase and probing of plasticity residues. *J. Nat. Prod.* **79**, 2455–2463 (2016).

51. Yoshikuni, Y., Ferrin, T. E. & Keasling, J. D. Designed divergent evolution of enzyme function. *Nature* **440**, 1078–1082 (2006).

52. Li, R. *et al.* Reprogramming the chemodiversity of terpenoid cyclization by remolding the active site contour of epi-isozizaene synthase. *Biochemistry* **53**, 1155–1168 (2014).

53. Blank, P. N. *et al.* Substitution of Aromatic Residues with Polar Residues in the Active Site Pocket of epi-Isozizaene Synthase Leads to the Generation of New Cyclic Sesquiterpenes. *Biochemistry* **56**, 5798–5811 (2017).

54. Vedula, L. S., Jiang, J., Zakharian, T., Cane, D. E. & Christianson, D. W. Structural and mechanistic analysis of trichodiene synthase using site-directed mutagenesis: Probing the catalytic function of tyrosine-295 and the asparagine-225/serine-229/glutamate-233-Mg2 +B motif. *Arch. Biochem. Biophys.* **469**, 184–194 (2008).

55. Greenhagen, B. T., O'Maille, P. E., Noel, J. P. & Chappell, J. Identifying and manipulating structural determinates linking catalytic specificities in terpene synthases. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 9826–9831 (2006).

56. Starks, C. M., Back, K., Chappell, J. & Noel, J. P. Structural basis for cyclic terpene biosynthesis by tobacco 5-epi- aristolochene synthase. *Science (80-.).* **277**, 1815–1820 (1997).

57. Blank, P. N. *et al.* Substitution of Aromatic Residues with Polar Residues in the Active

Site Pocket of epi-Isozizaene Synthase Leads to the Generation of New Cyclic Sesquiterpenes. *Biochemistry* **56**, 5798–5811 (2017).

58. Gennadios, H. A. *et al.* Crystal structure of (+)-δ-cadinene synthase from Gossypium arboreum and evolutionary divergence of metal binding motifs for catalysis. *Biochemistry* **48**, 6175–6183 (2009).

59. Salmon, M. *et al.* ARTICLE Emergence of terpene cyclization in Artemisia annua. (2015) doi:10.1038/ncomms7143.

60. Der, J. P. & Nickrent, D. L. A molecular phylogeny of Santalaceae (Santalales). *Syst. Bot.* **33**, 107–116 (2008).

61. Howes, M.-J. R., Simmonds, M. S. J. & Kite, G. C. Evaluation of the quality of sandalwood essential oils by gas chromatography-mass spectrometry. *J. Chromatogr. A* **1028**, 307–312 (2004).

62. Kuriakose, S., Thankappan, X., Joe, H. & Venkataraman, V. Detection and quantification of adulteration in sandalwood oil through near infrared spectroscopy †. doi:10.1039/c0an00261e.

63. (15) (PDF) Anticancer Effects of Sandalwood (Santalum album). https://www.researchgate.net/publication/277405939_Anticancer_Effects_of_Sandalwood_Santalum_album.

64. Zhang, X. *et al.* Identification and functional characterization of three new terpene synthase genes involved in chemical defense and abiotic stresses in Santalum album. *BMC Plant Biol.* **19**, 115 (2019).

65. Kibbe, W. A. OligoCalc: An online oligonucleotide properties calculator. *Nucleic Acids Res.* **35**, (2007).

66. Mullis, K. B. The unusual origin of the polymerase chain reaction. *Sci. Am.* **262**, 56–65 (1990).

67. Bitinaite, J. *et al.* USER™ friendly DNA engineering and cloning method by uracil excision. *Nucleic Acids Res.* **35**, 1992–2002 (2007).

68. Bitinaite, J. & Nichols, N. M. DNA Cloning and Engineering by Uracil Excision. *Curr. Protoc. Mol. Biol.* **86**, 3.21.1-3.21.16 (2009).

69. Sambrook, J. & Russell, D. W. The Inoue Method for Preparation and Transformation of Competent E. Coli : "Ultra-Competent" Cells . *Cold Spring Harb. Protoc.* **2006**, pdb.prot3944 (2006).

70. Wilfinger, W. W., Mackey, K. & Chomczynski, P. Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity. *Biotechniques* **22**, 474–481 (1997).

71. *Introduction of Electrophoresis Process*.

72. Newman, J., Xu, J. & Willis, M. C. Initial evaluations of the reproducibility of vapor-diffusion crystallization. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **63**, 826–832 (2007).

73. Matthews, B. W. Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491–497 (1968).

74. Chayen, N. E. Comparative studies of protein crystallization by vapour-diffusion and microbatch techniques. *Acta Crystallographica Section D: Biological Crystallography* vol. 54 8–15 (1998).

75. Chayen, N. E. Methods for separating nucleation and growth in protein crystallisation. *Progress in Biophysics and Molecular Biology* vol. 88 329–337 (2005).

76. Lu, Q. Q. *et al.* Replacing a reservoir solution with desiccant in vapor diffusion protein crystallization screening. *J. Appl. Crystallogr.* **43**, 1021–1026 (2010).

77. Carter, C. W. Efficient factorial designs and the analysis of macromolecular crystal growth conditions. *Methods* **1**, 12–24 (1990).

78. Cudney, R., Patel, S., Weisgraber, K., Newhouse, Y. & McPherson, A. Screening and optimization strategies for macromolecular crystal growth. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **50**, 414–423 (1994).

79. Prater, B. D., Tuller, S. C. & Wilson, L. J. Simplex optimization of protein crystallization conditions. *J. Cryst. Growth* **196**, 674–684 (1999).

80. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science (80-.).* **336**, 1030–1033 (2012).

81. Kabsch, W. *XDS*. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 125–132 (2010).

82. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution?

1204 doi:10.1107/S0907444913000061.

83. Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination. *FEBS J.* **280**, 5705–5736 (2013).

84. Navaza, J. *AMoRe* : an automated package for molecular replacement. *Acta Crystallogr. Sect. A Found. Crystallogr.* **50**, 157–163 (1994).

85. Vagin, A. & Teplyakov, A. MOLREP: An Automated Program for Molecular Replacement. *J. Appl. Crystallogr.* **30**, 1022–1025 (1997).

86. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

87. Collaborative Computational Project, N. 4 & IUCr. The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **50**, 760–763 (1994).

88. Bulk Solvent Correction: Practical Application and Effects in Reciprocal and Real Space. http://legacy.ccp4.ac.uk/newsletters/newsletter34/bsdk_text.html.

89. Terwilliger, T. C. Rapid automatic NCS identification using heavy-atom substructures. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **58**, 2213–2215 (2002).

90. Weiss, M. Biomolecular Crystallography: Principles, Practice, and Applications to Structural Biology . By Bernhard Rupp. New York: Garland Science, Taylor and Francis Group, 2010. Pp. xxi + 809. Price (hardback) USD 145.00. ISBN 978-0-8153-4081-2. . *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 640–641 (2010).

91. Emsley, P. & Cowtan, K. Biological Crystallography Coot: model-building tools for molecular graphics. doi:10.1107/S0907444904019158.

92. Brünger, A. T. Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–475 (1992).

93. Tickle, I. J., Laskowski, R. A. & Moss, D. S. Rfree and the Rfree ratio. I. Derivation of expected values of cross-validation residuals used in macromolecular least-squares refinement. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **54**, 547–557 (1998).

94. Kleywegt, G. J. & Brünger, A. T. Checking your imagination: Applications of the free

R value. *Structure* **4**, 897–904 (1996).

95. Brünger, A. T. Assessment of phase accuracy by cross validation: the free R value. Methods and applications. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **49**, 24–36 (1993).

96. Shabalin, I. G., Porebski, P. J. & Minor, W. Refining the macromolecular model–achieving the best agreement with the data from X-ray diffraction experiment. *Crystallography Reviews* vol. 24 236–262 (2018).

97. Merritt, E. A. Expanding the model: Anisotropic displacement parameters in protein structure refinement. *Acta Crystallographica Section D: Biological Crystallography* vol. 55 1109–1117 (1999).

98. Merritt, E. A. To B or not to B: A question of resolution? *Acta Crystallogr. Sect. D Biol. Crystallogr.* **68**, 468–477 (2012).

99. Read, R. J. Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr. Sect. A* **42**, 140–149 (1986).

100. Urzhumtsev, A., Afonine, P. V., Lunin, V. Y., Terwilliger, T. C. & Adams, P. D. Metrics for comparison of crystallographic maps. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70**, 2593–2606 (2014).

101. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica Section D: Biological Crystallography* vol. 53 240–255 (1997).

102. Kantardjieff, K. A. & Rupp, B. Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Sci.* **12**, 1865–1871 (2003).

103. Fokine, A. & Urzhumtsev, A. Flat bulk-solvent model: Obtaining optimal parameters. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **58**, 1387–1392 (2002).

104. Kleywegt, G. J. Validation of protein crystal structures. *Acta Crystallographica Section D: Biological Crystallography* vol. 56 249–265 (2000).

105. Chen, V. B. *et al.* MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 12–21 (2010).

106. Abagyan, R. & Totrov, M. High-throughput docking for lead generation. *Current Opinion in Chemical Biology* vol. 5 375–382 (2001).

107. Friesner, R. A. *et al.* Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. (2006) doi:10.1021/jm051256o.

108. Huang, S. Y. & Zou, X. Advances and challenges in Protein-ligand docking. *International Journal of Molecular Sciences* vol. 11 3016–3034 (2010).

109. Salmaso, V. Exploring protein flexibility during docking to investigate ligand-target recognition. (2018).

110. Bowers, K. J. *et al.* Scalable algorithms for molecular dynamics simulations on commodity clusters. in *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, SC'06* 84 (ACM Press, 2006). doi:10.1145/1188455.1188544.

111. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. | Profiles RNS. https://profiles.uchicago.edu/profiles/display/6274832.

112. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* vol. 181 223–230 (1973).

113. Süel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.* **10**, 59–69 (2003).

114. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 2021 (2011).

115. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science (80-.).* **286**, 295–299 (1999).

116. Lichtarge, O., Bourne, H. R. & Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358 (1996).

117. Smock, R. G. *et al.* An interdomain sector mediating allostery in Hsp70 molecular chaperones. *Mol. Syst. Biol.* **6**, (2010).

118. Reynolds, K. A., McLaughlin, R. N. & Ranganathan, R. Hot spots for allosteric

regulation on protein surfaces. *Cell* **147**, 1564–1575 (2011).

119. Ota, N. & Agard, D. A. Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. *J. Mol. Biol.* **351**, 345–354 (2005).

120. Ferguson, A. D. *et al.* Signal transduction pathway of TonB-dependent transporters. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 513–518 (2007).

121. McLaughlin, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491**, 138–142 (2012).

122. Starks, C. M. Structural Basis for Cyclic Terpene Biosynthesis by Tobacco 5-Epi-Aristolochene Synthase. *Science (80-.).* **277**, 1815–1820 (1997).

123. Whittington, D. A. *et al.* Bornyl diphosphate synthase: Structure and strategy for carbocation manipulation by a terpenoid cyclase. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 15375–15380 (2002).

124. Mezulis, S., Yates, C. M., Wass, M. N., E Sternberg, M. J. & Kelley, L. A. The Phyre2 web portal for protein modeling, prediction and analysis. (2015) doi:10.1038/nprot.2015.053.

125. Mccoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Cryst* **40**, 658–674 (2007).

126. Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Cryst* **75**, 861–877 (2019).

127. 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* **14**, 1011–1019 (2019).

128. Starks, C. M., Back, K., Chappell, J. & Noel, J. P. Structural basis for cyclic terpene biosynthesis by tobacco 5-epi- aristolochene synthase. *Science (80-.).* **277**, 1815–1820 (1997).

129. McAndrew, R. P. *et al.* Structure of a three-domain sesquiterpene synthase: A prospective target for advanced biofuels production. *Structure* **19**, 1876–1884 (2011).

130. Baer, P. *et al.* Induced-Fit Mechanism in Class I Terpene Cyclases. *Angew. Chemie Int. Ed.* **53**, 7652–7656 (2014).

131. Li, R. *et al.* Reprogramming the chemodiversity of terpenoid cyclization by remolding the active site contour of epi-isozizaene synthase. *Biochemistry* **53**, 1155–1168 (2014).

132. Köksal, 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.* **402**, 363–373 (2010).

133. Li, J. X. *et al.* Rational engineering of plasticity residues of sesquiterpene synthases from Artemisia annua: Product specificity and catalytic efficiency. *Biochem. J.* **451**, 417–426 (2013).

134. Peters, R. J. & Croteau, R. B. *Abietadiene synthase catalysis: Mutational analysis of a prenyl diphosphate ionization-initiated cyclization and rearrangement*. www.pnas.orgcgidoi10.1073pnas.022627099 (2001).

135. Greenhagen, B. T., O'maille, P. E., Noel, J. P. & Chappell, J. *Identifying and manipulating structural determinates linking catalytic specificities in terpene synthases*. vol. 103 www.pnas.orgcgidoi10.1073pnas.0601605103 (2006).

136. O'Maille, P. E. *et al.* Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nat. Chem. Biol.* **4**, 617–623 (2008).

137. Koo, H. J. *et al.* Biosynthetic potential of sesquiterpene synthases: Product profiles of Egyptian Henbane premnaspirodiene synthase and related mutants. *J. Antibiot. (Tokyo).* **69**, 524–533 (2016).

138. Kersten, R. D., Diedrich, J. K., Yates, J. R. & Noel, J. P. Mechanism-Based Post-Translational Modification and Inactivation in Terpene Synthases. *ACS Chem. Biol.* **10**, 2501–2511 (2015).

139. Abdallah, I. I., Czepnik, M., Van Merkerk, R. & Quax, W. J. Insights into the Three-Dimensional Structure of Amorpha-4,11-diene Synthase and Probing of Plasticity Residues. *J. Nat. Prod* **79**, 42 (2016).

140. Seemann, M. *et al.* Pentalenene Synthase. Analysis of Active Site Residues by Site-Directed Mutagenesis. (2002) doi:10.1021/ja026058q.

141. Srivastava, P. L. *et al.* Functional Characterization of Novel Sesquiterpene Synthases from Indian Sandalwood, Santalum album OPEN. (2015) doi:10.1038/srep10095.

142. Noel, J. P. *et al.* Structural elucidation of cisoid and transoid cyclization pathways of a sesquiterpene synthase using 2-fluorofarnesyl diphosphates. *ACS Chem. Biol.* **5**, 377–392 (2010).

143. Vedula, L. S., Cane, D. E. & Christianson, D. W. Role of arginine-304 in the diphosphate-triggered active site closure mechanism of trichodiene synthase. *Biochemistry* **44**, 12719–12727 (2005).

144. Matos, J. O. *et al.* Mechanism Underlying Anti-Markovnikov Addition in the Reaction of Pentalenene Synthase. *Biochemistry* **59**, 3271–3283 (2020).

145. McAndrew, R. P. *et al.* Structure of a three-domain sesquiterpene synthase: A prospective target for advanced biofuels production. *Structure* **19**, 1876–1884 (2011).

146. Aaron, J. A., Lin, X., Cane, D. E. & Christianson, D. W. Structure of epi-isozizaene synthase from streptomyces coelicolor A3(2), a platform for new terpenoid cyclization templates. *Biochemistry* **49**, 1787–1797 (2010).

147. Durairaj, J. *et al.* An analysis of characterized plant sesquiterpene synthases. *Phytochemistry* **158**, 157–165 (2019).

148. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).

149. Rivoire, O., Reynolds, K. A. & Ranganathan, R. Evolution-Based Functional Decomposition of Proteins. *PLoS Comput. Biol.* **12**, 1004817 (2016).

150. Fiser, A. & Šali, A. MODELLER: Generation and Refinement of Homology-Based Protein Structure Models. *Methods Enzymol.* **374**, 461–491 (2003).

151. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

152. Hü, P. H., Mark, A. E. & Van Gunsteren, W. F. *Fluctuation and Cross-correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations*. *J. Mol. Biol* vol. 252 (1995).

153. Grant, B. J., Rodrigues, A. P. C., Elsawy, K. M., Mccammon, J. A. & Caves, L. S. D. Bio3d: an R package for the comparative analysis of protein structures. **22**, 2695–2696

(2006).

154. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).

155. Lakhani, B., Thayer, K. M., Black, E. & Beveridge, D. L. Journal of Biomolecular Structure and Dynamics Spectral analysis of molecular dynamics simulations on PDZ: MD sectors Spectral analysis of molecular dynamics simulations on PDZ: MD sectors. (2019) doi:10.1080/07391102.2019.1588169.

156. Baer, P. *et al.* Hedycaryol Synthase in Complex with Nerolidol Reveals Terpene Cyclase Mechanism. *ChemBioChem* **15**, 213–216 (2014).

157. Li, R. *et al.* Reprogramming the Chemodiversity of Terpenoid Cyclization by Remolding the Active Site Contour of epi-Isozizaene Synthase. (2014) doi:10.1021/bi401643u.

158. Chou, W. K. W. *et al.* Substitution of Aromatic Residues with Polar Residues in the Active Site Pocket of epi-Isozizaene Synthase Leads to the Generation of New Cyclic Sesquiterpenes. *Biochemistry* **53**, (2014).

Appendix A

Structural Studies of members of DOCK family of proteins and their interacting partners

A.1 Background:

Abnormal cell invasion is a feature of cancer cells leading to metastasis. The tumour cells metastasize through the tissues. This process requires cell mobility, interaction with the extracellular matrix and remodelling of cell-cell contacts[1]. Cell migration and adhesion are regulated by a number of physical and chemical factors like growth factors, tension, cytokines and shear stress. Membrane receptors, like growth factor and cytokine receptors, transmit these signals through the intracellular membrane. Furthermore, these receptors cascade the signal to activate Rho GTPases to attain the required actin cytoskeleton reorganization for cell migration and or adhesion.

Over 150 small GTPases comprise the RAS superfamily. These GTPases are approximately 20 kDa monomeric G-proteins and based on functional and structural conservation divided into five sub families, RHO, RAS, RAB, RAN and ARF[2,3]. These GTPases share the conserved biochemical function and act as biomolecular switches regulating a number of signaling cascades like protein and vesicle transport, gene expression and remodelling of cytoskeleton[4,5]. GTPases cycle between an active, guanosine-5'-triphosphate (GTP) bound state, and an inactive, guanosine-5'-diphosphate (GDP) bound state and thus are known as binary molecular switches[6] (Fig. A.1). Amongst these, members of RHO sub-family are known to be the key regulators of cytoskeleton remodelling and hence they control cell migration. There are three prototypical members of this family, Rho, Rac and Cdc42, that regulate the cell movement. Rac1 promotes lamellipodia formation for cell motility, whereas, RhoA promotes the formation of actin stress fibres and generates actomyosin contractile force for cell migration[7,8]. Under the physiological conditions, they regulate the cell movement by modulating actin assembly, microtubules and actomyosin contractility[9]. In many cancers, the expression of Rho family members is deregulated and is associated with progression of disease[10]. Due to the low

intrinsic GTPase activity of these G proteins, regulatory proteins are a requisite for spatiotemporal governance of the transition between inactive and active states. These regulatory proteins include GTPase activating proteins (GAPs) that trigger GTP hydrolysis resulting in an inactive GTPase, guanine nucleotide exchange factors (GEFs) that facilitates GDP dissociation to promote active (ON) state of the GTPases and the GDP dissociation inhibitors (GDIs) that sequester the inactive GTPase state of RHO and RAB subfamilies in the cytoplasm[6]. The regulation of GTPases by these regulatory proteins varies amongst cell types and hence forms intricate regulatory signalling networks [11].

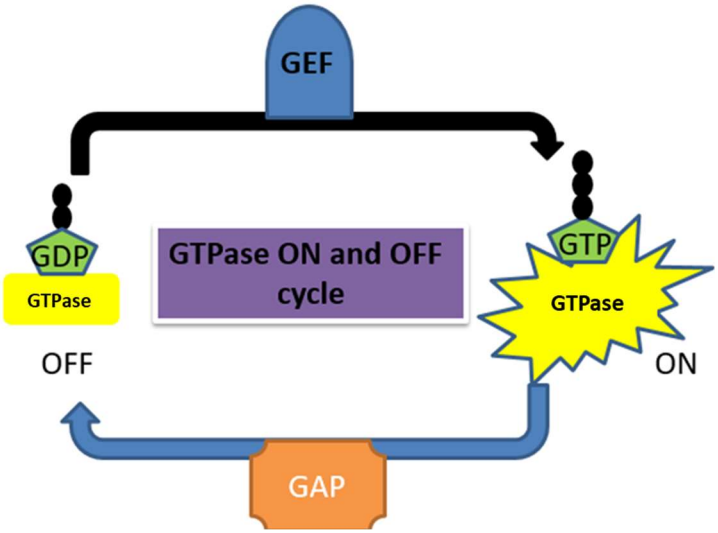


Fig. A.1. *'On' and 'Off' state of GTPases regulated by GEF (guanine nucleotide exchange factor) and GAP (GTPase activating protein), respectively by the exchange of GDP and GTP.*

Recently, amongst the three regulators, GEFs have grabbed lot of attention due to their potential druggability. There are two class of GEFs that regulate the activation of Rho GTPases, the Dbl homology (DH) domain containing or the typical (canonical) GEFs and the dedicator of cytokinesis (DOCK) or the atypical GEFs (Fig. A.2). The typical GEFs have a pleckstrin homology (PH) domain which along with DH domain confers GEF activity. On the contrary, the DOCK family GEF possess two DOCK homology (DHR-1 and DHR-2) domains. Interestingly, in human beings, out of 22 known Rho GTPases, DOCK GEFs regulate just two members, Rac1 and Cdc42.

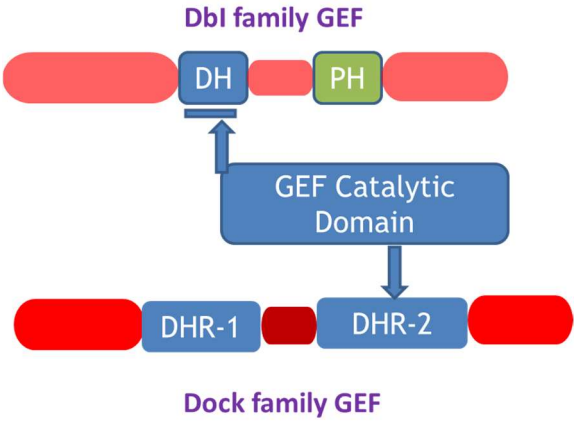


Fig. A.2. *Classes of GEF including the typical Dbl family and the atypical DOCK family.*

Depending on GTPase specificity and functional domains, the DOCK family is divided into four groups that include, DOCK-A, DOCK-B, DOCK-C and DOCK-D (Fig. A.3). The DOCK-A, including (DOCK1, DOCK2 and DOCK5), and DOCK-B, including (DOCK3 and DOCK4), subfamilies are specific for Rac1. These subfamilies possess an N-terminal Src homology 3 (SH3) domain, DOCK homology regions (DHR-1 and DHR-2) and a proline rich motif at C-terminus. It was proposed that the SH3 domain and the proline rich motif bind each other leading to an autoinhibited state of the protein. This state is relieved by interaction with other proteins like the adaptor engulfment and cell motility protein (ELMO1)[12,13]. The DHR-1 domain helps in localization of the protein at the plasma membrane which results in initiation of membrane protrusion and thus cell migration[14]. Whereas the DHR-2 domain is the catalytic domain, carrying out the GEF activity. Specific residues in the DHR-2 domain are suggested to be important for substrate binding and GEF activity[15,16].

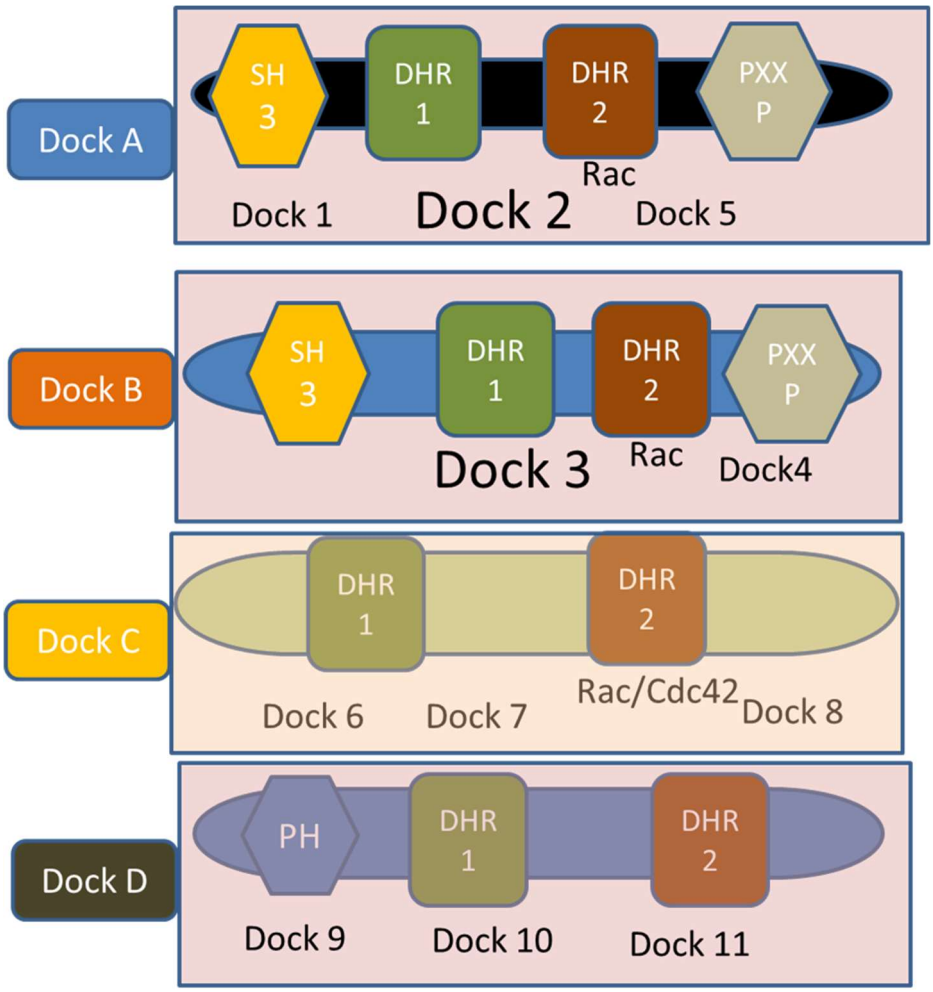


Fig. A.3. *Different classes of DOCK family that differ in their GTPase specificity and domain organization.*

Structural studies of members of DOCK family, especially of their DHR-2 domains, have helped in understanding the mechanism of nucleotide dissociation by the DHR-2 domain. This eventually aids in developing strategies to design strong interacting small molecule inhibitors that act as chemical probes to investigate the cellular function and provide initial phase for drug discovery. However, the role of other proteins in the activation and signalling of DOCK GEFs has still remained elusive. In the present work, we try to study the interaction of two important members of family, DOCK2 and DOCK3, with their partners ELMO1 and NEDD9 (Neural Precursor Developmentally Downregulated 9), respectively, with an objective to unravel the role of DOCK2 and DOCK3 in cell migration. To achieve this objective, we set out to elucidate structures of these proteins in *apo* form and in complex with their interacting partners.

Members of DOCK family act as exchange factors in cell migration:

A.1.1 DOCK2: DOCK2, involved in activation of Rac GTPases, expresses predominantly in hematopoietic cells[17,18]. It regulates immune system by activating both innate and adaptive

immune cells[19,20]. DOCK2 mutations have been reported in numerous immunodeficiencies[21]. DOCK2 expresses exclusively in microglia in the brain and acts as a biomarker for this cell type. It has implications in the pathogenesis of Alzheimer as it enhances formation of amyloid beta (Aβ) plaque[22]. In human esophageal adenocarcinoma cells, mutations in DOCK2 were observed, indicating involvement of aberrant Rac1 regulation by mutant DOCK2 in esophageal adenocarcinoma tumorogenesis[23].

The N-terminal region of DOCK2, containing the SH3 domain, interacts with ELMO1 which is a mammalian homolog of CED-12 of C. elegans[24] (Fig. A.4). ELMO1 comprises of RhoG binding domain, PH domain and ELMO domain at the N-terminus and three PxxP motifs at the C-terminus (Fig. A.5). The SH3 containing region of DOCK2 interacts with the C-terminal region of ELMO1 that includes the Proline rich sequence and further activates Rac[24].

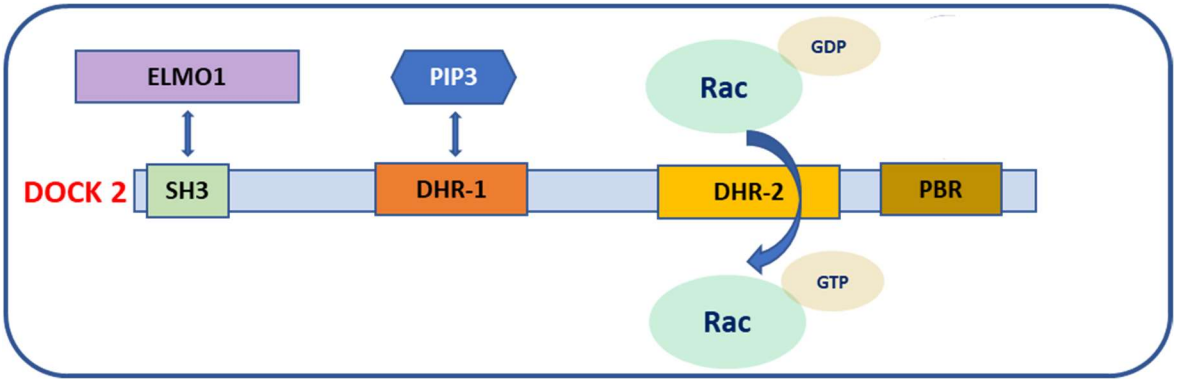


Fig. A.4. *Domain organization of DOCK2 with DHR-1 domain that interacts with phosphatidylinositol (3,4,5)-triphosphate (PIP3) that helps in membrane localization, DHR-2 domain that possess the GEF activity and a polybasic region (PBR).*

Fig. A.5. *Arrangement of different domains of ELMO1.*

The N-terminal SH3 domain of DOCK2 interacts with the DHR-2 domain that leads to autoinhibited state of the protein. Also, the interaction between the ELMO inhibitory domain (EID), which comprise of the PH domain of the protein, and the autoregulatory domain (EAD) renders ELMO to an autoinhibited state. However, the engagement of PH domain of ELMO with the SH3 domain of DOCK2 has been shown to mutually relieve their state of autoinhibition (Fig. A.6). In the present work, attempts were made to understand the molecular basis of GEF activity of DOCK2 mediated by ELMO1 and thus the signal transduction events leading to activation of DOCK2.

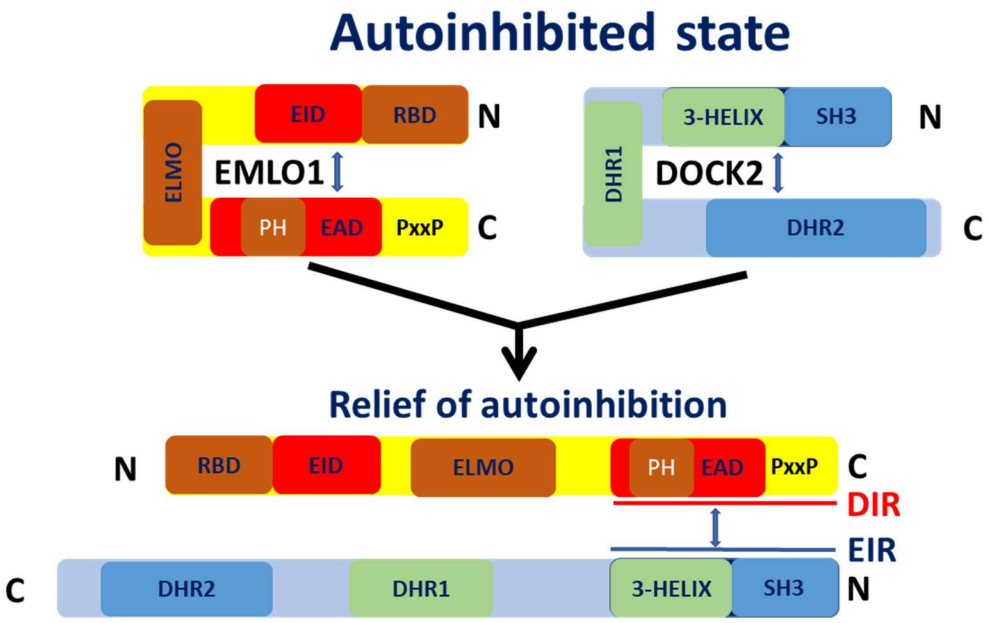


Fig. A.6. *Interaction of domains that leads to autoinhibition of ELMO1 and DOCK2 and relief of this autoinhibition states of both the proteins when they interact through their ELMO1 interacting region (EIR) and DOCK2 interacting region (DIR).*

A.1.2 DOCK3: DOCK3 was identified as an interacting partner of presenilin 1, a protein found to be mutated in various familial Alzheimer's disease cases. Thus, DOCK3 was known as presenilin binding protein (PBP)[25]. It is a cytoplasmic protein that expresses mainly in brain, retina and spinal cord[26]. DOCK3 has been observed to be associated with Alzheimer's disease and has diverse roles in neural development. It belongs to the DOCK-B family and activates specifically Rac1.

DOCK3 in complex with the adaptor protein NEDD9 activates Rac1 that leads to elongated mesenchymal movement[27]. Neural precursor expressed developmentally downregulated 9 or NEDD9 belongs to the Crk-associated substrate (Cas) family involved in signal transduction[28]. Proteins of Cas family possess a N-terminal SH3 domain, a serine rich domain, a substrate binding domain with numerous tyrosine phosphorylation site motifs and a C-terminal helix-loop-helix motif (Fig. A.7).

Fig. A.7. *Domain organization of NEDD9.*

In T and B cells, the integrin β1 signalling leads to the phosphorylation of this substrate domain of NEDD9[29,30]. Elevation in NEDD9 expression has been found to be linked to increased

migration and tumour progression. Overexpression of NEDD9 is observed in adult T-cell leukemia[31], ovarian cancer[32] and colorectal cancer[33]. The DOCK3-NEDD9-Rac signalling, leads to the actin assembly and proteolysis for extra-cellular matrix degradation[34] which is essential for mesenchymal & elongated cell movement. DOCK3 mediated signalling is shown to be dysregulated in melanoma cells, making them highly metastatic[27]. This signalling cascade is mediated by integrin αvβ3 which along with Src Kinase drives NEDD9 dependant elongation[35] (Fig. A.8). To understand the mechanism of DOCK3-NEDD9 regulated cell migration, there is a need to understand the structural basis of interaction of these two proteins, which would provide insights on the role of this interaction of the activity and localization of DOCK3.

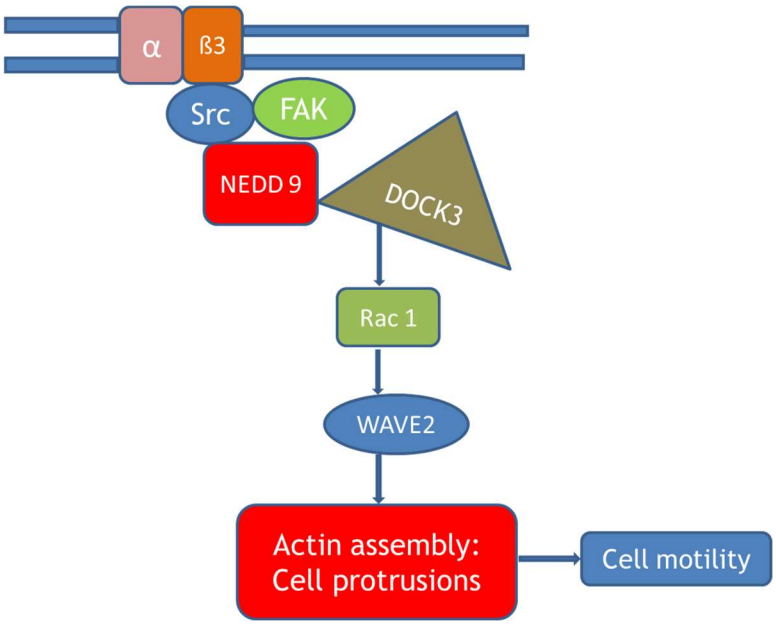


Fig. A.8. *Activation of Rac1 by interaction of DOCK3 and NEDD9. The integrin* αvβ3 *through Src kinase and FAK (focal adhesion kinase) drives NEDD9 dependent Rac1 activation.*

A.2 Methodology:

A.2.1 Cloning:

DOCK2 gene from *Homo sapiens* and ELMO1 gene from *Mus musculus* were cloned in the vector pFastBacDUAL (pFBDM) under p10 and polyhedrin promoter, respectively. The vector facilitates the heterologous gene recombination with the engineered baculovirus genome. DOCK2 has a 6XHis-SUMO tag at the N-terminus and a Strep tag at the C-terminus. ELMO1 has a N-terminal 6XHis-SUMO tag followed by a TEV site. Different full-length gene and shorter constructs of DOCK3 and NEDD9 from *Homo sapiens* were designed (Table A.1) based on the disorder prediction from PHYRE[2] [36]. These constructs cloned in pFBDM were tested for expression in *Spodoptera frugiperda* (Sf9) insect cells (Invitrogen).

A.2.2 Bacmid Preparation:

The cloned plasmid was transformed in chemically competent *E. coli* DH10MultiBac cells. These cells contain the baculovirus genome as a bacterial artificial chromosome (BAC)[37]. Blue white screening and antibiotic selection was performed to select the DH10MultiBac cells transformed with the desired plasmid. To amplify the recombinant viral DNA containing gene of interest, the white colony obtained was inoculated in 10-20 mL LB broth with the selection antibiotics (Ampicillin, Kanamycin, Gentamycin and Tetracycline). The baculoviral DNA was isolated using ethanol precipitation. The bacmid thus obtained was resuspended in 50 µL sterile ultra-pure distilled water.

A.2.3 Protein expression and Purification:

The Sf9 cells are transfected with the bacmid using commercial transfection reagent Gene Juice (Merck Millipore) in a six well plate. This forms the passage 1 (P1) stage of viral particles. The viral titres were further amplified to P2 and P3 with 50-100 mL culture. For protein expression, Sf9 cells were infected with P3 stage baculovirus and cultured in a shaker incubator for 65 hours. The cell viability was constantly monitored and the cells were harvested before the viability dropped below 80%. The Sf9 cells were cultured in Sf-900 III Serum Free Medium (Gibco) at 27^0C, 100 rpm. Viral titre and protein expression standardizations were first performed at a small scale (50-100 mL). Subsequently, the culture was scaled up to 2-3 litre for protein expression using the optimal viral titre and expression time.

The cells were harvested by centrifugation at 600 x g and the cell pellet was washed with phosphate buffer saline (PBS). For the constructs with strep tag, the cells were lysed by sonication in buffer containing 50 mM Tris-HCl (pH 8.0), 150 mM NaCl, 5% glycerol, 2 mM DTT, 3 mM ethylenediaminetetraacetic acid (EDTA) and one tablet of complete EDTA-free protease inhibitor (Roche). The lysate was centrifuged at 20,000 rpm for 30 minutes and loaded on the streptavidin resin (Qiagen) for binding. The bound protein was eluted with 10 mM desthiobiotin. The protein was further purified by gel filtration chromatography using Superose 6 10/300 GL (GE Healthcare Biosciences, USA) with buffer containing 10 mM Tris-HCl (pH 8.0), 250 mM NaCl, 0.1 mM EDTA and 2 mM DTT.

A.2.4 Crystallization trials of NEDD9:

NEDD9 full length and the shorter construct encoding 1-400 amino acids were tested for crystallization at a concentration of 4.5 mg/ml and 4.8 mg/ml, respectively. Different commercial crystallization suites were used for screening. Sitting drop vapor diffusion method was used using 100 nL of both protein and precipitating solution. The plates were stored at 20 ^{0}C and were screened for the hits.

A.2.5 Sample preparation and Data collection for Cyo EM:

Vitrification of protein was done using FEI vitrobot. C-Flat Holey carbon 300 mesh grids were glow discharged for 20 seconds at 20 mA. The grids were placed on a pair of tweezers in the vitrobot blotting chamber. The temperature of the chamber was set to 8 ^{0}C with 100% humidity to avoid evaporation of the sample. 3 µL of the protein sample with appropriate concentration was applied on the grid and the protein was allowed to absorb on the grid holes for 30 sec. The grid was blotted automatically for 7 sec, followed by plunge-freezing in liquid ethane. Frozen grids were stored liquid nitrogen in cryo grid boxes.

A.2.6 Data collection and Processing:

Grids were transferred to Talos Arctica at 200 kV and images were collected at 40,400 X (corresponding to a pixel size of 1.24 Å/pixel) with a K2 direct electron detector. The total exposure time was 8 sec with a flux of 30 $e^-/Å^2$.s. The defocus range used for automated data collection was 1.0 µm to 2.75 µm with intervals of 0.25 µm. The micrographs were analysed in RELION 3.1[38] and subjected to beam-induced motion, drift correction and dose-weighting with MotionCor235. Gctf-v.1.1836[39] was used to do contrast transfer function (CTF) fitting and phase-shift estimation without dose weighting. Micrographs were checked manually.

A.3 Results and Discussion:

A.3.1 Interaction studies of ELMO1-DOCK2-Rac1:

To study the structural basis of interaction of DOCK2 and ELMO1 and their role in activation of Rac1, attempts were made to express and purify the ternary complex. The gene for both DOCK2 and ELMO1 were cloned in pFBDM vector under different promoters and were expressed in Sf9 cells mediated by baculovirus infection. This heterodimer complex of around 600 kDa was purified using streptavidin pull down. Both the proteins were observed in the elute fraction (Fig. A.9a).

To obtain the ternary complex of this heterodimer with Rac1, the protein was purified using bacterial expression system. Rac1, cloned in the vector pOPINF with a N-terminal 6X-HIS tag, was expressed in the *E.coli* B834 rare cells. The protein was purified using NiNTA affinity chromatography with the elution buffer comprising of 50 mM Tris-HCl (pH 8.0), 500 mM Nacl, 5% glycerol, 2 mM βMe, and 2 mM EDTA. The eluted protein was further purified by gel filtration using Sephacryl S-200 16/60 with buffer containing 10 mM Tris-HCl (pH 8.0), 250 mM Nacl, 5% glycerol, 2 mM DTT and 2 mM EDTA (Fig. A.9b). The heterodimer DOCK2- ELMO1 was mixed with the purified Rac1 in a molar ratio of 1:3, respectively. The

mixture was incubated on ice for 1 hour and subjected to gel-filtration. The formation of the ternary complex, DOCK2-ELMO1-Rac1, was validated on SDS-PAGE (Fig. A.9c).

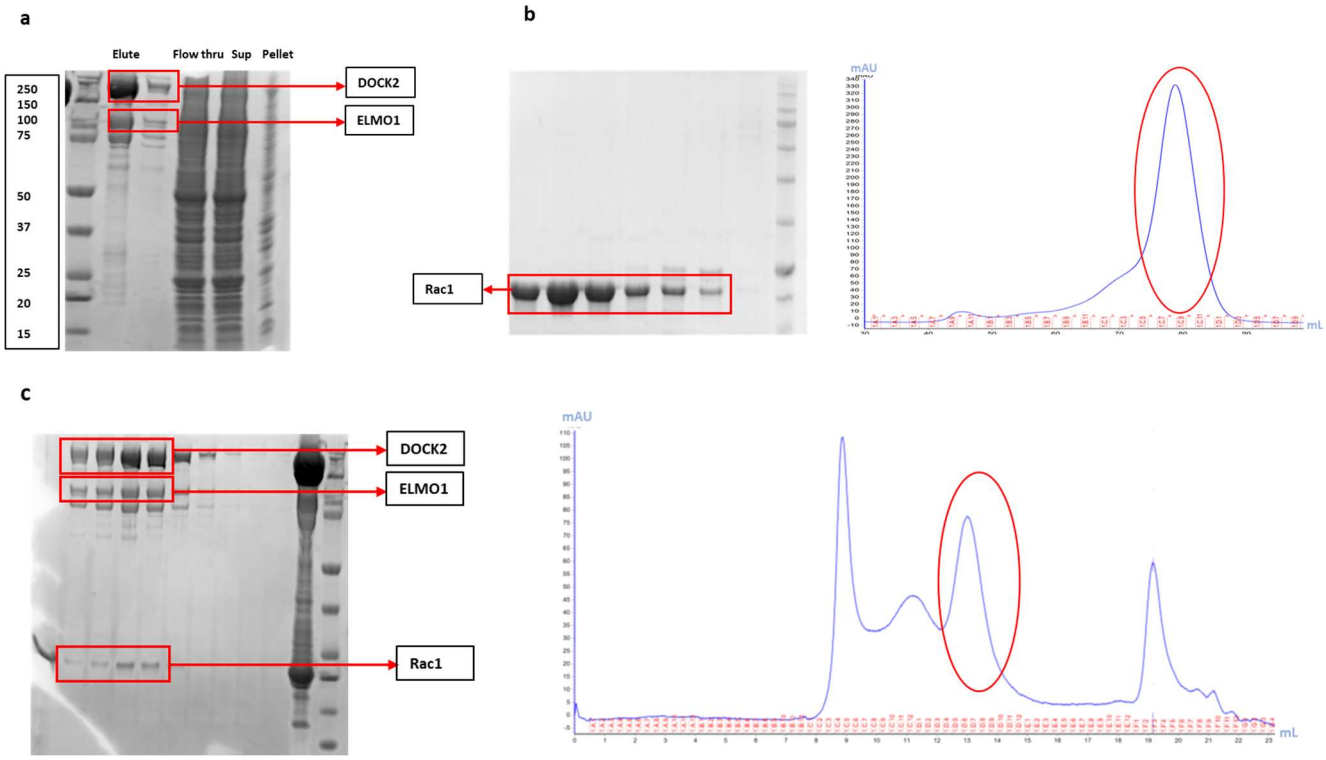


Fig. A.9. *(a) Strep pull down of ELMO1 and DOCK2. (b) Purification of Rac1 (21kDa) from bacterial system. (c) Formation of ternary complex of DOCK2-ELMO1-Rac1 confirmed by size exclusion chromatography. The proteins are observed to be in the same elution fraction and peak.*

Furthermore, cryo-electron microscopy was used to perform the structural studies of the complex. Due to its large size and the often-metastable nature, it is difficult to study the structure of the complex by X-ray crystallography and NMR. Hence, different concentrations of the protein were screened to obtain well resolved particles on the grid. Fewer particles were observed at a protein concentration of 2 mg/ml, which were not adequate for further processing (Fig.A.10). Also, crowding of particles was observed in the micrographs. Hence, further optimization is required for the concentration of the protein and grid preparation in order to obtain well resolved homogeneous particles for data processing.

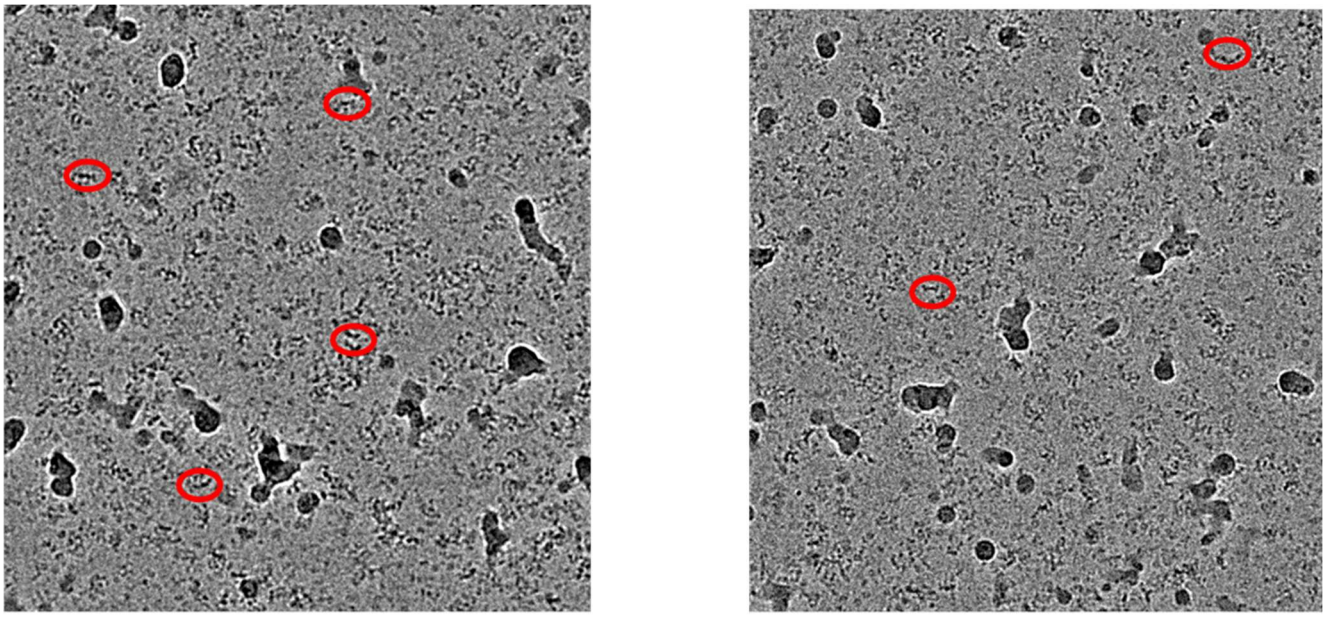


Fig. A.10. *Micrographs showing particle distribution of DOCK2-ELMO1-Rac1. The particles are highlighted with red circles.*

A.3.2 Sample preparation to study the interaction of DOCK3 and NEDD9 using Cryo EM:

To understand the mechanism of DOCK3-NEDD9 regulated cell migration, optimization trials for the expression of both the proteins were performed in Sf9 cells. Expression of both the proteins were regulated by different promoters in the vector pFBDM. However, the solubility and expression of the huge complex of approximately 600kDa was low (Fig. A.11a). Different expression and solubilization trials that include changing the viral titre for protein expression and using 0.1% Tween 20 were tried. However, these optimizations did not aid in enhancing the solubility or the expression of the complex. Thus, a green fluorescent protein (GFP) tag was introduced at the C-terminus of DOCK3, which improved the solubility of the protein (Fig. A.11b). Afterwards, the interaction of DOCK3-NEDD9 was studied by performing pull-down studies on the STREP column. DOCK3 has double strep tag at the C-terminus, therefore it was immobilized on the STREP column and purified NEDD9 was passed through the column to assess the complexation. However, substantial proportion of NEDD9 was observed to be in the unbound fraction (Fig. A.11c). Furthermore, the protein was found to be precipitating when the tag was cleaved with the TEV protease. Therefore, the attempts to study the interaction of DOCK3 and NEDD9 on strep column were met with limited success. This indicates that the interaction might require post-translational modifications of one of the proteins (perhaps of NEDD9) or both, which might not be happening in the current expression system (insect cells).

Therefore, to obtain stable DOCK3-NEDD9 complex different expression system could be tried. Co-expression of DOCK3-NEDD9 with FAK (focal adhesion kinase) might help in the formation of the stable DOCK3-NEDD9 complex as the latter is known to be involved in the phosphorylation of NEDD9. There are reports that suggest the role of FAK in DOCK3-NEDD9 signalling[35].

Fig. A.11. *(a) Expression of DOCK3 and NEDD9 was observed, however, the proteins were majorly insoluble. (b) Introduction of GFP tag in DOCK3 enhanced the fraction of protein in supernatant (sup) fraction. (c) Strep pull down of DOCK3 and NEDD9 in which NEDD9 was observed majorly in flow through.*

A.3.3 Structural studies of DOCK3 and NEDD9:

Based on the outcomes of our attempts on DOCK3-NEDD9 complexation, we planned to study the structure of DOCK3 and NEDD9, individually. To achieve this, full length and shorter constructs of both the proteins were expressed in insect cells. DOCK3 full length was observed to be less in the soluble fraction. Hence, based on the disorder prediction, a shorter construct encoding amino acids 1-1920 was cloned in pFBDM and expressed in Sf9 cells. However, no enhancement in the solubility levels of the protein was observed (Fig. A.12a). Thus, efforts to investigate the structure of full length DOCK3 were carried out by tweaking the purification conditions (Fig. A.12b). Different concentrations of salt and percentage of glycerol in buffers were tried to overcome the problem of protein precipitation during concentration and size exclusion. However, the protein was found to be precipitating.

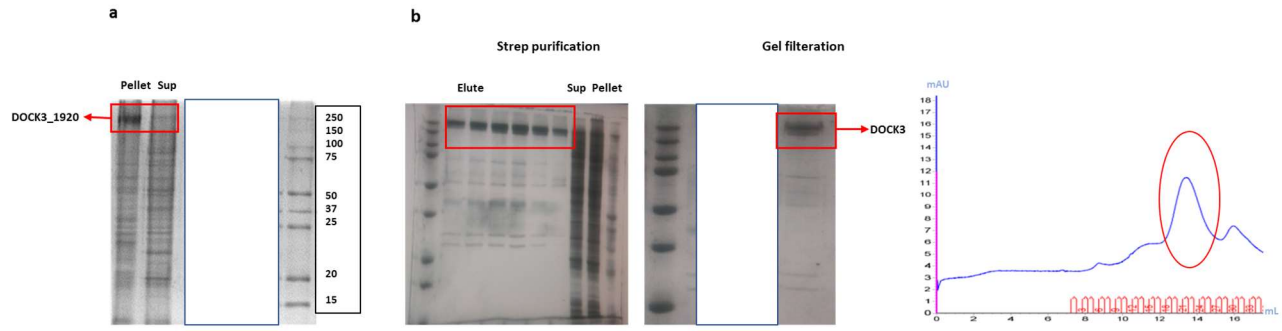


Fig. A.12. *(a) Expression trial of DOCK3_1920 with little amount of protein in supernatant (Sup) fraction.. (b) Strep purification of DOCK3 followed by size exclusion chromatography.*

The protein thus obtained was subjected to cryo electron microscopy studies. Grids were prepared with a concentration of 1 mg/ml and were tested for data collection. Similar to DOCK2-ELMO1-Rac1 complex, in this case also, fewer particles were observed on the micrograph that were insufficient for data processing to acquire 2D classes (Fig. A.13).

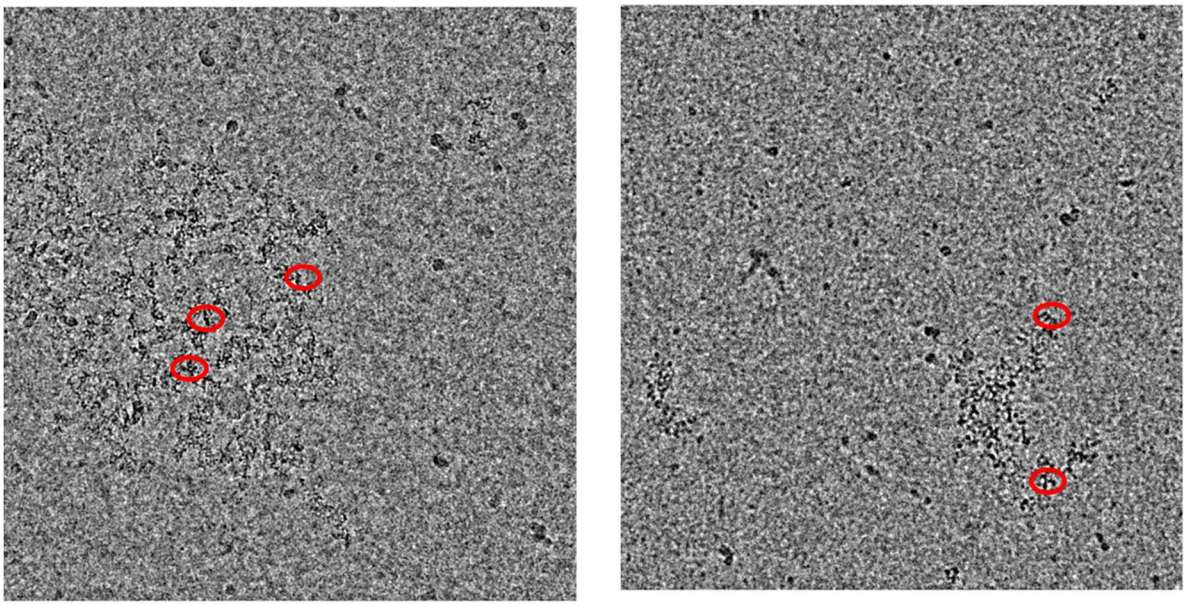


Fig. A.13. *Micrographs showing particle distribution of DOCK3. The particles observed were few in number and are highlighted with red circles.*

On the other front, NEDD9 was purified using NiNTA affinity chromatography by lysing the cells using sonication and buffer containing 50 mM Tris-HCl (pH 8.0), 300 mM Nacl, 5% glycerol, 1 mM DTT, 2 mM EDTA, 5mM imidazole, 0.05% triton X-100 and one tablet of complete EDTA-free protease inhibitor. The bound protein was eluted with 100 mM imidazole. This was followed by gel filtration using Superdex 200 increase 10/300 GL with buffer comprising of 10 mM Tris-HCl (pH 8.0), 300 mM Nacl, 5% glycerol, 2 mM DTT, 2 mM EDTA

(Fig. A.14a). NEDD9 crystallization trials were carried out for both full length and shorter construct, coding for 1-400 amino acids with the substrate binding domain of the protein. Various commercial screens were used for this purpose. No substantial crystal hits were obtained for the NEDD9 (1-400) construct. However, for the full-length construct few conditions were found to be promising, and improvisation of these yielded small crystals. However, they were not optimal for diffraction studies (Fig. A.14b).

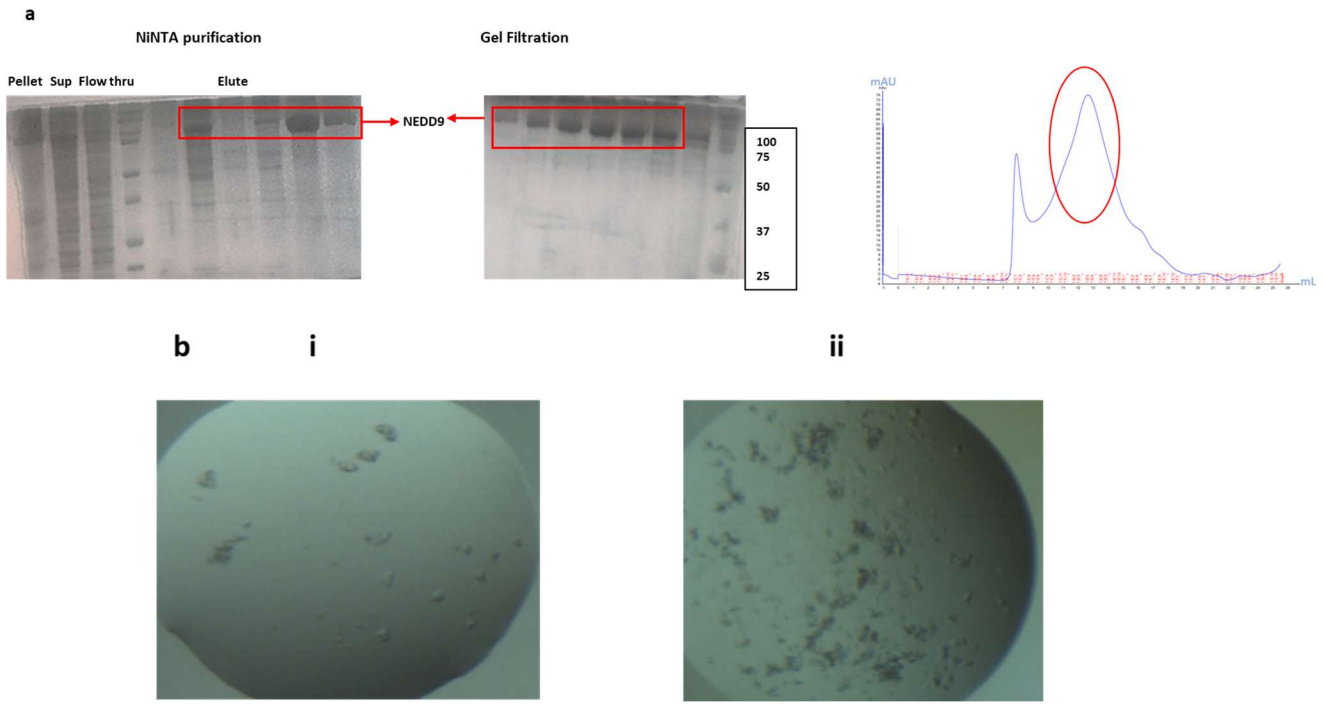


Fig. A.14. *(a) Purification of NEDD9 full length through NiNTA affinity chromatography. (b) Crystal hits obtained for NEDD9 in the condition: (i) 0.1M Mg Acetate, 0.1M Na cacodylate (pH-6.5), 15% PEG6000 (ii) 0.2M Ca Acetate, 0.1M Na cacodylate (pH-6.5), 18% PEG8000.*

A.4 Conclusion:

In the present work, we tried to understand the regulation of cell migration by the members of DOCK family. Two important members of DOCK-A family, DOCK2 and DOCK3 are known to regulate the activation of Rac1 that eventually leads to cell migration and motility. Thus, to study this regulation at the molecular level, attempts for structural studies of these proteins with their interacting partners were carried out. A stable ternary complex of DOCK2-ELMO1-Rac1 was observed. To study the structural aspect of this ternary complex, cryo electron grids were prepared. However, the particles of the protein on the grids appeared to be crowded together and those isolated as single particles were not adequate for data processing. Similar attempts were performed for DOCK3. However, a stable complex of DOCK3 with its interacting partner NEDD9 was not obtained. Thus, we tried to elucidate the structure of DOCK3. Though the

protein expression was optimal, due to precipitation the final protein yield was low. The particles of DOCK3 on the grid were observed to be clustered together. Concomitantly, we tried to express and crystallize NEDD9. The crystals obtained for the protein were small and could not be subjected to diffraction. Table A.1 lists the different constructs of DOCK3 and NEDD9 with their expression details. Our experimental results suggest that further optimization trials are required to obtain high yield of DOCK3 and subsequently good particle distribution on the grid to elucidate the structure of the proteins and their complexes. These structures would provide a wealth of information on the mechanism of action of DOCK2 and DOCK3 at molecular level in exerting cell migration. Meanwhile, recently, structures of Dock2-Elmo1 and Dock2-Elmo1-Rac1 complexes were reported[40].

Construct	Vector	Expression Tag	Expression
DOCK3_FL	pFBD	N terminus- FLAG C terminus- HA, double strep	Expression observed
DOCK3(1-1920)	pFBD	N terminus- FLAG C terminus- HA, double strep	Expression observed (Less soluble)
NEDD9_FL	pFBD	C terminus- 6XHIS	Expression observed
NEDD9(1-400)	pFBD	C terminus- 6XHIS	Expression observed
DOCK3_NEDD9	pFBD	DOCK3 (N terminus- FLAG C terminus- HA, double strep) NEDD9 (C terminus- 6XHIS)	Expression observed (Less soluble)
DOCK3_NEDD9_GFP	pFBD	DOCK3 (N terminus- FLAG C terminus- HA, double strep, GFP, 6XHIS) NEDD9 (C terminus- 6XHIS)	Expression observed

Table A.1. *Different clones of DOCK3 and NEDD9 with their construct details and their expression levels.*

References

1. Fidler, I. J. Critical determinants of cancer metastasis: Rationale for therapy. in *Cancer Chemotherapy and Pharmacology, Supplement* vol. 43 (Springer Verlag, 1999).

2. Boureux, A., Vignal, E., Faure, S. & Fort, P. Evolution of the Rho family of Ras-like GTPases in eukaryotes. *Mol. Biol. Evol.* **24**, 203–216 (2007).

3. Wennerberg, K., Rossman, K. L. & Der, C. J. The Ras superfamily at a glance. *J. Cell Sci.* **118**, 843–846 (2005).

4. Hanna, S. & El-Sibai, M. Signaling networks of Rho GTPases in cell motility. *Cellular Signalling* vol. 25 1955–1961 (2013).

5. Hoon, J., Tan, M. & Koh, C.-G. The Regulation of Cellular Responses to Mechanical Cues by Rho GTPases. *Cells* **5**, 17 (2016).

6. Gray, J. L., von Delft, F. & Brennan, P. E. Targeting the Small GTPase Superfamily through Their Regulatory Proteins. *Angewandte Chemie - International Edition* vol. 59 6342–6366 (2020).

7. Kimura, K. *et al.* Regulation of myosin phosphatase by Rho and Rho-associated kinase (Rho- kinase). *Science (80-.).* **273**, 245–248 (1996).

8. Amano, M. *et al.* Formation of actin stress fibers and focal adhesions enhanced by Rho- kinase. *Science (80-.).* **275**, 1308–1311 (1997).

9. Rho GTPases and cell migration - PubMed. https://pubmed.ncbi.nlm.nih.gov/11683406/.

10. Sahai, E. & Marshall, C. J. RHO - GTPases and cancer. *Nature Reviews Cancer* vol. 2 133–142 (2002).

11. DeGeer, J. & Lamarche-Vane, N. Rho GTPases in neurodegeneration diseases. *Experimental Cell Research* vol. 319 2384–2394 (2013).

12. Hanawa-Suetsugu, K. *et al.* Structural basis for mutual relief of the Rac guanine nucleotide exchange factor DOCK2 and its partner ELMO1 from their autoinhibited forms. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 3305–3310 (2012).

13. Shi, L. Dock protein family in brain development and neurological disease.

Communicative and Integrative Biology vol. 6 (2013).

14. Côté, J. F., Motoyama, A. B., Bush, J. A. & Vuori, K. A novel and evolutionarily conserved Ptdlns(3,4,5)P3-binding domain is necessary for DOCK180 signalling. *Nat. Cell Biol.* **7**, 797–807 (2005).

15. Kukimoto-Niino, M. *et al.* Structural Basis for the Dual Substrate Specificity of DOCK7 Guanine Nucleotide Exchange Factor Article Structural Basis for the Dual Substrate Specificity of DOCK7 Guanine Nucleotide Exchange Factor. *Struct. Des.* **27**, 741-748.e3 (2019).

16. Kulkarni, K., Yang, J., Zhang, Z. & Barford, D. Multiple factors confer specific Cdc42 and Rac protein activation by dedicator of cytokinesis (DOCK) nucleotide exchange factors. *J. Biol. Chem.* **286**, 25341–25351 (2011).

17. Ushijima, M. *et al.* The rac activator DOCK2 mediates plasma cell differentiation and IgG antibody production. *Front. Immunol.* **9**, (2018).

18. Kikuchi, T. *et al.* Dock2 participates in bone marrow lympho-hematopoiesis. *Biochem. Biophys. Res. Commun.* **367**, 90–96 (2008).

19. Ippagunta, S. K. *et al.* The inflammasome adaptor ASC regulates the function of adaptive immune cells by controlling Dock2-mediated Rac activation and actin polymerization. *Nat. Immunol.* **12**, 1010–1016 (2011).

20. Kunimura, K., Uruno, T. & Fukui, Y. DOCK family proteins: Key players in immune surveillance mechanisms. *International Immunology* vol. 32 5–15 (2020).

21. Dobbs, K. *et al.* Inherited DOCK2 Deficiency in Patients with Early-Onset Invasive Infections. *N. Engl. J. Med.* **372**, 2409–2422 (2015).

22. Cimino, P. J. *et al.* Ablation of the microglial protein DOCK2 reduces amyloid burden in a mouse model of Alzheimer's disease. *Exp. Mol. Pathol.* **94**, 366–371 (2013).

23. Dulak, A. M. *et al.* Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat. Genet.* **45**, 478–486 (2013).

24. Sanui, T. *et al.* DOCK2 regulates Rac activation and cytoskeletal reorganization through interaction with ELMO1. *Blood* **102**, 2948–2950 (2003).

25. Kashiwa, A. *et al.* Isolation and characterization of a novel presenilin binding protein. *J. Neurochem.* **75**, 109–116 (2000).

26. Namekata, K., Enokido, Y., Iwasawa, K. & Kimura, H. MOCA Induces Membrane Spreading by Activating Rac1. *J. Biol. Chem.* **279**, 14331–14337 (2004).

27. Sanz-Moreno, V. *et al.* Rac Activation and Inactivation Control Plasticity of Tumor Cell Movement. *Cell* **135**, 510–523 (2008).

28. Astier, A. *et al.* The related adhesion focal tyrosine kinase differentially phosphorylates p130(Cas) and the Cas-like protein, p105(HEF1). *J. Biol. Chem.* **272**, 19719–19724 (1997).

29. Manié, S. N. *et al.* Involvement of p130(Cas) and p105(HEF1), a novel Cas-like docking protein, in a cytoskeleton-dependent signaling pathway initiated by ligation of integrin or antigen receptor on human B cells. *J. Biol. Chem.* **272**, 4230–4236 (1997).

30. Minegishi, M. *et al.* Structure and function of Cas-L, a 105-kD Crk-associated substrate- related protein that is involved in β1 integrin-mediated signaling in lymphocytes. *J. Exp. Med.* **184**, 1365–1375 (1996).

31. Iwata, S. *et al.* HTLV-I Tax induces and associates with Crk-associated substrate lymphocyte type (Cas-L). *Oncogene* **24**, 1262–1271 (2005).

32. Donninger, H. *et al.* Whole genome expression profiling of advance stage papillary serous ovarian cancer reveals activated pathways. *Oncogene* **23**, 8065–8077 (2004).

33. Li, Y. *et al.* HEF1, a novel target of Wnt signaling, promotes colonic cell migration and cancer progression. *Oncogene* **30**, 2633–2643 (2011).

34. Friedl, P. & Wolf, K. Plasticity of cell migration: A multiscale tuning model. *Journal of Cell Biology* vol. 188 11–19 (2010).

35. Ahn, J., Sanz-Moreno, V. & Marshall, C. J. The metastasis gene NEDD9 product acts through integrin β3 and Src to promote mesenchymal motility and inhibit amoeboid motility. *J. Cell Sci.* **125**, 1814–1826 (2012).

36. Mezulis, S., Yates, C. M., Wass, M. N., E Sternberg, M. J. & Kelley, L. A. The Phyre2 web portal for protein modeling, prediction and analysis. (2015) doi:10.1038/nprot.2015.053.

37. Berger, I., Fitzgerald, D. J. & Richmond, T. J. Baculovirus expression system for heterologous multiprotein complexes. *Nat. Biotechnol.* **22**, 1583–1587 (2004).

38. Scheres, S. H. W. *Single-particle processing in relion-3.1*. (2019).

39. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).

40. Chang, L. *et al.* Structure of the DOCK2−ELMO1 complex provides insights into regulation of the auto-inhibited state. doi:10.1038/s41467-020-17271-9.

Appendix B

List of Primers

Primer Name	5' - 3' sequence
Fw_SaSQS1	CGGGGAU ATG TTT AAT GGA GAC GAT TCG TCG GTG GT
Rv_SaSQS1	CCCCGAU TTACTCCTCATCTAGCGTAATTGGGTGAATT
Fw_T313S	GAGCTTTAATATCGACAATCGATGATGTCT
Rv_T313S	AGACATCATCGATTGTCGATATTAAAGCTC
Fw_T313V	GAGCTTTAATAGTAACAATCGATGATGTCT
Rv_T313V	AGACATCATCGATTGTTACTATTAAAGCTC
Fw_Q387S	AAAGTGTGGGCTGATTCACTAAAAAGTTAC
Rv_Q387S	GTAACTTTTTAGTGAATCAGCCCACACTTT
Fw_Q387L	AAAGTGTGGGCTGATCTTCTAAAAAGTTAC
Rv_Q387L	GTAACTTTTTAGAAGATCAGCCCACACTTT
Fw_Y391F	ATCAACTAAAAAGTTTCACCAAAGAAGCAA
Rv_Y391F	TTGCTTCTTTGGTGAAACTTTTTAGTTGAT
Fw_Y391V	ATCAACTAAAAAGTGTCACCAAAGAAGCAA
Rv_Y391V	TTGCTTCTTTGGTGACACTTTTTAGTTGAT
Fw_S416A	AATGCTCTGGTCGCCATAGGATTCCCA
Rv_S416A	TGGGAATCCTATGGCGACCAGAGCATT
Fw_I417F	CTCTGGTCTCCTTTGGATTCCCAAACCTT
Rv_I417F	AAGGTTTGGGAATCCAAAGGAGACCAGAG

Fw_G418A	AATGCTCTGGTCTCCATAGCATTCCCAAA
Rv_G418A	TTTGGGAATGCTATGGAGACCAGAGCATT
Fw_F419A	TGGTCTCCATAGGAGCCCCAAACCTTCTT
Rv_F419A	AAGAAGGTTTGGGGCTCCTATGGAGACCA
Fw_R457K	TCTTGCATTCTTTGTAAGATCATCAACGA
Rv_R457K	TCGTTGATGATCTTACAAAGAATGCAAGA

Abstract

Name of the Student: Sneha Singh **Registration No. :** 10BB14A26043

Faculty of Study: Biological Science **Year of Submission:** 2021

AcSIR academic centre/CSIR Lab: CSIR-National Chemical Laboratory, Pune

Name of the Supervisor(s): Dr. Kiran Kulkarni

Name of the Co-Supervisor(s): Dr. H.V. Thulasiram

Title of the thesis: Structural studies on Sesquisabinene Synthase 1: Enzyme involved in Terpene Biosynthesis Pathway

Terpenes or Isoprenoids are a diverse group of natural products that function as primary and secondary metabolites in plants. This huge diversity can be attributed to terpene synthases. Amongst these class of enzymes, Sesquiterpene synthases catalyze the cyclization of FPP (farnesyl pyrophosphate) to form sesquiterpenes. All the sesquiterpene synthases share a conserved catalytic mechanism. However, there is a lesser understanding of the structural rationale on the dynamics of the enzyme, which results in product diversity. The present work focuses on structural studies of sesquisabinene synthase 1 (SaSQS1), a sesquiterpene synthase found in Indian sandalwood, which catalyzes the cyclization of FPP to form sesquisabinene, a key component of sandalwood oil. We elucidated the crystal structure SaSQS1 to understand its mechanism and product specificity. The structure provides insights to the domain architecture of the enzyme. Furthermore, biochemical studies of SaSQS1 show that the classical approach based on sequence and structure comparative studies fails to identify the residues that may regulate its product specificity. Structural studies of two important inactive mutants of SaSQS1 provide insights to the differential dynamics of these mutants that may play role in the quantum of product formed. Due to various limitations of the classical method, we developed a novel approach combining sequence, structural and dynamical information of plant sesquiterpene synthases to predict product modulating residues (PMRs). We tested this approach on the sesquiterpene synthases with known PMRs and also on SaSQS1. Our results show that the dynamical sectors of sesquiterpene synthases with their hydrophobicity and vicinity indices provide leads for the identification of PMRs. Together, the work reported in this thesis attempts to study the structural basis of mechanism of action of SaSQS1 and understand the product specificity of the enzyme using a novel approach.

Publications emanating from the thesis work

1. Dynamic coupling analysis on plant sesquiterpene synthases provides leads for the identification of product specificity determinants.
   S. Singh, H. V. Thulasiram, D. Sengupta, K. Kulkarni, Biophys. Res. Commun. 536 (2021) 107–114. https://doi.org/10.1016/j.bbrc.2020.12.041.
2. Biochemical and structural studies of mutants of sesquisabinene synthase 1 indicating their role in product specificity (manuscript under preparation).

List of papers with abstract presented (poster) at National/International conferences or seminars

1. Sneha Singh, H.V. Thulasiram*, Kiran Kulkarni*

Structural Studies of Sesquiterpene Synthase 1.

Science Day, CSIR-NCL, 2015

Abstract

Isoprenoids(terpenes) consist of highly diverse group of natural products which have numerous functions in primary and secondary metabolism of plants. These metabolites have found a number of applications as pharmaceutical drugs. Isoprenoids are classified into different groups according to the number of carbons they contain: Monoterpenes(C10), Sesquiterpenes(C15), Diterpenes(C20),Triterpenes(C30). Prenyl diphosphate precursors are converted into isoprenoids through the action of terpene synthases and terpene modifying enzymes. One such synthase is Sesquiterpene synthase 1 that cyclizes FPP (Farnesyl DiPhosphate) into sesquisabinene and sesquiphellandrene. Structural studies of sesquiterpene synthases have provided insights to the mechanism of action of these enzymes. Here, we have tried to elucidate the structure of Sesquiterpene synthase 1 in order to understand its mechanism of action and product specificity.

Biochemical and Biophysical Research Communications 536 (2021) 107–114

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Dynamic coupling analysis on plant sesquiterpene synthases provides leads for the identification of product specificity determinants

Sneha Singh [a, d], Hirekodathakallu V. Thulasiram [b, d, **], Durba Sengupta [c, d], Kiran Kulkarni [a, d, *]

[a] Division of Biochemical Sciences, CSIR – National Chemical Laboratory, Pune, 411008, India
[b] Division of Organic Chemistry, CSIR – National Chemical Laboratory, Pune, 411008, India
[c] Division of Physical and Materials Chemistry, CSIR – National Chemical Laboratory, Pune, 411008, India
[d] Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, 201002, India

ARTICLE INFO

Article history:
Received 2 December 2020
Accepted 13 December 2020

Keywords:
Sesquiterpene synthases
Product specificity
X-ray crystallography
Statistical coupling analysis
Molecular dynamics

ABSTRACT

Sesquiterpene synthases catalyse cyclisation of farnesyl pyrophosphate to produce diverse sesquiterpenes. Despite utilising the same substrate and exhibiting significant sequence and structural homology, these enzymes form different products. Previous efforts were based on identifying the effect of divergent residues present at the catalytic binding pocket on the product specificity of these enzymes. However, the rationales deduced for the product specificity from these studies were not generic enough to be applicable to other phylogenetically distant members of this family. To address this problem, we have developed a novel approach combining sequence, structural and dynamical information of plant sesquiterpene synthases (SSQs) to predict product modulating residues (PMRs). We tested this approach on the SSQs with known PMRs and also on sesquisabinene synthase 1 (SaSQS1), a SSQ from Indian sandalwood. Our results show that the dynamical sectors of SSQs obtained from molecular dynamics simulation and their hydrophobicity and vicinity indices together provide leads for the identification of PMRs. The efficacy of the technique was tested on SaSQS1 using mutagenesis. To the best of our knowledge, this is a first technique of this kind which provides cues on PMRs of SSQs, with divergent phylogenetic relationship.

1. Introduction

Terpenes are a diverse group of natural products found in plants, fungi and bacteria. In plants, these compounds function as metabolites like carotenoids, sterols, hormones, which are also known to play crucial role in plant defense mechanism and communication [1]. Depending on the number of fundamental isoprene units (C_5), terpenes are classified as monoterpenes (C_{10}), sesquiterpenes (C_{15}), diterpenes (C_{20}), triterpenes (C_{30}) and tetraterpenes (C_{40}). Till date, more than 7000 chemically diverse forms of sesquiterpenes are known, with greater than 300 different carbon skeletons which have applications as pharmaceuticals, highly valued essential oils and fragrances [2–5]. These compounds are formed from farnesyl pyrophosphate (FPP), which undergoes a series of chemical reactions catalyzed by sesquiterpene synthases (SSQs). The enzyme produces highly reactive carbocation intermediates, which subsequently leads to formation of particular skeletal forms [6,7]. The contour of active site serves as a template for the catalysis and controls the conformations attained by these intermediates, providing product diversity.

Indian sandalwood (*Santalum album*) oil has great commercial value due to its pleasant and woody fragrance including applications in industries like cosmetics, perfumery and aromatherapy and acts as an anti-inflammatory, antidepressant, antifungal and antiseptic agent [8]. Sesquisabinene, one of the important components of sandalwood oil, is formed from FPP through electrophilic intramolecular cyclisation reaction, catalyzed by sesquisabinene synthase. In sandalwood, two isoforms of sesquisabinene synthases, sesquisabinene synthase 1(SaSQS1) and sesquisabinene synthase

* Corresponding author. Division of Biochemical Sciences, CSIR – National Chemical Laboratory, Pune, 411008, India.
** Corresponding author. Division of Organic Chemistry, CSIR – National Chemical Laboratory, Pune, 411008, India.
E-mail addresses: hv.thulasiram@ncl.res.in (H.V. Thulasiram), ka.kulkarni@ncl.res.in (K. Kulkarni).

https://doi.org/10.1016/j.bbrc.2020.12.041
0006-291X/

2(SaSQS2) have been reported [3]. We (PDB ID: 6K16) and others [9] have simultaneously elucidated structure of SaSQS1 to gain insights on the catalytic mechanism and product specificity of the enzyme.

Despite exhibiting significant sequence and structural homology, especially in the catalytic pocket nestling α domain, SSQs form diverse products. The factors that determine the product specificity of this family of enzymes have still remained unclear. Previous efforts in determining the product modulating residues (PMRs) were focused on extensively testing structure based mutants of divergent residues at the binding pocket, especially those which lie within a radius of 6 Å from the ligand [10]. However, given the size of the binding pocket, the possible combinations of residues that could influence type of the product are astronomical [10,11]. Although, this approach seems to have worked for few SSQs like 5-*epi*-aristolochene synthase (TEAS) [11], Vetispiradiene synthase [12], β-farnesene synthase [10] and α-bisabolol synthase (AaBOS) [13], the rationale underlying the identification of PMRs appear to be failing when applied across the other divergent members of the family (Fig. S1). Furthermore, in all of the above cases, the methodology employed in devising the subset of the divergent residues markedly varies. Thus, a generalized approach that could provide putative set of PMRs of SSQs is lacking. Here, we have attempted to develop a technique that employs a combination of sequence, structure and dynamical features of SSQs. Using spectral decomposition technique [14] we have identified dynamical "sectors" that harbor PMRs, which were further, refined by combining sequence based hydrophobicity index and structure based ligand vicinity index to identify most plausible PMRs. The methodology developed was validated against the SSQs with known PMRs and also provide putative PMRs of SaSQS1.

2. Materials and methods

2.1. Cloning, expression, purification, crystallization, mutagenesis and activity assay

SaSQS1 construct was designed based on the disorder prediction from PHYRE^2 [15]. The gene (GenBank accession number: KJ665776.1) encoding 17–566 amino acids was cloned in pOPINss vector with a cleavable N-terminal 6× His-SUMO tag. The protein was expressed in C41 strain of *Escherichia coli* and purified by gel filtration chromatography following the NiNTA chromatography. SaSQS1 was crystallized using the sitting-drop vapor diffusion method and structure was solved by molecular replacement technique. Site directed mutagenesis was performed. Activity assay was performed following the reported protocol [3]. Detailed methodology employed in all these steps can be found in supplementary methods section.

2.2. Molecular dynamics (MD) simulation

MD simulation for the structures of SaSQS1 (PDB ID: 6K16), TEAS (PDB ID: 5IK0), AaBOS (PDB ID: 4FJQ) and mutant *M2* of AaBOS (PDB ID: 4GAX) was performed using Maestro - Desmond of Schrödinger suite 2020–2. Details of the simulation are given in the supplementary methods section.

2.3. Sequence based statistical coupling analysis (sSCA)

The python version of SCA module, PySCA (http://reynoldsk.github.io/pySCA/) was used to generate sequence based sectors [14]. For sSCA calculation, sequences from curated plant SSQs database (www.bioinformatics.nl/sesquiterpene/synthasedb) were used [16]. The database contains 248 sequences, out of which 139 were identified as unique sequences by the program. The multiple sequence alignment (MSA) was trimmed at both N and C terminus such that it contains the α domain, starting from residue G271 to Q537 (residue numbers as per SaSQS1 sequence) and further processed as described by Rivoire et al. [14]. Briefly, the coevolution matrix (sC_{ij}) was calculated from the MSA. The group of coevolving residues, known as sectors (sICs) were identified from the spectral decomposition (diagonalization) of C_{ij} [14]. The significant eigenmodes (K^*) obtained from decomposition represent a set of coevolving amino acids, which were used for calculating the independent sectors.

2.4. Molecular dynamics based sector identification

The dynamical cross-correlation matrix (DCCM), dC_{ij} [17], was calculated as:

$$dC_{ij} = < r_i r_j > - < r_i > < r_j >$$

where, r_i and r_j are the position vectors of the *i*th and *j*th Cα atoms, respectively, at time *t*. The R implementation of Bio3D program (version 2.4–1) [18] was used for the DCCM calculations. The dynamical sectors were obtained from the spectral decomposition algorithm developed by Rivoire et al. [14]. The MATLAB implementation of SCA was employed for this purpose. The dimensions of dC_{ij} were kept identical to that of sC_{ij}.

The hydrophobicity and vicinity indices were calculated for these dynamical sectors. The hydrophobicity index (HI) is the average kd hydrophobicity scores of the sectors [19]. To calculate the vicinity index (VI), a pseudo reference atom was fixed at the center of mass of the Cα atoms of residues A289, I312, G418 and C456 (the residue type and numbers are according to the SaSQS1 structure and for other structures, equivalent residues were taken) (Fig. S2). The vicinity index for a dynamical sector, dIC, is calculated as:

$$V = \frac{1}{N} \sum_{i \forall dIC} \delta_i$$

where, *N* is total number of residues in sector dIC and $\delta_i = 1$ if the Cα atom of the *i*th residue lies within 12 Å from the reference atom, else $\delta_i = 0$.

3. Results and Discussion

3.1. Structural deviations suggest dynamical interplay in the catalytic mechanism of SSQs

We and others [9] determined structure of SaSQS1 to elucidate its catalytic mechanism and product specificity. The enzyme exhibits classical αβ helical domain architecture characteristic of plant class I terpenoid cyclases. The catalytic C-terminal α domain comprises of fourteen helices (α′1- α′14) out of which seven helices form the core substrate binding pocket and nestle the conserved DDXXD and NSE/DTE motifs (Fig. 1). These motifs are known to coordinate the Mg^{2+} ions, essential for catalysis. The conserved catalytic mechanism involves stabilization of the bound FPP by forming enzyme - Mg^{2+} - ligand ternary complex, leading to the formation of highly reactive carbocation intermediates, which undergo hydride transfer and several rearrangement steps. The final product is formed on quenching of the carbocation intermediates, mediated either by water molecule or through deprotonation by the residues of the enzyme. The structure reported here (PDB ID: 6K16) is by and large identical to the other reported structures (PDB ID: 6O9Q, 6O9P, 6A3X, 6A1I, 6A1E, 6A1D), however, conformational deviations are seen at the conserved DDXXD and RXR(G) regions

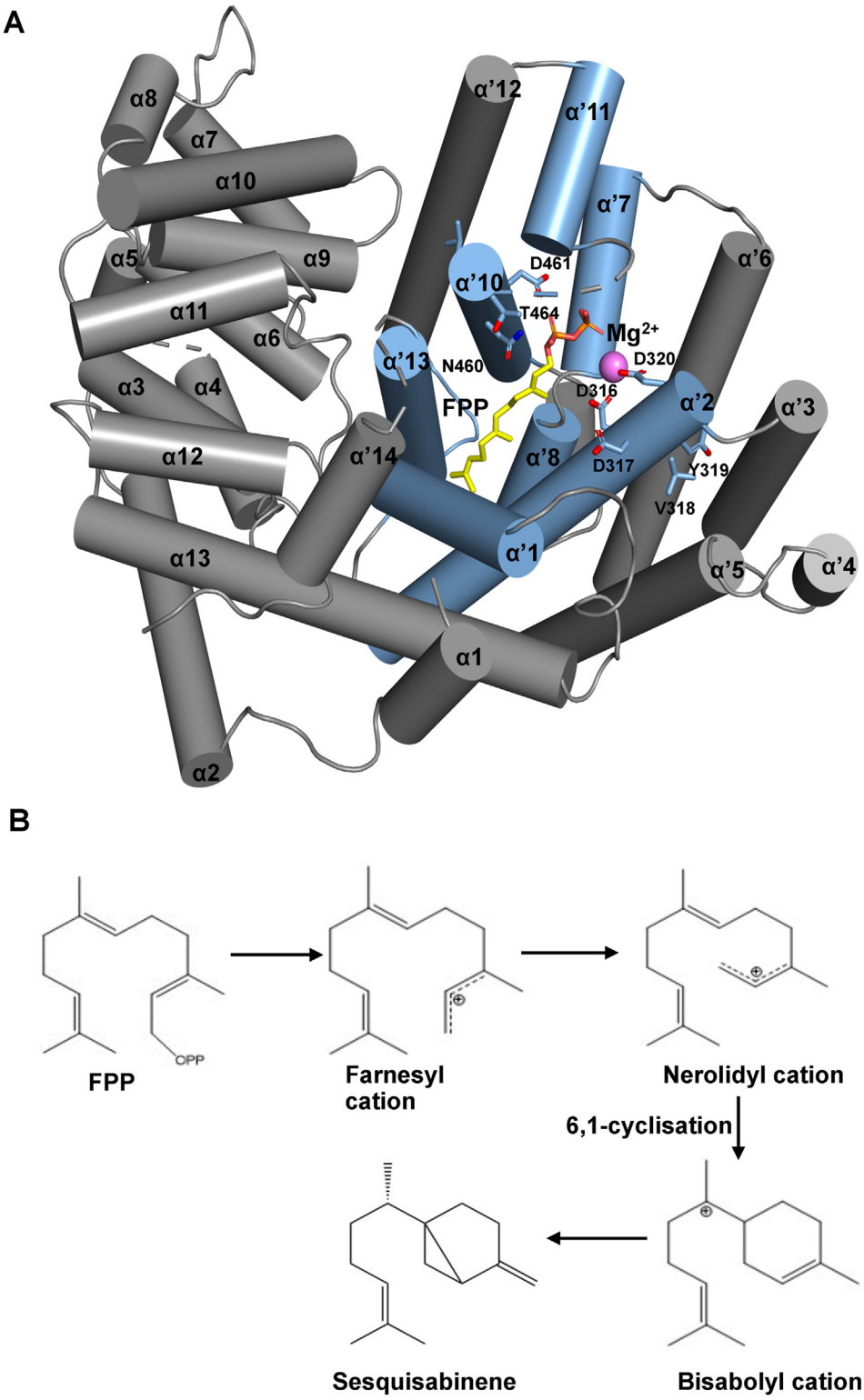


Fig. 1. Structure and mechanism of SaSQS1. (A) Overall structure of SaSQS1 docked with FPP. The core of the binding pocket colored in blue and the ligand & Mg^{2+} ions are shown in ball and stick. The conserved motifs, DDXXD and NSE/DTE, are highlighted at helices α′2 and α′10, respectively. (B) Proposed cyclisation mechanism for conversion of FPP to sesquisabinene involving generation of carbocation intermediates.

(Fig. S3 A, C). However, the reason behind these conformational deviations is not clear from the structures.

Recently, from earlier studies on TEAS, Blank et al. have proposed that SaSQS1 exhibits open and closed states upon ligand binding. These states are attributed to the conformational changes in the RXR loop leading to the interaction between the first Arg and second Asp of RXR and DDXXD motifs, respectively [9,20]. This transition was suggested to be essential for protecting the reactive carbocation intermediates from water mediated nucleophilic attack. However, comparative structural analysis on SSQs, such as

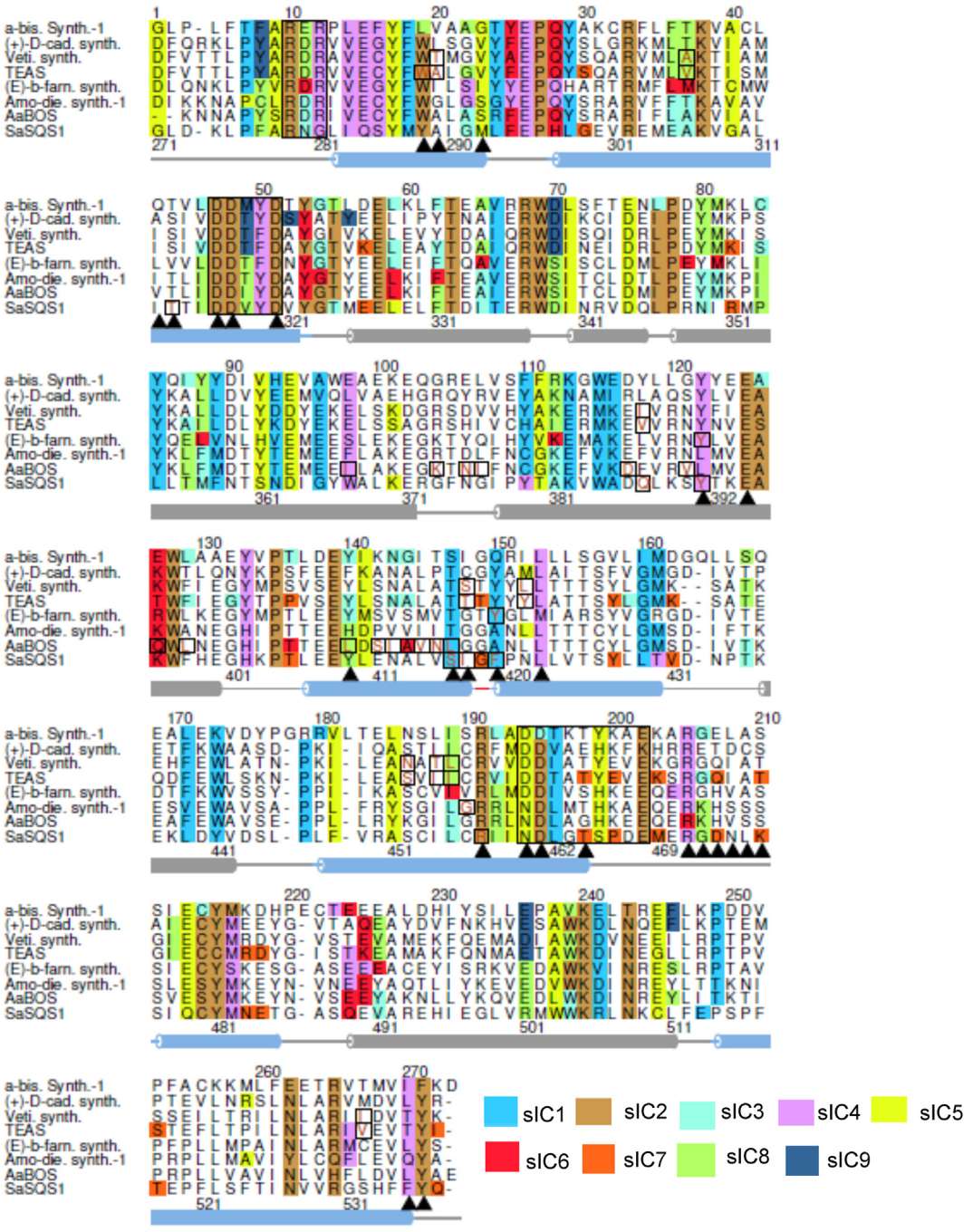


Fig. 2. Sequence based sectors (sICs) mapped on multiple sequence alignment of plant SSQs (α domain). Conserved motifs like DDXXD, NSE/DTE and RXR are shown in boxes. The core and adjacent catalytic site helices, corresponding to the SaSQS1 structure, are shown in blue and grey colors, respectively. The residues corresponding to the sectors (sICs), from sIC1-sIC9, are highlighted as per the color-code given in the figure. The residues of SaSQS1 that lie within 6 Å from the ligand are marked with arrows. The known PMRs of SSQs are highlighted in brown color with box. For SaSQS1 these boxes indicate the residues considered for mutagenesis study.

wild type & *M2* mutant of AaBOS (AaBOS^{M2}) (Fig. S4A) and *apo* and ligand bound form of α-bisabolene synthase provides contrary observations (Fig. S4B). Furthermore, other reported structures of SaSQS1 also suggest that this transition may have little influence on the catalytic efficiency of the enzyme (Fig. S4C). However, the comparative structural analysis of SaSQS1 with its other forms (Fig. S3A) and also with its homologs (Fig. S3B), suggest that the catalytic pocket is highly dynamic and could play some role in other aspects of the enzyme such as product specificity.

3.2. Role of evolutionary coupling of sequences in defining the PMRs of plant SSQs

Although, sequence and structure based approaches in identifying PMRs were limited to few members of SSQs, the proposed rationale behind the methodology seems to be inapplicable when extrapolated to other more divergent members of the family. For example, the products of TEAS and Vetispiradiene synthase were switched by introducing mutation by combination of 9 residues

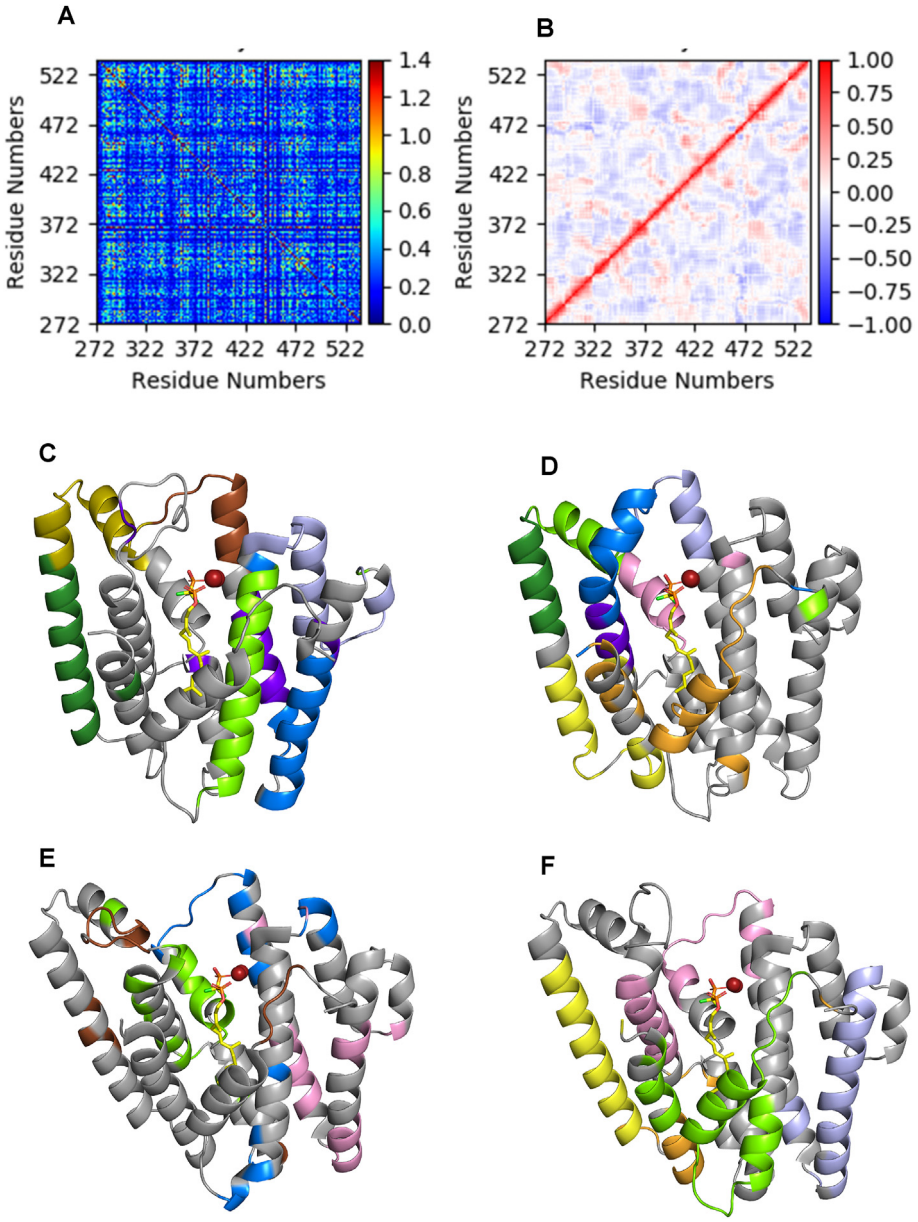


Fig. 3. Dynamical sectors of SSQs. (A) Sequence and (B) MD simulation based covariance matrices for SaSQS1. Dynamical sectors mapped on the α domain structures (C) SaSQS1(PDB ID: 6K16) (D) TEAS (PDB ID: 5IK0) (E) AaBOS (PDB ID: 4FJQ) (F) AaBOS^{M2} (PDB ID: 4GAX). For all the structures following scheme for dICs color coding is used: dIC1- dark blue, dIC2- brown, dIC3- lime, dIC4- pink, dIC5- yellow, dIC6- violet, dIC7- orange, dIC8- green, dIC9- purple, dIC10- golden. Ligand & Mg^{2+} ions are shown in ball and stick.

(A274, V291, V372, T402, Y406, S436, I438, I439, V516; sequence numbering is as per TEAS) [11,12,21]. In a similar exercise, mutations of two residues (Y402, Y430) in β -farnesene synthase were observed to alter the product from β-farnesene to cyclic products [10] (Table S3). Likewise, domain swapping experiments between AaBOS and Amorpha-4,11-diene synthase led to the identification of different combinations of a set of residues in AaBOS (I350, K356, N358, I359, D369, V373, Q379, L381, L392, S394, I395, A396, V397, N398 & L399), which when mutated generate an additional product, γ-humulene [13] (Table S3). It is interesting to note that product modulation requires alterations in a set of residues rather than one or two specific positions, sometimes including residues distant from the active site (Fig. S5). These observations suggest that an intrinsic coupling amongst the catalytic site residues alter the product formed. We probed the positional coupling in SSQs using sequence based statistical coupling analysis (sSCA). As explained in the methods section, sSCA computes covariance matrix (sCij) based on the sequence conservation and decomposes to identify sectors in the protein, which are shown to have functional correlation [14]. We explored this approach to identify PMRs in SSQs. We mapped the sICs on MSA and respective structures of eight plant SQSs (α-bisabolene synthase, Delta cadinene synthase, Vetispiradiene synthase, TEAS, β-farnesene synthase, Amorpha-4,11-diene synthase 1, Amorpha-4,11-diene synthase 2 or AaBOS and SaSQS1) (Fig. 2, Fig. S7). Interestingly, first five sICs, except sIC3, can be mapped to the conserved residues, encompassing functionally important motifs such as DDXXD, RXR and NSE/DTE, across all of the eight SSQs (Fig. 2). Rest of the three sICs (sIC6 to sIC8) largely map on the divergent residues at the catalytic pocket of SSQs (Fig. 2). Since different sICs are suggested to indicate divergent functional states of proteins, we checked whether these sectors indicate PMRs [10–13,22,23]. Surprisingly, the sICs of SSQs do not

Table 1
Hydrophobicity index and vicinity index (in the parenthesis) values of different dICs of SSQs. The sectors which contain PMRs are highlighted in orange. Putative PMRs containing dICs, obtained from the current analysis, are highlighted in violet. NU include the dICs which are not unique and mapped the entire α domain of SSQs, while NE are not existing sectors in a particular SSQ.

SSQ	IC1	IC2	IC3	IC4	IC5	IC6	IC7	IC8	IC9	IC10
SaSQS1	-0.43 (0.00)	-1.85 (0.00)	0.47 (0.54)	NU	NU	-0.54 (0.00)	NU	-0.29 (0.05)	-0.18 (0.12)	-0.47 (0.00)
TEAS	-0.97 (0.00)	NU	-0.25 (0.00)	0.57 (0.45)	-0.30 (0.00)	0.19 (0.00)	0.18 (0.46)	-0.39 (0.00)	2.84 (0.43)	NE
AaBOS (*wt*)	-0.36 (0.05)	-1.70 (0.00)	0.24 (0.33)	0.46 (0.04)	NU	NE	NE	NE	NE	NE
AaBOS (*M2*)	NU	NU	0.14 (0.40)	-0.76 (0.13)	-0.88 (0.00)	-0.26 (0.00)	1.14 (0.06)	NE	NE	NE

show any clustering of their PMRs. Although a subset of PMRs appears to fall in few sICs, they do not show any pattern in the hierarchy of sICs (Fig. 2). For example, PMRs of Vetispiradiene synthase (A298 & L446) and of TEAS (V291 & I439) can be mapped to sIC8. However, in both, just these two residues are not adequate enough to modulate their product [11,12,21]. Even the intra-sector divergent residues do not show any well-defined correlation between their respective sICs and their ability to modify the product. This can be seen for AaBOS, where members of sIC6, Q379 and A396, exhibit sequence divergence within the sector but do not change the cognate product of the enzyme, when mutated (Fig. 2). In summary, sICs help to identify functionally important residues of the SSQ family, however, do not indicate any particular patterns for recognizing PMRs.

3.3. Dynamical sectors provide leads to identify product-defining residues

Besides sequence and three-dimensional structure, the dynamical features of proteins are equally consequential in dictating their function. Therefore, the dynamical coupling in SSQs were explored to identify the PMRs, as few members of this family exhibit dynamical rearrangement of catalytic pocket upon ligand binding. To achieve this, we performed MD simulation on the FPP bound structures of SaSQS1, TEAS, AaBOS and AaBOS^{M2} (Fig. S6). In case of non-availability of ligand bound structures, FPP was docked on these structures. All the MD simulations were performed in triplicates and trajectories were pooled for further analysis, which resulted in 18,000 frames for each system (protein). DCCM matrix for the residues belonging to the α domain were generated from the pooled trajectories. The spectral decomposition of the DCCM provides set of residues, what we call dynamical sectors (dICs), which are dynamically coupled. The hierarchy of dICs for different SSQs are shown in Fig. 3C–F. Unlike for the PDZ domain [24], the sICs and dICs of SSQs hardly overlap (Fig. 3, Fig. S7). Furthermore, despite having homologous structures, TEAS and AaBOS (rmsd is 2.81) do not have identical dICs, *i.e.*, the corresponding structures do not have structurally equivalent residues (Fig. 3D and E).

Interestingly, the PMRs of TEAS and AaBOS are found to be clustering within particular dICs (Table S4). However, the hierarchy of these dICs between the two structures is not unique. For example, the PMRs of AaBOS are found to be falling in dIC3 (L392, S394, I395, A396, V397, N398, L399) and dIC4 (I350, I359, D369). Similarly, the PMRs of TEAS belong to dIC7 (W273, A274, V516) and dIC9 (I438 & I439). Furthermore, the dICs of AaBOS (Fig. 3E) and AaBOS^{M2} (Fig. 3F) are markedly different. Since, AaBOS^{M2} is still capable of producing terpenoids, γ-humulene, which is a non-cognate product of AaBOS, we are tempted to argue that the mutant has been essentially transformed into a new type of SSQ, which results in dICs different from the wild type. This indicates that there is correlation between the product formed and the dICs of SSQs. Another interesting observation is that the PMR containing dICs extend beyond the catalytic pocket, indicating structural allostery in SSQs. Thus, dICs provide a pattern for PMRs for two highly sequentially divergent SSQs, which was otherwise not clear from sequence and structure comparison studies.

Although, dICs cluster the PMRs, their hierarchy varies amongst structures. For instance, PMRs of TEAS are in dIC7 and dIC9 whereas for AaBOS they are found in dIC3 and dIC4 (Table S4). Therefore, to further narrow down on the PMR containing dICs, we augmented dynamical information with structural and sequence data. Ligand vicinity and hydrophobicity of the catalytic site residues are known to play important role in the reaction cascade by protecting the carbocation intermediates [6,25] and hence the product formed [25–28]. Therefore, we calculated average kd hydrophobicity values (hydrophobicity index, HI) and vicinity index (VI) for all dICs, as defined in the methods section. Clearly, dICs with positive HI and VI (>0) are likely to harbor the PMRs and interestingly, that is the case for both TEAS and AaBOS (Table 1). Thus, it appears that the group of catalytic site residues of plant SSQs, which are dynamically coupled and have positive HI and VI values modulate their product.

3.4. Effect of individual divergent residues at the binding pocket of SaSQS1 on the product plasticity

To identify PMRs of SaSQS1, we followed the classical approach of mutating divergent residues, which lie within a radius of 6 Å from the ligand. Based on sequence and structure comparative studies of SaSQS1 with its closest homolog TEAS, eleven mutants of SaSQS1 (T313S, T313V, Q387S, Q387L, Y391F, Y391V, S416A, I417F, G418A, F419A and R457K) were tested for product specificity. Here, the rationale for the mutation was based on the earlier observation that both nature of the residue (hydrophobic or hydrophilic) and length affect the product formed. However, in our case no change in the product was seen in these mutants but some of the mutants such as Q387S, Q387L, Y391F, Y391V, S416A and F419A change the product yield (Fig. 4) and few others make the enzyme inactive. Particularly, F419A and Y391V increase the product yield, whereas,

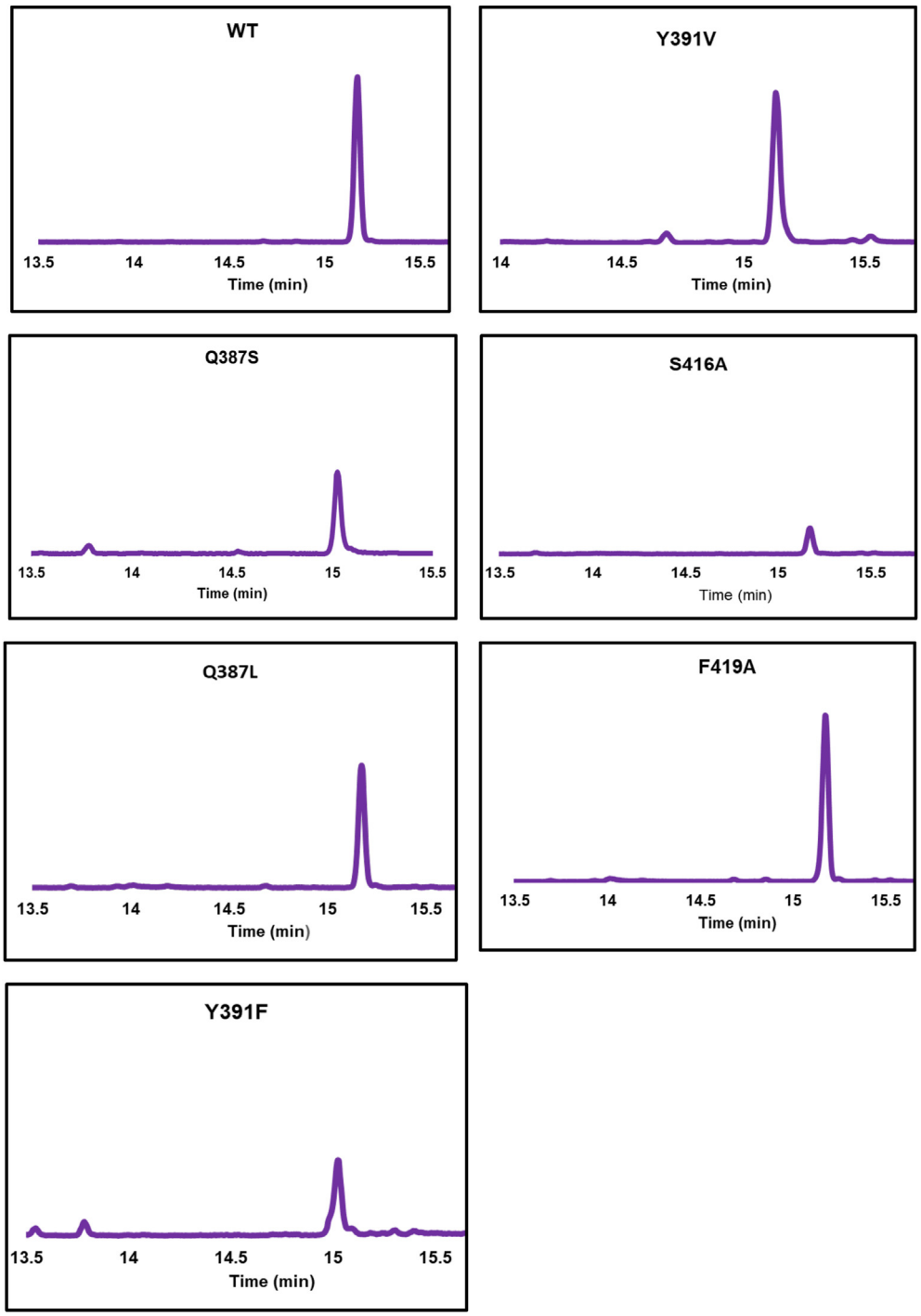


Fig. 4. The GC-MS profiles of products from *wild* type and mutant forms of SaSQS1. Only those mutants, which exhibit the activity are represented here. In all of the chromatograms, the peak corresponds to sesquisabinene.

all others showed the reduction in product level when equal quantity assay mixtures were compared. However, T313S, T313V, I417F, G418A and R457K abolish the enzyme activity. In line with our observations for TEAS and AaBOS, these residues of SaSQS1 do not cluster in any of the dICs (Table S4). However, the residues which abrogate the activity of SaSQS1 on substitution, such as T313, belongs to dIC3. Interestingly, dIC3 has positive HI and VI value greater than zero, which is the criteria developed for identifying potential PMRs. Further, to test the technique developed here on SaSQS1, a combination of residues from dIC3 were tried. Unfortunately, these SaSQS1 mutants were found to be insoluble, forming inclusion bodies. Although, a direct validation of the technique was not possible with SaSQS1, mutation of individual members of dIC3 suggest that indeed they play significant role in influencing the product in terms of its quantity.

4. Conclusion

Engineering SSQs to produce desired products will have enormous impact on the biotechnological applications of the enzymes.

This requires precise knowledge of the residues of the enzyme that control the product plasticity. Earlier sequence and structure based methods to determine PMRs were less generic and involved bottom up approach of testing extensive combination of residues. To address this problem, we developed a new approach combining sequence, structure and dynamical information to predict PMRs which correctly predicts the biochemically validated PMRs of SSQs and also provides leads on identification of PMRs for SaSQS1. However, efficacy of the method could not be tested on SaSQS1 as the combination of predicted mutants were found to be insoluble. In summary, we present a novel approach to determine product modulating residues of plant sesquiterpene synthases which could be further validated by testing it on other members of the family.

Author contributions

S.S. performed all the experiments. H.V.T. provided reagents and supervised the assays. D.S. provided inputs on MD simulations. K.K. conceptualised the problem, supervised overall work, analysed the data and wrote the paper with inputs from S.S.

Accession code

Coordinates and structure factors have been deposited in the Protein Data Bank under accession code 6K16.

Funding

This work was supported by the Department of Biotechnology [grant number: BT/PR12502/BRB/10/1387/2015].

Declaration of competing interest

The authors declare no conflicts of interest.

Acknowledgement

The authors acknowledge XRD1 beamline facility, ELETTRA synchrotron and Dr. Maurizio Polentarutti for diffraction data collection. S.S. acknowledges CSIR, India for fellowship. The authors acknowledge Ms. Aiswarya Pawar for her help with R scripts.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.bbrc.2020.12.041.

References

[1] J. Kirby, J.D. Keasling, Biosynthesis of plant isoprenoids: perspectives for microbial engineering, Annu. Rev. Plant Biol. 60 (2009) 335–355, https://doi.org/10.1146/annurev.arplant.043008.091955.
[2] M.A. Asadollahi, J. Maury, K. Møller, K.F. Nielsen, M. Schalk, A. Clark, J. Nielsen, Production of plant sesquiterpenes inSaccharomyces cerevisiae: effect ofERG9 repression on sesquiterpene biosynthesis, Biotechnol. Bioeng. 99 (2008) 666–677, https://doi.org/10.1002/bit.21581.
[3] P.L. Srivastava, P.P. Daramwar, R. Krithika, A. Pandreka, S. Shiva Shankar, H. V Thulasiram, Functional characterization of novel sesquiterpene synthases from Indian sandalwood, santalum album, Sci. Rep. 5 (2015) 10095, https://doi.org/10.1038/srep10095.
[4] L. Moujir, O. Callies, P.M.C. Sousa, F. Sharopov, A.M.L. Seca, Applications of sesquiterpene lactones: a review of some potential success cases, Appl. Sci. 10 (2020) 3001, https://doi.org/10.3390/app10093001.
[5] A. Bommareddy, B. Rule, A.L. Vanwert, S. Santha, C. Dwivedi, α-Santalol, a derivative of sandalwood oil, induces apoptosis in human prostate cancer cells by causing caspase-3 activation, Phytomedicine 19 (2012) 804–811, https://doi.org/10.1016/j.phymed.2012.04.003.
[6] D.W. Christianson, Structural and Chemical Biology of Terpenoid Cyclases, Chem Rev., 2017, pp. 11570–11648, https://doi.org/10.1021/acs.chemrev.7b00287.
[7] D.J. Miller, R.K. Allemann, Sesquiterpene synthases: passive catalysts or active players? Nat. Prod. Rep. 1 (2011) 60–71, https://doi.org/10.1039/c1np00060h.
[8] J.A. Teixeira Da Silva, M.M. Kher, Deepak Soner, Tony Page, X. Zhang, M. Nataraj, G. Ma, D. Soner, M. Nataraj, M.A. Cn, Sandalwood: basic biology, tissue culture, and genetic transformation, Planta 243 (2016) 847–887, https://doi.org/10.1007/s00425-015-2452-8.
[9] P.N. Blank, S.A. Shinsky, D.W. Christianson, Structure of sesquisabinene synthase 1, a terpenoid cyclase that generates a strained [3.1.0] bridged-bicyclic product, ACS Chem. Biol. 14 (2019) 1011–1019, https://doi.org/10.1021/acschembio.9b00218.
[10] M. Salmon, C. Laurendon, M. Vardakou, J. Cheema, M. Defernez, S. Green, J.A. Faraldos, P.E. O'maille, Emergence of terpene cyclization in Artemisia annua, Nat. Commun. 6 (2015) 6143, https://doi.org/10.1038/ncomms7143.
[11] B.T. Greenhagen, P.E. O'maille, J.P. Noel, J. Chappell, Identifying and manipulating structural determinates linking catalytic specificities in terpene synthases, Proc. Natl. Acad. Sci. Unit. States Am. 103 (2006) 9826–9831. www.pnas.orgcgidoi10.1073pnas.0601605103.
[12] H.J. Koo, C.R. Vickery, Y. Xu, G.V. Louie, P.E. O'Maille, M. Bowman, C.M. Nartey, M.D. Burkart, J.P. Noel, Biosynthetic potential of sesquiterpene synthases: product profiles of Egyptian Henbane premnaspirodiene synthase and related mutants, J. Antibiot. (Tokyo) 69 (2016) 524–533, https://doi.org/10.1038/ja.2016.68.
[13] J.X. Li, X. Fang, Q. Zhao, J.X. Ruan, C.Q. Yang, L.J. Wang, D.J. Miller, J.A. Faraldos, R.K. Allemann, X.Y. Chen, P. Zhang, Rational engineering of plasticity residues of sesquiterpene synthases from Artemisia annua: product specificity and catalytic efficiency, Biochem. J. 451 (2013) 417–426, https://doi.org/10.1042/BJ20130041.
[14] O. Rivoire, K.A. Reynolds, R. Ranganathan, Evolution-based functional decomposition of proteins, PLoS Comput. Biol. 12 (2016) 1004817, https://doi.org/10.1371/journal.pcbi.1004817.
[15] S. Mezulis, C.M. Yates, M.N. Wass, M.J. E Sternberg, L.A. Kelley, The Phyre 2 web portal for protein modeling, prediction and analysis, Nat. Protoc. 10 (2015) 845–858, https://doi.org/10.1038/nprot.2015.053.
[16] J. Durairaj, A. Di Girolamo, H.J. Bouwmeester, D. de Ridder, J. Beekwilder, A.D. van Dijk, An analysis of characterized plant sesquiterpene synthases, Phytochemistry 158 (2019) 157–165, https://doi.org/10.1016/j.phytochem.2018.10.020.
[17] P.H. Hü, A.E. Mark, W.F. Van Gunsteren, Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations, JMB (J. Mol. Biol.) 4 (1995) 492–503.
[18] B.J. Grant, A.P.C. Rodrigues, K.M. Elsawy, J.A. Mccammon, L.S.D. Caves, Bio: an R package for the comparative analysis of protein structures 22 (2006) 2695–2696, https://doi.org/10.1093/bioinformatics/btl461.
[19] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157 (1982) 105–132, https://doi.org/10.1016/0022-2836(82)90515-0.
[20] C.M. Starks, Structural basis for cyclic terpene biosynthesis by tobacco 5-epi-aristolochene synthase, Science 277 (1997) 1815–1820, https://doi.org/10.1126/science.277.5333.1815, 80.
[21] P.E. O'Maille, A. Malone, N. Dellas, B. Andes Hess, L. Smentek, I. Sheehan, B.T. Greenhagen, J. Chappell, G. Manning, J.P. Noel, Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases, Nat. Chem. Biol. 4 (2008) 617–623, https://doi.org/10.1038/nchembio.113.
[22] R.D. Kersten, J.K. Diedrich, J.R. Yates, J.P. Noel, Mechanism-based post-translational modification and inactivation in terpene synthases, ACS Chem. Biol. 10 (2015) 2501–2511, https://doi.org/10.1021/acschembio.5b00539.
[23] I.I. Abdallah, M. Czepnik, R. Van Merkerk, W.J. Quax, Insights into the three-dimensional structure of amorpha-4,11-diene synthase and probing of plasticity residues, J. Nat. Prod. 79 (2016) 42, https://doi.org/10.1021/acs.jnatprod.6b00236.
[24] B. Lakhani, K.M. Thayer, E. Black, D.L. Beveridge, Spectral analysis of molecular dynamics simulations on PDZ: MD sectors, J. Biomol. Struct. Dyn. 38 (2019) 781–790, https://doi.org/10.1080/07391102.2019.1588169.
[25] M. Seemann, G. Zhai, J.-W. De Kraker, C.M. Paschall, D.W. Christianson, D.E. Cane, Pentalenene synthase. Analysis of active site residues by site-directed mutagenesis, J. Am. Chem. Soc. 124 (2002) 7681–7689, https://doi.org/10.1021/ja026058q.
[26] P. Baer, P. Rabe, C.A. Citron, C.C. de Oliveira Mann, N. Kaufmann, M. Groll, J.S. Dickschat, Hedycaryol synthase in complex with nerolidol reveals terpene cyclase mechanism, Chembiochem 15 (2014) 213–216, https://doi.org/10.1002/cbic.201300708.
[27] R. Li, W.K.W. Chou, J.A. Himmelberger, K.M. Litwin, G.G. Harris, D.E. Cane, D.W. Christianson, Reprogramming the chemodiversity of terpenoid cyclization by remolding the active site contour of epi-isozizaene synthase, Biochemistry 53 (2014) 1155–1168, https://doi.org/10.1021/bi401643u.
[28] W.K.W. Chou, J.A. Himmelberger, K.M. Litwin, G.G. Harris, D.E. Cane, D.W. Christianson, Substitution of aromatic residues with polar residues in the active site pocket of epi-isozizaene synthase leads to the generation of new cyclic sesquiterpenes, Biochemistry 56 (2014) 5798–5811, https://doi.org/10.1021/acs.biochem.7b00895.