UNIVERSITY OF NAIROBI

FACULTY OF SCIENCE & TECHNOLOGY

CHARACTERIZATION OF HEXACHLOROCYCLOHEXANE (HCH) DEGRADATION PATHWAY GENES IN TWO STRAINS OF *SPHINGOBIUM* BACTERIA PREVIOSULY ISOLATED FROM HCH-CONTAMINATED SOIL IN KITENGELA

BY

NAMAKWA PHERIS (BSc. Biochemistry, University of Nairobi)

Reg. No.: H56/8629/2017

A thesis submitted in partial fulfillment of the requirements for the award of the degree of Master of Science in Biochemistry of the University of Nairobi

MAY, 2022

DECLARATION

This thesis is my original work and has not been presented for a degree award in any other University.

Pheris Namakwa

Signature…………………………………………… Date…………………11/05/2022…………………

This thesis has been submitted for examination with our approval as University supervisors.

Dr. Edward K. Muge

Department of Biochemistry

University of Nairobi

Signature……………………………………………… Date…………………11/05/2022…………………

Dr. Evan E. Nyaboga

Department of Biochemistry

University of Nairobi

Signature…………………………………………….. Date…………………11/05/2022………………....

DEDICATION

To Mrs. Mary Namakwa and family at large

ACKNOWLEDGEMENTS

First and foremost, I express my utmost gratitude to God for His loving kindness, mercy, and grace. And for His wisdom, strength, and ability that works in me mightily, to propel me in the path of progress, absolute victory, and unending success. Secondly, I am extremely grateful to my supervisors, Dr. Edward Muge and Dr. Evans Nyaboga for your invaluable suggestions, guidance, and constructive criticism. Special thanks to Nelson Khan for allowing me to be a part of his research study. Thank you for the time, energy, and resources that you spent on my research and for mentorship in the various facets of the research study.

Many thanks to the technical and teaching staff in the Department of Biochemistry, especially, Mr. Thomas Arani, Mr. George Okumu, and Mrs. Zelipha Kabubii. To all my friends and colleagues, thank you for your moral support and encouragement.

Last but not least, I thank my family for being the source of inspiration and motivation thus far. Most importantly, I am sincerely grateful to Geoffrey, Alice, Boniface, Eva, and Simon, who’ve always been of immense help. Thank you for your labor of love, care, and support. God bless you and graciously reward you!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

α-HCH	alpha-Hexachlorocyclohexane
β-HCH	beta-HCH
γ-HCH	gamma-HCH
γ-HMSA	gamma-hydroxymuconic semi aldehyde
δ-HCH	delta-HCH
ε-HCH	epsilon-HCH
ζ-HCH	zeta-HCH
η-HCH	eta-HCH
θ-HCH	theta-HCH
t-HCH	technical HCH
BCIP	5-bromo-4-chloro-3-indoyl phosphate
CHQ	Chlorohydroquinone
Chr 1	Chromosome 1
Chr 2	Chromosome 2
CMA	Carboxymethyl acetate
CoA	Coenzyme A
DCCH	Dichlorocyclohexane
DCHQ	Dichlorohydroquinone
DCP	Dichlorophenol
DNA	Deoxyribonucleic acid
GABA	Gamma amino butyric acid
GC-MS	Gas Chromatography-Mass spectrophotometry
GSH	Reduced glutathione
GSSG	Oxidized glutathione
GSTs	Glutathione *S*-transferases
HCl	Hydrogen chloride
H_2O	Water

HQ	Hydroquinone
MA	Maleylacetate
Mb	Megabase
MCB	Monochlorobenzene
MCH	Monochlorocyclohexane
NAD^+	Nicotinamide adenine dinucleotide (oxidized)
NADH	Nicotinamide adenine dinucleotide (reduced)
NBT	Nitro blue tetrazolium
PCCH	Pentachlorocyclohexane
PCDD	Polychlorinated dibenzo-*p*-dioxins
PCDF	Polychlorinated dibenzofurans
PCHL	Pentachlorocyclohexanol
RMSD	Root-mean-square-deviation
TCCH	Tetrachlorocyclo-1-hexene
TCDL	Tetrachlorocyclohexanediol
2,3,5 TriCDL	2,3,5-trichloro-5-cyclohexene-1,4-diol

ABSTRACT

Hexachlorocyclohexane (HCH) continues to pose threat to the environment despite the restricted use or complete ban in most parts of the world due to its toxicity effects, environmental persistence, and bioaccumulation within the food chain. The extensive use of lindane (99% pure γ-HCH isomer) in agriculture has resulted in contamination of soil and water environments on a global scale. Microbes, particularly sphingomonads, can degrade HCH residues into non-toxic and environmentally safe metabolites. A variety of enzymes participate in the lindane degradation pathway, including dehydrochlorinase (LinA), dehalogenase (LinB), dehydrogenase (LinC & LinX), dechlorinase (LinD), dioxygenase (LinE), and transcriptional regulator (LinR). To develop efficient technologies for sustainable bioremediation of lindane, information on the organization and diversity of *Lin* genes among sphingomonads with the potential to degrade lindane is a prerequisite. Therefore, this study aimed to characterize *Lin* genes involved in the degradation of Hexachlorocyclohexane (HCH) in two strains of *Sphingobium* bacteria (*Sphingobium* sp. S6 and *Sphingobium* sp. S8). DNA was extracted from broth cultures of *Sphingobium* strains S6 and S8, and DNA hybridization was carried out to detect the *Lin* genes in the two strains using Digoxigenin (DIG)-labeled DNA probes synthesized by PCR. The *Lin* genes detected were amplified by PCR using their respective primers and purified PCR products sequenced by the Sanger sequencing method. The resulting DNA sequences were analysed by homology and phylogenetic analysis. Proteins of the respective *Lin* genes were modeled via homology modeling to predict their 3D structure and active sites. Both *Sphingobium* strains S6 and S8 were found to contain *LinA*, *LinB*, *LinC*, *LinD*, *LinE*, *LinR,* and *LinX* gene and IS*6100*, which were conserved. Single copies of *LinA*, *LinB*, and *LinD* gene, two copies of *LinC*, *LinE*, and *LinX* gene, and multiple copies of *LinR* gene and IS*6100* occurred within the genome of *Sphingobium* strain S6. Moreover, the DNA sequences of *LinA* to *LinX* from *Sphingobium* sp. S6 produced full-length polypeptides (LinA, LinB, LinC, LinD, LinE, LinR, and LinX) containing 156, 296, 250, 346, 321, 301, and 250 amino acid residues, respectively, whereas those of *Sphingobium* sp. S8 contained 150, 291, 250, 344, 317, 293, and 250 amino acids, respectively. The predicted protein models of LinA, LinB, LinC, LinD, and LinE comprised of one to four chains. Furthermore, the active binding site of LinA contained three conserved catalytic residues (Lys20, Asp25, and His73), that of LinB possessed a quintet consisting of the nucleophile−Asp108, catalytic acid−Glu132, and catalytic base−His272 and the halide stabilizing residues Asn38 and Trp109. Putative residues (Trp109, Val134, Phe143, Pro144, Gln146, Asp147, Phe151, Phe169, Val173, Leu177, Trp207, Pro208, Ile211, Ala247, Leu248, and Phe273) surrounding the active site of LinB were also conserved. The protein models of LinC, LinD, and LinE contained between two to four active binding sites whose catalytic residues have not been elucidated yet. This study demonstrated that *Sphingobium* sp. S6 and *Sphingobium* sp. S8 possess *Lin* genes with their copy numbers ranging from one to two to multiple copies and are highly conserved. Based on the presence of catalytic sites, information on the substrate-binding

properties of lindane by the various Lin proteins need to be elucidated, which could form the basis for designing enzyme mutants with improved lindane degradation capabilities. These can then be applied in the enzymatic bioremediation of HCH stockpiles and liquid contamination.

CHAPTER ONE

1.0 INTRODUCTION

1.1 Background of the study

Hexachlorocyclohexane (HCH) is a polychlorinated hydrocarbon that includes a benzene ring wherein hydrogen and chlorine are connected to every carbon (Kumari *et al*., 2002). It was widely used in agriculture and for pharmaceutical purposes under the commercial name "Lindane" (Vijgen, 2006; Vijgen *et al*., 2006). Lindane was widely used as an insecticide for the treatment of seeds, wood and timber, spraying of crops, control of ectoparasites in cattle and other farm animals (Girish & Kunhi, 2013). It is also used for topical treatment of head lice and control of scabies (Humphreys *et al*., 2008). Currently, all of the agricultural uses of lindane have been abrogated under the Stockholm Convention because of its toxicity to non-target living beings, inclusive of humans, wildlife, and invertebrates (Madadi *et al*., 2017; Vijgen *et al*., 2011; Somvanshi *et al*., 2008).

However, lindane is still being fabricated as cheap but effective insecticide in a few developing nations such as India, mainly for financial reasons (Cao *et al*., 2013). In addition, HCH residues are continually being detected in air, water, soils, and sediment in various parts of Kenya and beyond (Madadi *et al*., 2017). Two formulations of HCH exist namely; technical HCH (*t*-HCH) and lindane (γ-HCH) (Vijgen, 2006). Technical HCH consists of eight different isomers (**Fig. 1**), five (α-, β-, δ-, γ-, & ε-HCH) of which are stable and constitute the major components of the technical mixture (Kumari *et al*., 2002; Manickam *et al*., 2008; Okai *et al*., 2010). All HCH isomers differ in their stereochemistry and have high hydrophobicity and persistence, and are widely spread in the environment (Cao *et al*., 2013). Lindane is the 99% pure γ-HCH isomer and is typically purified from the technical combination but the rest of the isomers are commonly discarded (Humphreys *et al*., 2008).

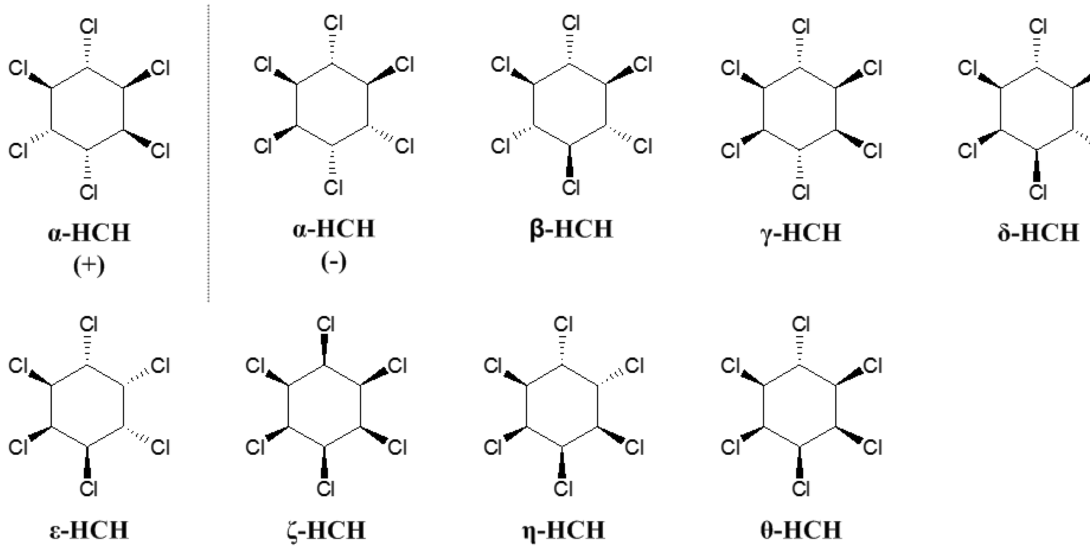


Figure 1. Positions of chlorine atoms around the cyclohexane ring in the different isomers of HCH. Adapted from Zdravkovski (2004).

Lindane was named after the Dutch Researcher, "Dr. Teunis van der Linden", who found its insecticidal property in 1912 (Girish & Kunhi, 2013). Lindane began to be produced commercially in 1945 and by reaction of chlorine and benzene in the presence of ultra-violet (UV) light to form eight isomers; the five most stable isomers occurring in the proportions of α- (~65%), β- (~8.5%), γ- (~11%), δ- (~8%) and ε-HCH (~4%) while the remaining isomers (ζ-, η-, and θ-HCH) occur in only small amounts (Girish & Kunhi, 2013). The mixture of isomers (known as technical HCH) undergoes several steps of purification and concentration to produce the 99% pure γ-HCH isomer (lindane) which accounts for only 10–15% of the total yield (Girish & Kunhi, 2013; Nayyar & Lal, 2016). Lindane manufacturing is hence inefficient due to the fact for each ton of γ-HCH isomer produced, about 8 to 12 tons of HCH residuals comprising α-, β-, δ- and ε-HCH are obtained (Girish & Kunhi, 2013; Lal *et al*., 2006; Vijgen, 2006). An estimated four to six million tons of HCH waste on account of the manufacturing of approximately 60,000 tons of lindane are stated to occur (Nayyar & Lal, 2016). Moreover, HCH waste isomers are regularly dumped inappropriately in lots of places across the world, and attempts to recycle those waste isomers (in particular β- and δ-HCH) have led to incredibly contaminated waste streams (Vijgen *et al*., 2006).

In addition to inefficient production, the purification system is likewise high priced and as a result of financial reasons, many countries opted for using technical HCH as opposed to lindane and in an indiscriminate manner (Kumari *et al*., 2002). Large-scale manufacturing of lindane coupled with the indiscriminate use of technical HCH considerably over the last 60 years has consequently created a severe problem of environmental contamination worldwide (Manickam *et al*., 2007). The greatest concern has been the open stockpiles of HCH waste, lack of proper management and disposal of HCH waste, and continual shifting of HCH waste from dumpsites (Nayyar & Lal, 2016). These sites act as reservoirs from whence HCH travels to remote locations via volatilization and transport by air, subsequently contaminating new areas and continually being a worldwide issue for several decades (Cao *et al*., 2013).

Following their listing as Persistent Organic Pollutants (POPs) under the Stockholm Convention of 2009, HCH isomers continued to receive numerous scrutiny in most developed countries (Nayyar & Lal, 2016). Later, a ban was imposed on the use of lindane and *t*-HCH in several countries while they were severely restricted in others. However, neither the ban on manufacturing nor the restriction of HCH use has decreased residues levels or stopped their access into the environment (Böltner *et al*., 2005; Kumari *et al*., 2002; Nayyar & Lal, 2016). In Kenya, HCH residues have been detected in samples of soil and air from regions of Nairobi and Mount Kenya (Aucha *et al*., 2017), sediment and water from Rusinga Island in Lake Victoria (Osoro *et al*., 2016), and in agricultural soils in Meru County (G *et al*., 2019). The problem of contamination has further been compounded by biotransformation and biomagnification of HCH residues through the food chain (Dogra *et al*., 2004; Nayyar & Lal, 2016).

Microorganisms with the capacity to degrade HCH have attracted interest due to the fact they may be used for *in situ* detoxification and bioremediation (Girish & Kunhi, 2013). Microbial degradation of lindane is observed to occur in aerobic as well as anaerobic ecosystems (Girish & Kunhi, 2013) however entire mineralization is only aerobic (Nagata *et al*., 2007). *Clostridium* species (including *C. sphenoides* and *C. rectum*) and some members of *Enterobacteriaceae* and *Bacillaceae* reportedly degrade γ-HCH anaerobically (Girish & Kunhi, 2013). *Sphingomonas/Sphingobium* species and a few white-rot fungi, including *Trametes hirsutus*, *Cyathus bulleri*, *Phanerochaete sordida* as well as *Phanerochaete chrysosporium* degrade γ-HCH aerobically (Fuentes *et al*., 2010). A plethora of HCH-degrading microorganisms exist but only members of the *Sphingomonadaceae* family play a significant role in the total mineralization of HCH. Three separate strains of *Sphingobium* namely; *S. japonicum* UT26 (Imai *et al*., 1989), *S. indicum* B90A (Sahu *et al*., 1990), and *S. francense* Sp+ (Dogra *et al*., 2004) previously isolated from soils contaminated with HCH in Japan, India, and France, respectively, are the most widely studied (Lal *et al*., 2006). Besides these, twelve other species of sphingomonads were later isolated from HCH-contaminated soils at Bilbao and Chemnitz in Spain and Germany, respectively (Mohn *et al*., 2006; Böltner *et al*., 2005).

Despite this significant number of HCH-degrading isolates, there is limited information regarding HCH-degrading microbes isolated from Africa and particularly, in Kenya. In addition, though the Lin pathway has been elucidated in several sphingomonads plus a number of other bacteria confined to HCH-contaminated locales, they are known to be highly variable in their organization and the coding sequences. Likewise, the extent, as well as biological relevance of these variations, especially of the key upstream genes (*LinA* and *LinB*), is not well understood. Available evidence additionally shows that there are high degrees of polymorphisms within the protein sequences of each LinA and LinB, the most important enzymes of the upstream pathway (Hu). To address these gaps, this study centered on the characterization of *Lin* genes in *Sphingobium* strains S6 and S8 isolated in Kenya by investigating their presence, copy numbers, and sequence variations, and modeling the three-dimensional (3D) structures of key Lin pathway enzymes. The study of *Lin* genes in the *Sphingobium* strains S6 & S8 isolated from Kenya would generate new information on the genetic variability and distribution of *Lin* genes amongst sphingomonads. Thus, increasing the number of HCH-degrading microbes and growing the pool of HCH-degrading genes (or enzymes) which can then be utilized for developing sustainable bioremediation technologies.

1.2 Problem statement

The use of HCH and other organochlorine pesticides (OCPs) became banned in the 1970s and 1990s in most developed countries and severely restricted in others due to their persistence in the environment, toxicity, and bioaccumulation in the food chain (Girish & Kunhi, 2013). Despite the ban or restricted use, HCH isomers remain a serious health and environmental threat, particularly in developing countries. This is especially due to high levels of contamination, accumulation of stockpiles following the ban, and illegal disposal of HCH waste due to remiss environmental laws in a number of nations (Lal *et al*., 2008; Nayyar & Lal, 2016). In addition, lindane is a neurotoxin that is known to interfere with gamma-aminobutyric (GABA) neurotransmitter function and is possibly carcinogenic (Girish & Kunhi, 2013). Considerable exposure can cause serious health problems, such as neurological derangements, seizures, convulsions, and different formative toxicities, and indeed death (Cao *et al*., 2013). Moreover, fish, mammals, commodities of food, and human blood samples and fat tissue have been found to contain HCH residues. Residues of HCH isomers have been detected in samples of soil, water, sediment, and air from Nairobi's Industrial area, Dandora and Kabete dumpsites, Mt. Kenya and Lake Victoria regions (Aucha *et al*., 2017; G *et al*., 2019; Osoro *et al*., 2016), showing the presence and extent of spread of HCH contamination in Kenya. Therefore, there is need to study *Lin* genes in *Sphingobium/Sphingomonas* species isolated from HCH-contaminated soil in Kenya to generate new information on the genetic variability and distribution of *Lin* genes amongst sphingomonads as well as the substrate binding properties of the enzymes encoded by these *Lin* genes. This information would then contribute to future development of sustainable bioremediation strategies that can be used to mitigate HCH contamination problems.

1.3 Justification of the study

The use of microorganisms for remediation of contaminated sites is considered to be an efficient and cost-effective approach because it is eco-friendly and poses minimum health risks (Pant *et al*., 2019). A significant number of HCH-degrading microbes have been discovered and isolated from HCH-contaminated soils in various parts of the world but there is limited information regarding HCH degraders isolated from Africa, particularly in Kenya. HCH-degrading microbes possess *Lin* genes which play a critical role in the pesticide (lindane) degradation pathway by converting very toxic metabolites to less toxic or non-toxic ones, and hence reduce the impact of HCH contamination in the environment. In this study, *Lin* genes were characterized using pure cultures of *Sphingobium* sp. S6 and *Sphingobium* sp. S8 isolated from Kenya to investigate their variability and distribution and the substrate binding properties of the various enzymes encoded by these genes. The study of genetic variability of *Lin* genes of would increase the pool of HCH-degrading microbes and genes (or enzymes) and facilitate know-how of the extent and biological significance of these variations. Moreover, the identification of catalytic residues in the active sites of key enzymes through comparative modeling of their 3D structures would provide an understanding of the Biochemistry of Lin enzymes. Through modern research technologies, this information can then be used to develop new strains with improved growth and viability properties and enzymes with a broader substrate range and better kinetics for enzymatic bioremediation of stockpiles and liquid contamination.

1.3 Objectives

1.3.1 Main objective

To characterize *Lin* genes in Hexachlorocyclohexane (HCH) degradation pathway of *Sphingobium* bacteria (*Sphingobium* sp. S6 & *Sphingobium* sp. S8) isolated from HCH-contaminated soil in Kitengela and study the substrate binding pockets of key enzymes.

1.3.2 Specific objectives

i) To investigate the presence and copy numbers of *LinA*, *LinB*, *LinC*, *LinD*, *LinE*, *LinR*, *LinX*, and IS*6100* in *Sphingobium* sp. S6 and *Sphingobium* sp. S8

ii) To evaluate the genetic variability of *LinA*, *LinB*, *LinC*, *LinD*, *LinE*, *LinR*, and *LinX* in *Sphingobium* sp. S6 and *Sphingobium* sp. S8

iii) To model the three-dimensional (3D) structure of key Lin pathway enzymes in the HCH degradation pathway

1.4 Research questions

i) Do *Sphingobium* sp. S6 and S8 harbor *Lin* genes with the same copy number?

ii) How do *Lin* genes of *Sphingobium* sp. S6 and *Sphingobium* sp. S8 compare with those from other HCH-degrading bacteria?

iii) How does sequence variation among Lin enzymes affect their active 3D structure?

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 Estimates of global lindane use, storage, and deposition

Lindane was largely utilized in agriculture as an insecticide for the treatment of seeds, lumber and timber, spraying of crops, and against ectoparasites in different farm animals including cattle (Girish & Kunhi, 2013). Lindane was also used for topical treatment of head lice and control of scabies (Humphreys *et al*., 2008) and public health control against malaria (Girish & Kunhi, 2013). However, using lindane was later deserted in most developed nations and seriously restrained in others due to its toxicity consequences on non-target organisms, such as humans (Somvanshi *et al.,* 2008). Lindane is likewise a neurotoxin regarded to intervene with Gamma-aminobutyric acid (GABA) neurotransmitter function and is a probable carcinogen (Girish & Kunhi, 2013).

About 600,000 metric tons of γ-HCH (lindane) were used worldwide in the period ranging from 1950−2010 (Nayyar & Lal, 2016). An estimated 4.8 million tons of HCH waste was stored as stockpiles and deposited indiscriminately in many locations around the globe (Vijgen *et al*., 2006). Stockpiles of HCH residues (called "white mountains", "HCH muck" or scum) were often left uncovered resulting in contamination of adjoining areas because of the semi-volatile nature of HCH residues. Furthermore, the residuals from the production processes were originally considered to be harmless and water insoluble and were therefore used for construction purposes. These actions in addition compounded the problem of contamination and attempts to recycle the stockpiles, on the contrary, led to contamination of waste streams with exceedingly polychlorinated compounds (Vijgen, 2006; Vijgen *et al*., 2006).

2.2 HCH residues levels in Kenya

In Kenya, the latest look at the levels of OCP (organochlorine pesticide) residues in sediment and water samples from Rusinga Island by Osoro *et al*. (2016) confirmed excessive levels of α-HCH (7.023±0.01µg/L of water and 22.624±3.23µg/Kg of sediment) followed by β-HCH (2.96±0.97µg/L of water and 21.94±4.21µg/Kg of sediment) and the least concentration was reported with γ-HCH (0.52±0.01µg/L water and 6.23±1.95µg/Kg sediment). Moreover, a high occurrence of α-, β-, and γ-HCH residues have been said to occur in air and soil at Nairobi's Industrial area, Dandora, and Kabete dump sites, and Mt. Kenya regions, with β-HCH being the most abundant and α-HCH the least abundant of the new POPs analyzed (Aucha *et al*., 2017). HCH residues were also detected in soil from Imenti South and Imenti North Sub Counties in Meru County. The mean HCH isomer concentrations ranged between Below Detectable Limit (BDL) to ~60 µg/Kg, ~50 µg/Kg, ~40 µg/Kg, and ~0.60 µg/Kg for α-, β-, γ-, and δ-HCH in North Imenti. In South Imenti, on the other hand, the mean HCH isomer concentrations ranged between BDL to ~2 µg/Kg, ~15 µg/Kg, ~3 µg/Kg, and ~0.5 µg/Kg for α-, β-, γ-, and δ-HCH (G *et al*., 2019).

2.3 Microbial HCH degradation

Microbial degradation of HCH isomers was considerably studied in the laboratory and pilot and/or *in situ* field settings (Phillips *et al.,* 2005). These studies involved the use of pure liquid cultures of bacteria (*Clostridium rectum* and *Pandoraea* spp.) and native soil microbes, white-rot fungi (*Phanerochaete chrysosporium*), and sewage sludge, in both anaerobic and aerobic conditions (Quintero *et al.,* 2005). Substrate concentration, temperature, external carbon source, soil inhomogeneity, and sorption and solubilization characteristics, according to Bachmann *et al*. (1988) are the necessary conditions for HCH biodegradation to occur and the rate of bioconversion decreases following the order; γ-, α-, δ-, and β-HCH.

Chemical properties which include polarity, solubility, volatility, and sorption characteristics affect the biodegradability and transport and therefore persistence of the isomers. Furthermore, how chlorine atoms are organized and placed across the cyclohexane ring additionally determines the relative persistence and the biodegradation rate of every HCH isomer (Phillips *et al*., 2005). Isomerization, on the other hand, alters the relative stability and persistence thereby influencing the success of bioremediation (Phillips *et al*., 2005). According to Girish & Kunhi (2013), both living and non-living factors determine the outcome of HCH in the environment and how efficiently it isdegraded by microbes. A diverse number of HCH-degrading microorganisms have been discovered (Böltner *et al*., 2005) and HCH degradation has been shown to occur in both aerobic and anaerobic conditions (Lal *et al.,* 2010).

2.3.1 Anaerobic degradation of HCH

HCH degradation became first of all pronounced in anaerobic conditions in soil microcosms, soil slurries, field studies in flooded soils, pure cultures, and groundwater plumes (Phillips *et al.,* 2005; Lal *et al.,* 2010). Mineralization of α-HCH, β-HCH, and γ-HCH observed in glass columns containing methanogens further provided evidence of anaerobic degradation of HCH (Phillips *et al*., 2005). Anaerobic biodegradation of γ-HCH produces chlorobenzene (**Fig. 2**) and proceeds through two dichloroeliminations successively forming two products, γ-TCCH (γ-3,4,5,6-tetrachloro-1-cyclohexene) and 5,6-dichlorocyclohexa-1,2-diene, followed by dehydrochlorination to form chlorobenzene (Lal *et al*., 2010). Two intermediates, γ-TCCH, and MCH (Monochlorocyclohexane), as well as small amounts of tri- and tetra-chlorobenzene (TCBs), are formed (Lal *et al*., 2010).

Similarly, α-HCH, β-HCH, and δ-HCH are mineralized additionally via successive dichloroeliminations accompanied by dehydrochlorination to produce chlorobenzene (Lal *et al*., 2010). Although α- and δ-HCH degradation produces chlorobenzene, that of α-HCH takes place via δ-TCCH (δ-3,4,5,6-tetrachloro-1-cyclohexane) while that of β-HCH (**Fig. 2**) takes place through successive dichloroeliminations via δ-TCCH to form dichlorocyclohexadiene, some of which is further broken down by dehydrochlorination to chlorobenzene (Doesburg *et al*., 2005).

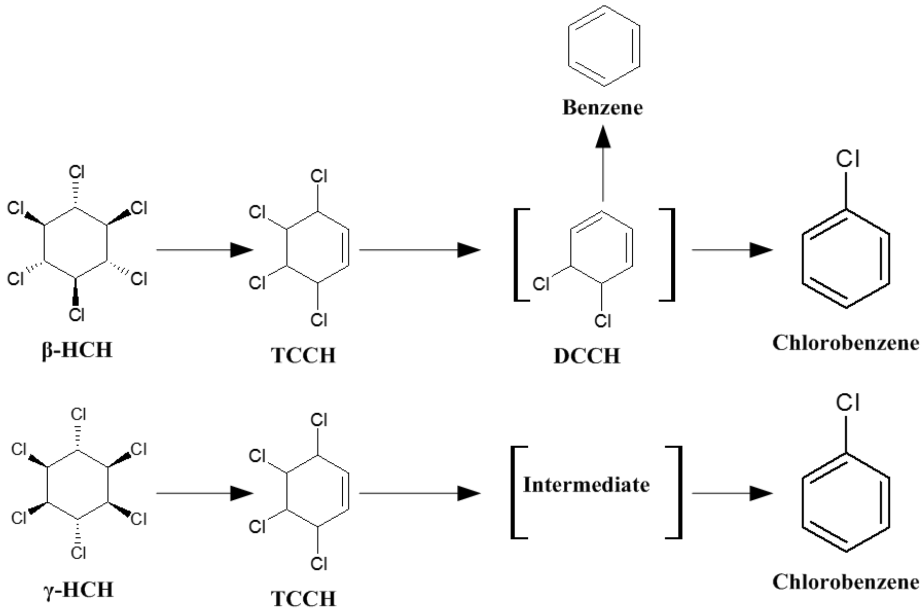


Figure 2. Anaerobic γ-HCH and β-HCH degradation pathway. Shown in square brackets are compounds proposed to occur but not yet determined empirically. Adapted from Lal *et al.* (2010).

2.3.2 Aerobic degradation of HCH

Mineralization of HCH aerobically was first reported in *Escherichia coli* and *Pseudomonas* strains (Girish & Kunhi, 2013). Sphingomonads are the most common aerobes known to date and about 30 species occur (Lal *et al*., 2010). Three of these are separate strains of *Sphingobium* namely; *S. japonicum* UT26 (Imai *et al*., 1989), *S. indicum* B90A (Sahu *et al*., 1990), and *S. francense* Sp+ (Dogra *et al*., 2004) previously isolated from soils contaminated with HCH in Japan, India, and France, respectively (Lal *et al*., 2006).

2.3.2.1 Degradation of γ-HCH

γ-HCH is completely mineralized aerobically via six steps of dechlorination (Nagata *et al.,* 2007). A detailed pathway has been elucidated in UT26 (**Fig. 3**) wherein γ-HCH is transformed to γ-1,3,4,6-tetrachloro-1,4-cyclohexadiene (1,4-TCDN) through γ-pentachlorocyclohexene (γ-PCCH) by two dehydrochlorination reactions (Lal *et al.,* 2010; Okai *et al.,* 2010; Nagata *et al.,* 2007). 1,4-TCDN is then metabolized via two steps of hydrolytic dechlorination in the second reaction to form 2,4,5-trichloro-2,5-cyclohexadiene-1-ol (2,4,5-DNOL) and 2,5-dichloro-2,5-cyclohexadiene-1,4-diol (2,5-DDOL). 2,5-dichlorohydroquinone (2,5-DCHQ) is then generated by dehydrogenation of 2,5-DDOL, thus completing the upstream HCH degradation pathway (Lal *et al*., 2010). Two minor products generated autonomously by dehydrochlorination of 1,4-TCDN and 2,4,5-DNOL, i.e., 1,2,4- trichlorobenzene (1,2,4-TCB) and 2,5-dichlorophenol (2,5-DCP), respectively, are dead-end products in strain UT26 (Lal *et al.,* 2010; Endo *et al.,* 2005).

In the downstream pathway, reductive dechlorination of 2,5-DCHQ produces chlorohydroquinone (CHQ) and through the second step of reductive dechlorination, hydroquinone (HQ) is formed. HQ so formed undergoes cleavage to form γ-hydroxymuconic semi-aldehyde (γ-HMSA) (Lal *et al*., 2010). The transformation of CHQ by direct ring cleavage to an acyl chloride, and subsequently to MA (maleylacetate) is the major route (Lal *et al*., 2010). Maleylacetate (MA) forms β-ketoadipate, which is cleaved to succinyl-CoA and acetyl-CoA, which then enter the citric acid cycle and are additionally broken down to release CO_2 and water (Girish & Kunhi, 2013; Nagata *et al*., 2007; Endo *et al*., 2005).

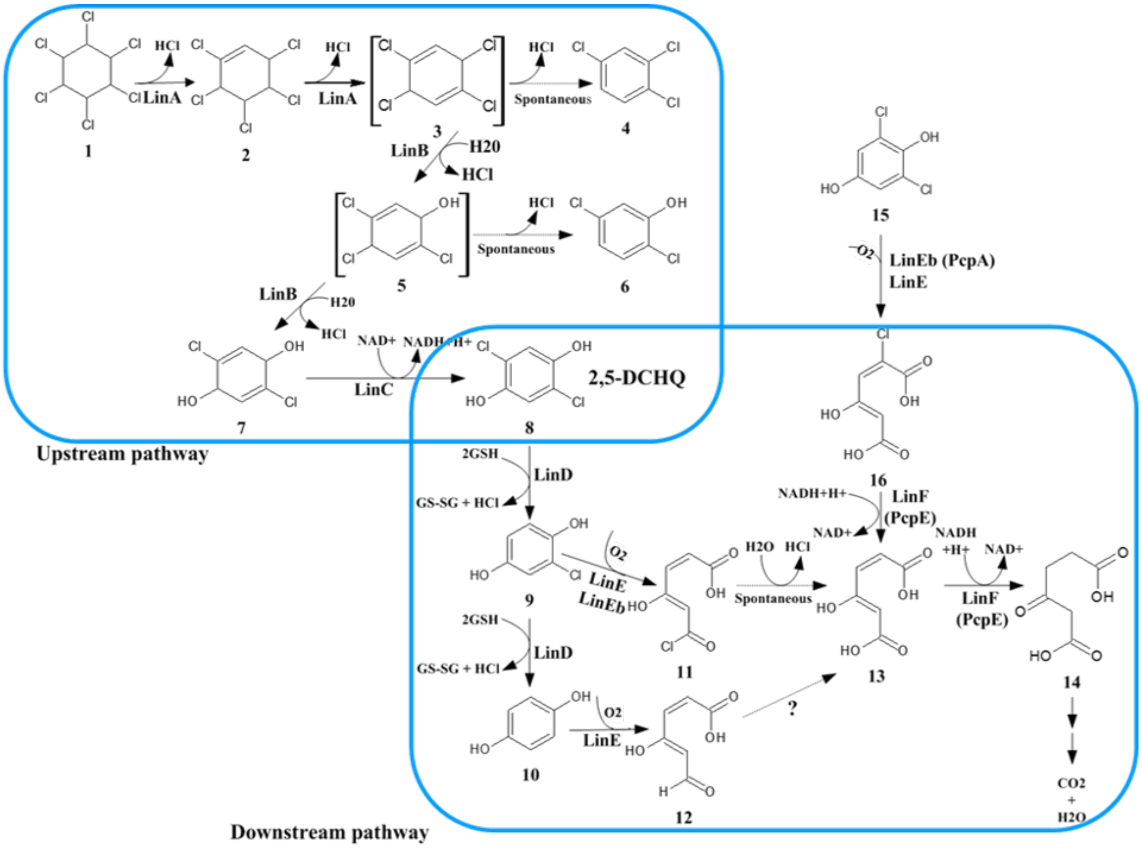


Figure 3. Proposed pathway of γ-HCH degradation in *S. japonicum* UT26
Compounds: **1 =** γ-Hexachlorocyclohexane (γ-HCH), **2 =** γ-pentachlorocyclohexane (γ-PCCH), **3** = 1,3,4,6-tetrachloro-1,4-cyclohexadiene (1,4-TCDN), **4 =** 1,2,4-trichlorobenzene (1,2,4-TCB), **5** = 2,4,5-trichloro-2,5-cyclohexadiene-1-ol (2,4,5-DNOL), **6** = 2,5-dichlorophenol (2,5-DCP), **7** = 2,5-dichloro-2,5-cyclohexadiene-1,4-diol (2,5-DDOL), **8** = 2,5-dichlorohydroquinone (2,5-DCHQ), **9 =** chlorohydroquinone (CHQ), **10** = hydroquinone (HQ), **11 =** acylchloride, **12** = γ-hydroxymuconic semialdehyde (γ-HMSA), **13** = maleylacetate (MA), **14** = β-ketoadipate (3-oxoadipate), **15** = 2,6-dichlorohydroquinone (2,6-DCHQ), and **16** = 2-chloromaleylacetate (2-CMA); GSH, reduced form of glutathione; GSSG, oxidized form of glutathione. Compounds shown in square brackets are those that are unstable. Adapted from Endo *et al*. (2005).

2.3.2.2 Degradation of α-HCH

Mineralization of α-HCH is exemplified in strain B90A (**Fig. 4**) and dehydrochlorination also occurs as with γ-HCH in UT26 (Lal *et al*., 2010). Since α-HCH has two enantiomers, two β-PCCH products; β-1,3,4,5,6-PCCH corresponding to (+)-α-HCH, and β-1,3,4,5,6-PCCH corresponding to (-)-α-HCH, are formed. These two products are suggested to be further broken down to form the dead-end product, 1,2,4-TCB in both UT26 and B90A. However, there is no empirical proof by any means to verify this assertion, and the route through TCDN is simply through inference from mineralization of δ- and γ-HCH (Lal *et al*., 2010).

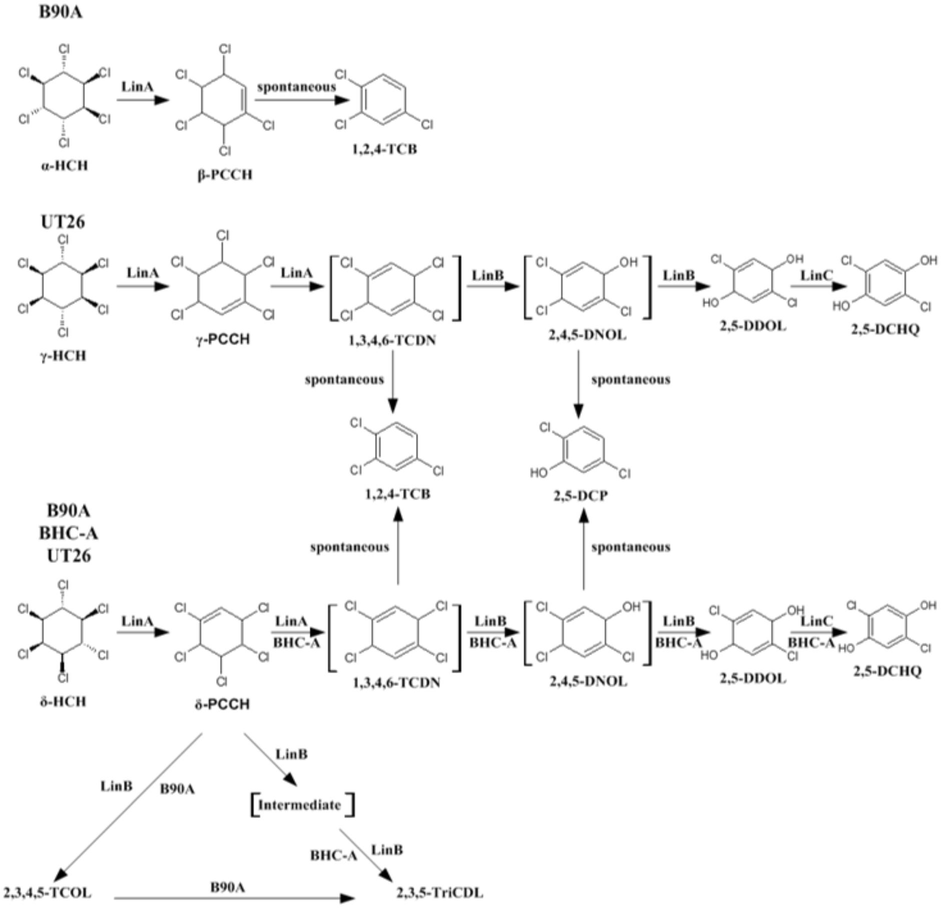


Figure 4. Aerobic degradation of α-, γ-, and δ-HCH in the upstream pathway in three separate species of *Sphingobium*; UT26, B90A, and BHC-A. Adapted from Lal *et al*. (2010).

2.3.2.3 Degradation of β-HCH

β-HCH is the most persistent and is not always absolutely mineralized aerobically (Lal *et al.,* 2010; Doesburg *et al.,* 2005). It isrelatively stable due to the equatorial chlorine atoms which might be a barrier to dehydrochlorination reactions that require axial chlorine atoms. Hydrolytic dechlorination reactions are more probable and in two strains (Sp+ and

UT26), among the five that were tested, 2,3,4,5,6-pentachlorocyclohexanol (PCHL) was the end-product formed in the transformation of β-HCH. In other strains, PCHL was further broken down to 2,3,4,5,6-tetrachlorocyclohexanediol (2,3,4,5,6-TCDL), showing that differences within and between strains account for the observed differences in the mineralization of β-HCH in comparison with α-HCH and δ-HCH (Lal *et al*., 2010).

2.3.2.4 Degradation of δ-HCH

There are two possible routes for δ-HCH degradation in the upstream pathway (**Fig. 4**); a dehydrochlorinase-led route in BHC-A, similar to that of γ-HCH in UT26 and hydrolytic dechlorination-led route, present in these strains and similar to that of β-HCH described above (Lal *et al*., 2010). Mineralization of δ-HCH to PCHL occurs via the hydrolytic dechlorination-led route in UT26 and Sp+ but PCHL is not transformed further in these strains. However, in B90A and BHC-A, PCHL so formed is further transformed to TCDL. The dehydrochlorinase-led route on the other hand converts δ-HCH to δ-PCCH in BHC-A and B90A, and δ-PCCH is then transformed to 1,3,4,6-TCDN as for γ-HCH. 1,3,4,6-TCDN then follows the same route used by γ-HCH in UT26. But in BHC-A (as well as B90A), two hydrolytic dechlorination reactions successively convert δ-PCCH to form 2,3,5-TriCDL via 2,3,4,5-TCOL (Lal *et al.,* 2010).

2.4 The *Lin* genes

The catabolic genes associated with γ-HCH degradation (termed "*Lin*" genes) first were identified in UT26, and later in *Rhodanobacter lindaniclasticus* (Lal *et al*., 2006; Pal *et al*., 2005). Such genes have also been found in other sphingomonad species including; B90(A) and Sp+ from India and France, respectively (Girish & Kunhi, 2013; Lal *et al.,* 2010; Böltner *et al.,* 2005). Apart from γ- HCH, all other isomers lack a functional pathway although *Lin* genes may significantly play a role (Pearce *et al.,* 2015). In spite of the huge number of microorganisms known today, the γ-HCH pathway has been only comprehensively studied in UT26 (Böltner *et al*., 2005).

The γ-HCH pathway in UT26 comprises of eight structural genes (*LinA* to *LinJ*) plus a regulatory gene (*LinR or LinI*) (Lal *et al.,* 2006; Nagata *et al.,* 1999) (**Table 1**). *LinA*, *LinB*, and *LinC* encode enzymes of the upstream pathway namely; HCH dehydrochlorinase (LinA), halidohydrolase (LinB), and dehydrogenase (LinC), respectively whereas *LinD*, *LinE* (*LinEb*), *LinF*, *LinGH*, and *LinJ* encode enzymes of the downstream pathway namely; reductive dechlorinase (LinD), ring cleavage dioxygenase (LinE), maleylacetate reductase (LinF), acyl-CoA transferase (LinGH) and thiolase (LinJ), respectively (Lal *et al.,* 2010; Nagata *et al.,* 1999). The Lys-type transcriptional regulator (LTTR) or *LinR* regulates *LinD* and *LinE* expression and is located upstream of *LinE* (Böltner *et al.,* 2005). *LinD* and *LinE* form an operon together with other reading frames (Nagata *et al*., 1999). A *LinX*-encoded dehydrogenase occurs upstream of *LinA* (Böltner *et al*., 2005) and has similar activity to LinC (Dogra *et al*., 2004). The putative ABC-type transporter encoded by *LinE*, *LinL*, *LinM*, and *LinN* is important in the utilization of γ-HCH by strain UT26 (Nagata *et al.,* 2007).

Table 1. The *Lin* genes in the three separate *Sphingobium* species; B90A, UT26, and Sp+

***Lin* Genes**	**Nucleotides (aa) base pairs (bp)**			**G+C content (%)**			**Function**	**Expression in UT26**
	B90A	**UT26**	**Sp+**	**B90A**	**UT26**	**Sp+**	**B90A, UT26, Sp+**	
LinA1	462 (154)	**	**	52.7	**	**	Dehydrochlorinase	Constitutive
LinA2 /LinA[a]	468 (156)	468 (156)	468 (156)	53.9	53.9	53.9	Dehydrochlorinase	
LinB	888 (296)	888 (296)	888 (296)	62.5	62.5	62.5	Halidohydrolase	Constitutive
LinC	750 (250)	750 (250)	750 (250)	64.5	64.3	64.5	Dehydrogenase	Constitutive
LinD	1038 (346)	1038 (346)	1038 (346)	61.8	61.0	61.8	Reductive dechlorinase	Inducible
LinE	963 (321)	963 (321)	963 (321)	60.1	60.1	60.1	Ring-cleavage dioxygenase	Inducible
LinR	909 (303)	909 (303)	909 (303)	60.3	61.3	60.3	Transcriptional regulator	?
LinX1	750 (250)	750 (250)	750 (250)	64.5	64.5	64.5	Dehydrogenase	?
LinX2	750 (250)	**	**	64.5	**	**	Dehydrogenase	?
LinX3	750 (250)	**	**	64.5	**	**	Dehydrogenase	?
LinF	*	1056 (352)	*	*	68.1	*	Reductase	?
tnpA	792 (264)	792 (264)	792 (264)	61.0	61.0	61.0	Transposase	?

[a] *LinA* refers to the *LinA* gene (UT26 and Sp+). Asterisk (**): not detected; (*): not determined. Adapted from Lal *et al*. (2006) and Nagata *et al*. (1999).

2.4.1 Genomic organization of strain UT26

Three circular replicons (Chr 1 ~3.5Mb; Chr 2 ~682kb, pCHQ1 ~191kb) and two other small plasmids comprise the UT26 genome (Tabata *et al.,* 2011; Nagata *et al.,* 2007). *LinA*, *LinB*, *LinC*, *LinKLMN* occur on Chr 1 whereas *LinF* & *LinGHIJ* on Chr 2, and *LinDER* on pCHQ1 plasmid. Furthermore, *LinA*, *LinB*, *LinC*, *LinDER*, and *LinF* occur in association with IS*6100* on unique DNA regions in the UT26 genome, an indication that they were acquired by horizontal gene transfer (Pearce *et al.,* 2015; Tabata *et al.,* 2011; Nagata *et al.,* 2007). *LinGHIJ* and *LinKLMN* are conserved among sphingomonads, indicating that they perform core functions (Tabata *et al.,* 2011; Nagata *et al.,* 2007). However, *LinA* and *LinB* of Sp+ occur in plasmids thus showing that plasmids also contribute to the spread of most *Lin* genes (Nagata *et al.,* 2007).

2.4.2 Genomic organization of strain B90A

B90A genome comprises four replicons; a 3.6 Mb Chr and three plasmids (pSRL1, pSRL2, and pSRL3) (Verma *et al*., 2017). B90(A) contains two *LinA* copies; *LinA1* and *LinA2*, and both encode a functional LinA when cloned in *Escherichia coli*. Besides, their gene products (LinA1 and LinA2) have a 92% sequence similarity and are identical to LinA of UT26 (88% and 99%, respectively) (Dogra *et al*., 2004). According to Verma *et al.* (2017), *LinA2*, *LinB*, *LinRED*, *LinF*, *LinGHIJ*, and *LinKLMN* occur on the chromosome whereas *LinA1*, *LinC*, and *LinF* occur on plasmid pSRL1 and *LinDER* on plasmid pSRL3. About 26 copies of transposon IS*6100* were observed to occur, 15 of which were associated with *Lin* genes and several gene duplications including *LinA*, *LinDER*, *LinGHIJ*, and *LinF* alongside two other *LinE* variants (*LinEa* and *LinEb*). B90 is a mutant strain of B90A that lacks *LinD*, *LinE*, and *LinR* (Dogra *et al*., 2004).

2.4.3 *Lin* genes and their association with IS*6100*

IS*6100* insertion sequence is an 880bp transposable element belonging to the IS*6* family, previously isolated from *Mycobacterium fortuitum* (Nagata *et al.,* 2007; Lal *et al.,* 2006).

It is a constituent of the composite transposon Tn*610* responsible for sulfonamide resistance by bacteria (Dogra *et al*., 2004). It incorporates the *tpnA* gene encoding a transposase (Tabata *et al*., 2011) that mobilizes genes between two direct repeats (14 bp) of IS*6100* elements via replicon fusion followed by resolution through homologous recombination (Pearce *et al*., 2015). In other words, it transposes by the formation of co-integrate intermediates (Böltner *et al*., 2005). Because IS*6100* has a wide range of host strains, it isfound to occur on the chromosome, plasmids as well as in catabolic transposons. Most *Lin* genes in strains UT26, B90A, and Sp+ occur in association with IS*6100* and the number of copies may vary from strain to strain (Böltner *et al*., 2005; Dogra *et al*., 2004; Verma *et al*., 2017).

2.5 HCH degradation pathway enzymes: Upstream pathway

2.5.1 HCH dehydrochlorinase (LinA)

LinA is a 16.5kDa homo-trimeric, an extracellular secretory protein located within the periplasmic space of sphingomonads and homologous to LinA of *Rhodanobacter lindaniclasticus* (Lal *et al*., 2010). It catalyzes dehydrochlorination of γ-HCH plus γ-PCCH, α-HCH and δ-HCH but not β-HCH (Nagata *et al.,* 2007, 1999). The mechanism of dehydrochlorination by LinA is stereoselective, occurring at *trans* and biaxial hydrogen-chlorine pairs of α-, γ-, & δ-HCH. β-HCH and δ-PCCH, on the contrary, do not have the 1,2-biaxial H-Cl group hence are not acted upon by LinA (Lal *et al.,* 2006; Nagata *et al.,* 1999). Because of its narrow substrate specificity, LinA does not require cofactors, and consequently γ-HCH dehydrochlorination proceeds via 1,2-anti dehydrochlorination (Lal *et al.,* 2010; Nagata *et al.,* 2007, 1999). Only four variants of LinA have been reported to date, and the amino acid differences that exist among them have been attributed to changes at the level of DNA. Thus, the mutations that have been accumulated by LinA variants are non-synonymous (Nagata *et al.,* 2007).

2.5.2 Haloalkane dehalogenase (LinB)

LinB is a 32-kDa monomeric protein in the periplasmic space of sphingomonads and hydrolytically catalyzes the dechlorination of 1,4-TCDN forming 2,5-DDOL (Lal *et al*., 2010). Sequence similarity comparisons between B90A: Sp+; B90A: UT26; and Sp+:UT26 showed a 97%, 97%, and 98% sequence similarity, respectively and LinB of *R. lindaniclasticus* and UT26 were found to be strongly homologous (Lal *et al*., 2006). LinB (or halidohydrolase) belongs to the α/β-hydrolase fold family and has a broad substrate range (Nagata *et al.,* 2007). It issignificantly similar to haloalkane dehalogenase (Dh1A), haloacetate dehalogenase (DehH1), and 2-hydroxymuconic semi-aldehyde hydrolase (DmpD) in *Xanthobacter autotrophicus* GJ10, *Moraxella* species B, and *Pseudomonas* species CF600, respectively (Lal *et al.,* 2006; Nagata *et al.,* 1999).

Besides dehalogenation of γ-HCH in UT26, LinB also catalyzes β-HCH and δ-HCH transformation in B90A, Sp+, and UT26 (Lal *et al*., 2006). The transformation efficiency of LinB however differs from strain to strain. Nevertheless, a catalytic triad composed of H272, E132, and D108 in the same α/β family is thought to catalyze the reaction in conjunction with other residues (particularly N38 and W109) (Lal *et al.,* 2010; Nagata *et al.,* 2007). Amino acid substitutions amongst LinB variants are non-synonymous mutations arising among LinA variants and may be responsible for the observed variation in activity towards β-HCH (Nagata *et al.,* 2007).

2.5.3 HCH Dehydrogenases (LinC and LinX)

LinC and LinX are dehydrogenases (28kDa each) in the short-chain alcohol dehydrogenase superfamily with similar activity though LinX is only 33.1% identical to LinC (Lal *et al.,* 2010, 2006; Nagata *et al.,* 1999). Their respective genes (*LinC* & *LinX*) occur as single copies in UT26, two copies in B90A, and three copies in Sp+ (Lal *et al*., 2006). Two highly conserved regions are said to exist in this superfamily; the first is at the amino-terminal end where NAD^+ binds and the second region spans from position 150 to 153 on the consensus sequence and contains tyrosine and lysine at positions 150 and 153, respectively but tyrosine is conserved in these regions whereas lysine is not (Nagata *et al.,* 1999).

2.6 HCH degradation pathway enzymes: Downstream pathway

The enzymes in the γ-HCH's downstream pathway include reductive dechlorinase (LinD), ring-cleavage dioxygenase (LinE), transcriptional regulator (LinR) as well as maleylacetate reductase (LinF), each encoded by *LinD*, *LinE*, *LinR*, and *LinF,* respectively. Three of the genes (*LinD*, *LinE*, & *LinR*) share 99 to 100% sequence identity among the *Sphingobium* strains; UT26, B90A, and Sp+ and were detected in other isolates, thus showing that these strains use the same downstream pathway (Lal *et al*., 2006). LinD, LinE, and LinF show similarity to pentachlorophenol degradation pathway enzymes (PcpC, PcpA, and PcpE, respectively) in *Sphingomonas chlorophenolica* ATCC 39723 (Lal *et al*., 2006).

2.6.1 Reductive Dechlorinase (LinD)

LinD is 38.4kDa and is similar to the theta class of glutathione-*S*-transferases (θ-GSTs). There are four classes of GSTs (α, β, π, and θ) and all bacterial GSTs belong to class theta (Nagata *et al.,* 1999). PcpC catalyzes the transformation of tetrachlorohydroquinone (TCHQ) to 2,6-dichlorohydroquinone (2,6-DCHQ) in *S. chlorophenolicum* ATCC 39723 and is highly similar to LinD compared to PcpA and PcpE (Nagata *et al.,* 1999).

2.6.2 Ring-cleavage dioxygenase (LinE)

LinE (Chloro/hydroquinone-1,2-dioxygenase) is a 36kDa enzyme with minor resemblance to dioxygenases in the meta-cleavage family (Lal *et al*., 2006). It has a 51% sequence identity and 72% sequence similarity to PcpA (2,6-dichloro-*p*-hydroquinone 1,2-dioxygenase) from *S. chlorophenolicum*. Its role is not yet known but is thought to prefer hydroquinone to catechol, a major substrate for meta-cleavage dioxygenases (Nagata *et al.,* 1999).

2.6.3 Transcriptional regulator (LinR)

LinR is a LysR-type transcriptional regulator (LTTR), of molecular weight 3.6kDa, and is a positive regulator for *LinE* and *LinD* expression in UT26, inducing their expression when either HQ, CHQ, or 2,5-DCHQ are available (Lal *et al*., 2006). Some aromatic compounds including catechol, chlorocatechol, and naphthalene also use LTTRs in their degradation pathways and $TN_{11}A$, a palindromic sequence located upstream of *LinE*, is the recognition sequence for LTTRs (Nagata *et al.,* 1999).

2.7 Prospects of using HCH-degrading microorganisms for bioremediation

Considering the extensive use of lindane and the huge stockpiles of HCH waste generated, there is a compelling need to create sustainable technologies for remediation of HCH-contaminated soils (Garg *et al*., 2016). Because spontaneous degradation of HCH occurs slowly and conjointly depends on the natural conditions, existing strategies of disposal (such as incineration and landfills) have major drawbacks and are not eco-friendly (Girish & Kunhi, 2013). Microbial remediation of HCH, on the other hand, is a new and alternative approach and is attracting considerable attention since it is cost effective, eco-friendly and does not produce toxic by-products (Kaur *et al*., 2021).

Bioremediation, a term that describes the use of microorganisms (or their enzymes) for the clearance of contaminants from the environment, has proved to be feasible and a promising strategy. To date, two approaches have been used for bioremediation and these include biostimulation and bioaugmentation (Girish & Kunhi, 2013; Lal *et al*., 2010). Biostimulation entails adding limiting nutrients to support the growth of microorganisms whereas bioaugmentation involves adding microorganisms capable of degradation (Adams *et al*., 2015). These strategies have been demonstrated in the laboratory and in pilot or *in situ* field settings (Garg *et al*., 2016; Lal *et al*., 2010; Phillips *et al*., 2006; Raina *et al*., 2008). However, the two approaches have proved much less effective whilst utilized in isolation because successful bioremediation requires the combination of the two approaches and involves numerous other factors (Adams *et al*., 2015; Garg *et al*., 2016; Lal *et al*., 2010).

A number of researches focusing on the degradation of HCH provide insights on the new these technologies for bioremediation (Kaur *et al*., 2021). For instance, Garg *et al*. (2016) used both biostimulation and bioaugmentation approaches to successfully reduce the levels of α- and β-HCH at HCH dumpsites. In addition, they also found the use of bacterial consortia rather than single HCH-degrading bacterial strains to be found more effective in field applications. Egorova *et al*. (2017) used bioaugmentation alone to significantly reduce the toxicity of HCH-contaminated soil. In spite of the recent advances, bioremediation is an evolving field and newer approaches are continually being discovered and implemented. From a recent study, for example, plants have been shown to exert beneficial effects on the community of microbes at contaminated locales by decreasing pollutant concentration whereas at the same time supporting the growth of plant-associated indigenous microbes. Thus, microbe-assisted plant-based bioremediation technique has been supported to remediate HCH-contaminated locales and further totally re-establish such contaminated locales by keeping up vegetation cover of appropriate plant species (Kaur *et al*., 2021).

2.8 Homology modeling

Experimental protein structure determination is very costly and takes a much longer time and is often less successful (Jaroszewski, 2009). Because of the advances in genome sequencing technologies, numbers of protein sequences have grown rapidly in comparison to the numbers of experimentally determined protein 3D structures, about 736 times bigger as of 2018 (Muhammed & Aki-Yalcin, 2019). Computational techniques for structure prediction are reducing this growing gap between the numbers of known protein sequences and solved crystal structures (Waterhouse *et al*., 2018). Comparative (or homology) modeling techniques are routinely used to generate three-dimensional (3D) structures of proteins where there are no existing experimentally determined structures (Biasini *et al*., 2014). Protein 3D structures are essential due to the fact they provide insights into the functioning of proteins at the molecular level and have numerous applications including structured-based drug design, designing of mutagenesis experiments, and identification of protein's catalytic and binding sites (Benkert *et al*., 2011).

Homology modeling uses the amino acid sequence of the target protein to predict its 3D structure using a set of evolutionary related structures known as templates (Biasini *et al*., 2014; Muhammed & Aki-Yalcin, 2019). Here, two major assumptions are applied; first, the protein's 3D structure is determined by its amino acid sequence, and second, the protein's 3D structure is much more conserved and changes happen at a much slower rate in evolutionary time. Therefore, similar sequences adopt the same 3D fold and even less related sequences have identical structures (França, 2015; Muhammed & Aki-Yalcin, 2019). The general rule regarding quality is that models generated with over 50% target-template sequence identity are sufficiently accurate for drug discovery applications; 25–50%, the models could help design experiments in mutagenesis studies; and for 10–25% sequence identity (or <25%), the models are mere superlatives (França, 2015). The homology modeling process comprises of the following steps: (1) identification of proteins (either one or more) with known 3D structure(s) to serve as a template(s); (2) sequence alignments of target and template; (3) model construction; (4) model optimization or refinement, and (5) validation of the resulting model (França, 2015; Muhammed & Aki-Yalcin, 2019; Munsamy & Soliman, 2017; Vyas *et al.,* 2012). These steps are iterated until a suitable model is built (Munsamy & Soliman, 2017).

CHAPTER THREE

3.0 MATERIALS AND METHODS

3.1 To investigate the presence and copy number of *Lin* genes in *Sphingobium* sp. S6 and *Sphingobium* sp. S8 by Southern blot hybridization

3.1.1 Genomic DNA isolation and agarose gel electrophoresis

Two strains of *Sphingobium* bacteria namely; *Sphingobium* sp. S6 and *Sphingobium* sp. S8, were used in the study. The bacterial strains were previously isolated in the laboratory (Department of Biochemistry, University of Nairobi) from HCH-contaminated soil collected at Kitengela in Kajiado County. Genomic DNA from strains S6 and S8 and the positive control (*Sphingobium indicum* B90A) was isolated using Quick-DNA Fungal/Bacterial Kit (Zymo Research, Tustin, CA 92780, United States). Bacterial samples were added directly to a ZR Bashing Bead™ Lysis Tube containing Bashing Bead Buffer and rapidly and efficiently lysed by bead beating using Genomic Lysis Buffer. The DNA was then isolated and purified using Zymo-Spin™ technique. The resulting DNA samples were examined for quality on a 1% (w/v) TAE (Tris-acetate-EDTA)-agarose gel containing ethidium bromide ($6.25x10^{-4}$ µg/ml) and electrophoresed for 1 hour and 30 mins at 80 V. The gel was then examined under a UV Transilluminator (Benchtop Variable Transilluminator; M–20V, P/N 095-0452-02).

3.1.2 Synthesis of Digoxigenin (DIG)-labeled DNA probes

DNA fragments of *Lin* genes (*LinA*, *B*, *C*, *D*, *E, R*, and *X*) as well as IS*6100*, to be used as probes for DNA-DNA hybridization, were prepared by PCR (Polymerase Chain Reaction) amplification of genomic DNA of B90A (positive control). The PCR DIG-labeling of DNA probes for hybridization was done by use of a PCR-DIG Probe Synthesis Kit (Roche, Mannheim, Germany). Each PCR DIG-labeling reaction was carried out in a 50 µl mixture comprising 5 µl of 10x PCR buffer with $MgCl_2$, 5 µl of PCR DIG probe synthesis mix, 5 µl of each 100 µM forward and reverse primers for the respective *Lin* genes, 0.75 µl of Enzyme mix, 5 µl of DNA template of strain B90A and 24.25 µl of PCR grade water.

The PCR DIG-labeling reactions were done using a TProfessional Thermocycler (Biometra GmbH, Mannheim, Germany) following the thermal profile: initial denaturation (94 ºC, 5 mins); 30 cycles comprising of denaturation (94 ºC, 1 min), primer annealing (5 ºC below the melting temperature of each of the primers, 1 min) and extension (72 ºC, 1 min); and the very last extension step (72 ºC, 7 mins). The annealing temperature of each *Lin* gene fragment was determined from the average melting temperature of their respective primers as shown (**Table 2**). Successful Digoxigenin (DIG)-labeled DNA probes were confirmed by electrophoresis as described in section 3.1.1.

Table 2. PCR primers used in the amplification of *Lin* genes from *Sphingobium* sp. S6 and *Sphingobium* sp. S8

Primer Name	Sequences (5′ to 3′)	Annealing temp.
LinA-F	GCGGATCCGCATGAGTGATCTAGACAGACTT	60 ºC
LinA-R	GCCTCGAGTTATGCGCCGGACGGTGCGAAATG	
LinB-F	GCGGATCCGCATGAGCCTCGGCGCAAAGCCA	60 ºC
LinB-R	GCCTCGAGTTATGCTGGGCGCAATCGCCGGAC	
LinC-F	GCGGATCCGCATGTCTGATTTGAGCGGC	63 ºC
LinC-R	GCCTCGAGTCAGATCGCGGTAAAGCCGCCGTC	
LinD-F	GCGAATTCAATGAGCGCTGATACAGAA	58 ºC
LinD-R	GCCTCGAGTTAGGCGTTGCTCAGGAGATGGAT	
LinE-F	AGGAATTCCATGATGCAACTGCCCGAA	57 ºC
LinE-R	AGCTCGAGCTCAAATGACGATCGGATC	
LinR-F	TGGGATCCCCGTGAATATAGATGACCTGG	60 ºC
LinR-R	GGGTCGACTCACACTCGCGCGGACAG	
LinX-F	GCGGATCCGCATGGCTAACAGACTCGCAGGCA	65 ºC
LinX-R	GCCTCGAGTCAAACACCCACGGACCAGCCTCC	
IS*6100*-F	CAATGCCAAAAGCTCTCTCC	48 ºC
IS*6100*-R	GGCTCTGTTGCAAAAATCG	

3.1.3 Restriction digestion of genomic DNA

Genomic DNA of bacteria strains S6 and S8 were subjected to restriction digestion using 8 restriction enzymes (i.e. *Bam*HI, *Eco*RI, *Eco*RV, *Hind*III, *Pst*I, *Sac*I, *Sal*I, and *Xho*I). A 50 µl restriction digest was set up as follows: 20 µl genomic DNA, 5 µl SureCut buffer, 2 µl restriction enzyme and 23 µl water were added into a microfuge tube and incubated for 1 hour in the water bath at 37ºC. After that, incubation in an oven at 65ºC for 15 mins to inactivate the enzyme followed for certain enzymes requiring inactivation.

The restriction enzymes were chosen based on previous study by Manickam *et al.* (2008). All enzymes are the commercially available NEB (New England BioLabs®) single cutter restriction enzymes. The presence of recognition sites within the gene for each restriction enzyme was evaluated for all the *Lin* genes (*LinA*−*LinX* & IS*6100*) using NEBcutter V2.0 tool (Vincze *et al*., 2003) available at (https://nc2.neb.com/NEBcutter2/) and tabulated as shown (**Table 3**).

Table 3. Recognition sequences of the enzymes used and the restriction site(s) within the *Lin* genes of *Sphingobium* sp. S6 and *Sphingobium* sp. S8

Enzyme	Recognition sequence (5'−3')	Presence/absence of restriction site(s) in the *Lin* genes							
		LinA	*LinB*	*LinC*	*LinD*	*LinE*	*LinR*	*LinX*	IS*6100*
*Bam*HI	…G▼GATCC…	−	+	−	+	+	−	−	−
*Eco*RI	…G▼AATTC…	−	−	−	−	−	+	−	−
*Eco*RV	…GAT▼ATC…	−	−	+	−	+	+	−	−
*Hind*III	…A▼AGCTT…	−	−	−	−	−	−	−	+
*Pst*I	…CTGCA▼G…	−	−	+	−	−	−	−	−
*Sac*I	…G▼AGCTC…	−	−	−	−	−	−	−	−
*Sal*I	…G▼TCGAC…	−	+	−	+	−	−	−	++
*Xho*I	…C▼TCGAG…	−	−	−	−	−	+	−	−

NB: (−): No restriction site; (+): One restriction site; (++): Two restriction sites

3.1.4 Restricted DNA fragments transfer to nitrocellulose membrane

About 40 µl of restricted genomic DNA fragments from bacteria strains S6 and S8 were electrophoresed on a 1.5% (w/v) Tris-acetate-EDTA agarose gel containing ethidium bromide (6.25x10^{-4} µg/ml). An aliquot of 5 µl DIG-labeled DNA molecular weight marker was loaded and run alongside DNA samples for 1 hour and 45 mins at 80 V. Before transfer, the gel was submerged in 0.25 M HCl at room temperature for 15 mins to depurinate the DNA. After incubation in HCl, the gel became rinsed in sterile double distilled water followed by immersion in denaturation solution (0.5 M NaOH, 1.5 M NaCl) twice for 15 mins at room temperature (15-25 ºC) to denature the DNA.

Again, the gel became rinsed in sterile distilled water and then immersed in a neutralization solution (0.5 M Tris-HCl, pH 7.5; 1.5 M NaCl) twice for 15 mins at room temperature. Subsequently, equilibration of the gel in 20x SSC (saline-sodium citrate or transfer buffer: 3.0 M NaCl, 0.3 M sodium citrate; pH 7.0) for 10 mins was done before transfer onto the nitrocellulose membrane. To blot the DNA onto nitrocellulose membrane, the capillary transfer method was used and the blot transfer system was set up (**Fig. 5**) and allowed to run overnight. After an overnight transfer, a brief wash step in 2x SSC was accomplished before fixing DNA onto the blot by baking in the oven at 80 ºC for 2 hours. The membrane was then hybridized and the bands were detected by the chromogenic method following the DIG Application Manual for Filter Hybridization (Roche Diagnostics GmbH, Mannheim, Germany).

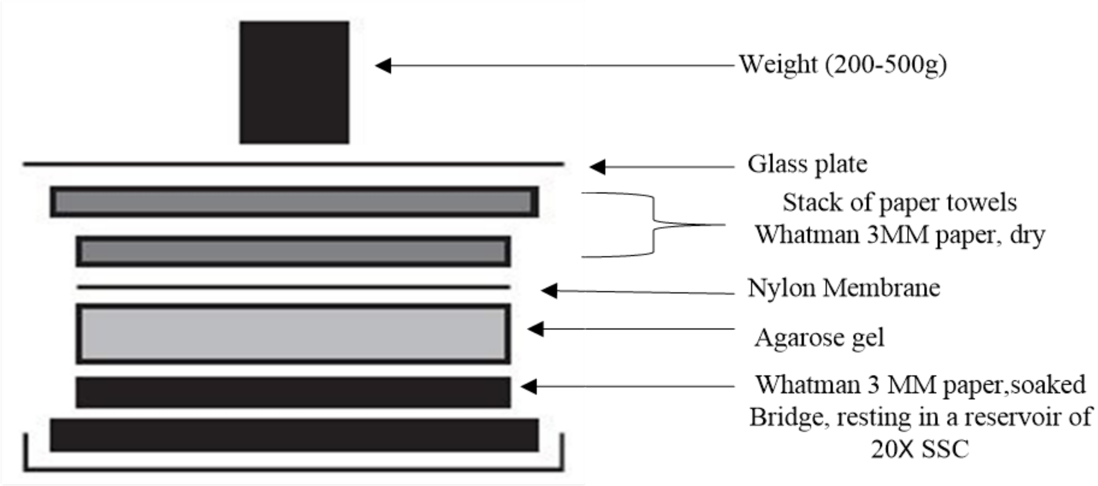


Figure 5. Capillary transfer method for Southern blotting (Sambrook & Russell, 2006). The soaked Whatman 3MM paper rests on absorbent bridge.

3.1.5 Prehybridization of a Southern blot

While the membrane was baking, 10 ml of pre-hybridization buffer (DIG Easy Hyb, Roche) was pre-warmed to the optimal hybridization temperature (50 ºC) calculated according to the formula: $\mathbf{T_m} = 49.82 + 0.41\ (\%G+C) - 600/L$; $\mathbf{T_{hyb}} = T_m - (20\ ºC\ \text{to}\ 25\ ºC)$ in which $\mathbf{T_m}$ is the melting temperature of target-probe hybrid, **(%G+C)** is the proportion of guanosine (G) and cytosine (C) residues in probe sequence, $\mathbf{T_{hyb}}$ is the optimal temperature for hybridization of the probe to target in DIG Easy Hyb, and **L** is the length of the hybrid in base pairs.

After baking, the membrane was pre-hybridized in a hybridization bag by addition of pre-warmed pre-hybridization buffer followed by incubation in an air bath at 50 ºC for 30 mins with mild agitation. At the same time, preparation of hybridization solution was done by addition of 7 µl of labeled DNA probe (Section 3.1.2) to 50 µl of sterile double distilled water in a microfuge tube. The microfuge tube was set in a boiling water bath for 5 mins to denature the DNA probe. After denaturation, the probe was rapidly chilled in ice and instantly added to 3.5 ml of pre-warmed hybridization buffer (DIG Easy Hyb; Roche

Diagnostics GmbH, Mannheim, Germany). After pre-hybridization, the pre-hybridization solution in the bag was replaced by a hybridization solution followed by incubation of the membrane in the hot air oven at 50 ºC for 16 hours under mild agitation. The hybridized membrane was then subjected to two 5 minute-washes in 50 ml of low stringency buffer (2x SSC, 0.1% sodium dodecyl sulfate or SDS) at 24 ºC, and two 15 minute-washes in preheated high stringency buffer (0.5x SSC, 0.1% SDS) at 65 ºC under constant agitation.

3.1.6 Chromogenic detection by NBT/BCIP stock solution

Target-probe hybrids were detected on the blot by the chromogenic method using DIG Wash and Block Buffer Set kit (Roche Diagnostics GmbH, Mannheim, Germany). After stringency washes, the membrane was put in 50 ml washing buffer (100 mM Maleic acid, 150 mM NaCl; pH 7.5; 0.3% (v/v) Tween 20) in a plastic container and incubated for 2 mins at room temperature (15-25 ºC). The washing buffer was poured off and 20 ml of blocking solution (made by dissolving 10% (w/v) blocking reagent in Maleic acid buffer [100 mM Maleic acid, 150 mM NaCl; pH 7.5]) was added and the membrane constantly agitated for 30 mins. The blocking solution was poured off and 10 ml of antibody solution (A freshly prepared solution of Anti-DIG-AP [Anti-Digoxigenin-Alkaline Phosphatase] 1:5000 (150 mU/ml) diluted in blocking solution) was added and the membrane was gently agitated for 30 mins.

Subsequently, the antibody solution was poured off and the membrane was washed two times for 15 mins in 50 ml washing buffer. This was followed by equilibration in 20 ml of detection buffer (100 mM Tris-HCl, 100 mM NaCl; pH 9.5). Finally, 10 ml of colour substrate solution (200 µl of BCIP/NBT in 10 ml detection buffer) was added to the membrane followed by incubation in the dark for 16 hours at 24 ºC without any agitation. After colour development reactions had generated bands of the desired intensity, the membrane was rinsed in 50 ml TE buffer (10 mM Tris-HCl, 1 mM EDTA; pH 8.0).

3.2 To evaluate the genetic variability of *Lin* genes in *Sphingobium* sp. S6 and *Sphingobium* sp. S8 by *Lin* gene sequencing, DNA sequence analysis, and phylogenetic investigation

3.2.1 Amplification of *Lin* genes by PCR

The *Lin* genes in the two bacterial strains were examined by PCR amplification of *Lin* gene fragments (*LinA*, *LinB*, *LinC*, *LinD*, *LinE*, *LinR,* and *LinX*) as well as the insertion sequence IS*6100*, using PCR Core Kit (Roche Diagnostics GmbH, Mannheim, Germany). Total genomic DNA extracted in section 3.1.1 served as the template for amplification by PCR using primers designed based on the conserved portion of each *Lin* gene fragment (**Table 2**; Böltner *et al*., 2005).

The PCR reactions were carried out in a 50 µl mixture that contained 5 µl 10x PCR buffer, 2 µl $MgCl_2$ (2 mM), 2 µl 10 mM dNTP mix (400 µM of each dATP, dCTP, dGTP, and dTTP), 2 µl of 1 µM forward and reverse primers for the respective *Lin* genes, 0.4 µl FastStart Taq DNA Polymerase (2U), 2 µl genomic DNA (up to 500 ng) and 34.6 µl PCR grade water. The PCR amplifications were performed as previously described (Section 3.1.2) and successful amplifications were confirmed by electrophoresis as outlined in section 3.1.1. The PCR amplicons were cleaned using the illustra™ GFX™ PCR DNA and Gel-Band Purification Kit (GE Healthcare GmbH, Germany) in line with the manufacturer's instructions. The amplicons were shipped to Macrogen Europe Laboratories (Amsterdam, The Netherlands) and commercially sequenced in both directions at Macrogen Europe Laboratories.

3.2.2 DNA sequence analysis

DNA Baser Sequence Assembler software v5.15.0 was used to check the quality of the chromatograms and build consensus sequences. DNA sequences obtained were translated using NCBI's (National Biotechnology Information) open reading frame (ORF) finder tool available at (https://www.ncbi.nlm.nih.gov/orffinder/). The resultant protein sequences were used as queries to search for similar sequences in protein databases including UniprotKB/TrEMBL and PDB (Protein Data Bank) using BLAST (Basic Local Alignment Search Tool) search tool available at (https://blast.ncbi.nlm.nih.gov/Blast.cgi) (Altschul, 2014). Highly homologous sequences were identified and selected based on Expectation value, query coverage, and percent identity and used for multiple alignments.

3.2.3 Multiple sequence alignment

Multiple sequence alignment (MSA) was performed using the translated protein sequences of *Lin* genes from *Sphingobium* sp. S6 and *Sphingobium* sp. S8 plus similar sequences selected from BLAST search. Selected protein sequences were first aligned in MUSCLE (Edgar, 2004) and manually modified using BioEdit V7.0.5.3 (Hall, 1999). The alignment graphics output generated were rendered as a Bitmap to show the conservation pattern and variations. From the alignment, percentage sequence identities and similarities were determined with respect to the query. Phylogenetic trees were constructed using the aligned sequences and evolutionary relatedness of the sequences and the organisms from whence the sequences originated inferred from the trees.

3.2.4 Phylogenetic analysis

Phylogenetic investigation of *Lin* genes was performed using the Bayesian inference method by MrBayes software v3.2.7 (https://nbisweden.github.io/MrBayes/) (Huelsenbeck & Ronquist, 2001; Ronquist *et al*., 2012; Ronquist & Huelsenbeck, 2003). The posterior distribution of model parameters used in estimating trees was approximated using Markov chain Monte Carlo (MCMC) (Ronquist *et al*., 2012; Ronquist & Huelsenbeck, 2003; Huelsenbeck & Ronquist, 2001). Analysis was performed using the GTR (Generalized Time Reversal) substitution model for proteins ([prset aamodelpr=mixed] lset nst=1 rates=invgamma); a mixture of amino acid models with fixed-rate matrices and in which the substitution rates and stationary state frequencies come from the mixture of models. A proportion of the sites were invariably distributed while the rest had a gamma distribution. The distribution of the shape parameter (shapepr) was exponential (1.00) whereas that of invariable sites (pinvarpr) was uniformly distributed on the interval (0.00, 1.00).

All tree topologies (topologypr) had equal probabilities whereas branch lengths (brlenspr) were unconstrained with a compound gamma Dirichlet distribution (unconstrained: gammadir [1, 1, 1, 1]). Two independent analyses (Nruns=2) were performed by MrBayes using four chains (3 "heated" chains and 1 "cold" chain). The MCMC sampling was performed over 3,000,000 generations at a sampling frequency of 1000 and the first 25% (relburnin=yes burninfrac=0.25) of samples were discarded when estimating the posterior probabilities of trees. After 3,000,000 generations, the analysis was stopped when the average standard deviation of split frequencies was less than 0.01 and tree parameters summarized. The resulting majority rule consensus phylogenetic trees constructed by MrBayes were visualized and modified by FigTree software v1.4.4 available at (http://tree.bio.ed.ac.uk/software/figtree/).

3.3 To model the three-dimensional (3D) structures of key enzymes in the HCH degradation pathway via comparative modeling

3.3.1 Comparative (homology) modeling of Lin proteins (LinA−LinE)

Three-dimensional (3D) structures of LinA, LinB, LinC, LinD, and LinE were determined by homology modeling approach as implemented in the SWISS-MODEL modeling pipeline (http://swissmodel.expasy.org/) (Biasini *et al*., 2014). A typical workflow in SWISS-MODEL comprises five steps namely; input data, template search, template selection, model building, and estimation of model quality.

3.3.1.1 Input data

Input data was an amino acid sequence was FASTA format or plain text and here appropriate, a specific UniprotKB accession code was used.

3.3.1.2 Template search

Templates showing significant similarity to target were identified by both BLAST and HHBlits searches against SMTL (Swiss Model Template Library) and ranked based on the expected quality of the resultant models via Global Mean Quality Estimate (GMQE) (Biasini *et al*., 2014) and Quaternary Structure Quality Estimate (QSQE) (Sadowski & Jones, 2007). GMQE scores are expressed as numerical values in the range of 0 and 1 reflecting the anticipated accuracy of the model such that the higher the GMQE score, the highly reliable the model is. A comparison of top-ranking templates, as well as alignments, was also made to determine if they represented different arrangements of the same structure or covered separate regions of the query (Waterhouse *et al*., 2018).

3.3.1.3 Template selection

The best template was chosen based on the characteristics of the target-to-alignment including sequence similarity, sequence identity and alignment score (Biasini *et al*., 2014). In general, a sequence identity >25% suggests that the template and target are likely to adopt the same 3D fold and the template is therefore appropriate for use in modeling (Muhammed & Aki-Yalcin, 2019). However, a greater than 30% (about 30–35%) sequence similarity is the limit for generating accurate 3D models (França, 2015; Jaroszewski, 2009; Munsamy & Soliman, 2017).

3.3.1.4 Model building

After template selection (either manually or automatic), a 3D model was built by the ProMod3 modeling engine in SWISS-MODEL using the target-template alignment generated (Waterhouse *et al*., 2018). The SWISS-MODEL employs rigid-frame assembly approach for model building whereby a protein structure is dissected into conserved core regions, adjoining loops, and side chains readorning the backbone and a 3D structure is built by bringing together those inflexible bodies (Muhammed & Aki-Yalcin, 2019; Vyas *et al*., 2012).

3.3.1.5 Model quality estimation

Modeling errors were quantified and the anticipated accuracy of the resultant model was approximated using SWISS-MODEL's QMEAN (Qualitative Model Energy ANalysis) scoring function (Biasini *et al*., 2014). Steric clashes or structural distortions that arise in the process of modeling were eliminated by energy minimization (Waterhouse *et al*., 2018). QMEAN estimated the "degree of nativeness" of the structural characteristics of the model on a global scale and ranged between -4 and 0 (zero). A score near zero indicated that the modeled structure was agreeable with experimental structures of the same size and therefore, a model of high quality while scores of -4 and below indicated models of low quality (Benkert *et al*., 2011). Global QMEAN scores indicated the overall quality of the model and were computed as Z-scores in comparison to scores obtained with highly resolved X-ray crystal structures (Biasini *et al*., 2014).

3.3.2 Validation of predicted homology models

The predicted 3D structures were checked for correctness by analyzing the stereochemistry, folding reliability, and packing quality (Damborský & Koča, 1999). Structure refinement and validation tools used in assessing the accuracy of the 3D models included PROCHECK (Laskowski *et al*., 2006, 1993), MOLPROBITY (Chen *et al*., 2010), ProSAII (Sippl, 1993), VERIFY3D (Eisenberg *et al*., 1997), ERRAT (Colovos & Yeates, 1993), and ANOLEA (Melo *et al*., 1997). Stereochemical quality was assessed by PROCHECK, quality of packing was assessed via "bump checks" and also by visually inspecting the hydrophilic and hydrophobic residues in the 3D protein model. Reliability in folding was assessed via 3D to 1D profile analysis by VERIFY3D (Eisenberg *et al*., 1997) and overall energy profile estimation by ProSAII (Sippl, 1993).

Models were subjected to energy minimization by molecular dynamics (MD) simulation in UCSF Chimera (Pettersen *et al*., 2004) using AMBER force field at 100 steps of steepest descent and 10 conjugate gradients steps, and Gasteiger-Huckel charges were assigned to the 3D protein model. Root Mean Square Deviations (RMSDs) were computed to verify the symmetry of the target (query) and template. The atomic coordinates (PDB format) of the 3D models generated were submitted to the PMDB (Protein Model Database) database (Castrignano *et al*., 2006) and assigned accession numbers.

CHAPTER FOUR

4.0 RESULTS

4.1 Investigating the presence and copy numbers of *Lin* genes in *Sphingobium* sp. S6 and *Sphingobium* sp. S8 by Southern blot hybridization

4.1.1 DIG-labeled DNA probes synthesized by PCR

The approximate sizes of the DNA probes for each *Lin* gene as deduced from the agarose gel (**Fig. 6**) were as follows: 700 and 500 bp for labeled and unlabeled *LinA*, respectively; 1000 and 900 bp for labeled and unlabeled *LinB*, respectively; 1000 and 800 bp for labeled and unlabeled *LinC*/*LinX*, respectively; 1200 and 1000 bp for labeled and unlabeled *LinD*/*LinE*, respectively; 1100 and 1000 bp for labeled and unlabeled *LinR*, 1000 and 700 bp for labeled and unlabeled IS*6100*, respectively.

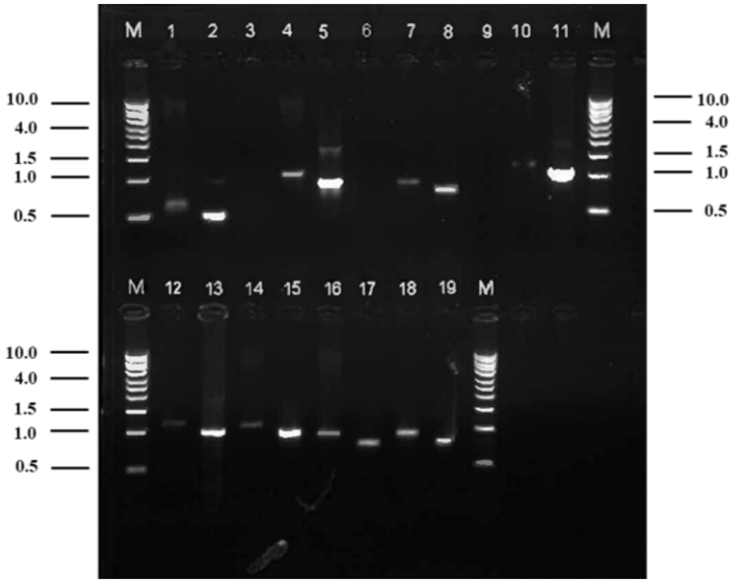


Figure 6. Agarose gel electrophoresis profile of DIG-labeled DNA probes.
M = 1 kb DNA Ladder (0.5–10 kb), Lane 1 and 2 = labeled and unlabeled *LinA*, respectively; Lane 4 and 5 = labeled and unlabeled *LinB*, respectively; Lane 7 and 8 = labeled and unlabeled *LinC*, respectively; Lane 10 and 11 = labeled and unlabeled *LinD*, respectively; Lane 12 and 13 = labeled and unlabeled *LinE*, respectively; Lane 14 and 15 = labeled and unlabeled *LinR*, respectively; Lane 16 and 17 = labeled and unlabeled *LinX*, respectively; Lane 18 and 19 = labeled and unlabeled IS*6100*, respectively. Lanes 3, 6, and 9 were empty.

4.1.2 Restriction digestion and Southern blot hybridization

The approximate copy numbers of *Lin* genes and the IS*6100* in *Sphingobium* sp. S6 were identified following the rationale that a probe hybridizing to a single portion of unfractionated target DNA produces only one band on the Southern blot. Alternatively, a probe hybridizing to various highly similar target DNA sequences (resulting from gene duplication) likely produces multiple bands. The maximum numbers of hybridizing bands produced by all restriction enzymes in each Southern blot are considered the copy numbers of *Lin* genes and the IS*6100*. Thus, the occurrence of one hybridizing band on the Southern blot in all restriction enzyme-digests indicates a single copy *Lin* gene is present whereas multiple bands show the presence of a multi-copy *Lin* gene. The same set of restriction enzymes was used for all *Lin* genes and IS*6100* hybridizations to maintain the pattern of restriction digestion. Hybridization of *Lin* genes of *Sphingobium* sp. S8 was not attempted as envisaged initially due to challenges encountered with the use of nitrocellulose membrane and hence such data could not be obtained.

4.1.2.1 Hybridization of *LinA*, *LinB*, and *LinC*

Southern blot hybridization of *LinA*, *LinB*, and *LinC* gene from *Sphingobium* sp. S6 with their respective DNA probes produced positive hybridization bands (**Fig. 7**). One hybridizing band each of sizes around 9416, 6557, and 4361 bp was observed in the S6 DNA probed with the *LinA* gene for all sets of restriction enzymes used in the blot. Similarly, one band was observed in the S6 DNA probed with *LinB* gene for all sets of restriction digests except *Bam*HI (Lane 1), *Pst*I (Lane 5), and *Sal*I digests (Lane 7) which generated two (9416 and 564 bp), three (9416, 4361, and 2027 bp), and two fragments (2027 and 564 bp), respectively. As for S6 DNA probed with *LinC* gene, all restriction enzymes used in the blot generated one band each of sizes around 9416, 6557, and 4361 bp except *Pst*I (Lane 5) which produced two fragments of sizes 9416 and 2027 bp (**Table 4**).

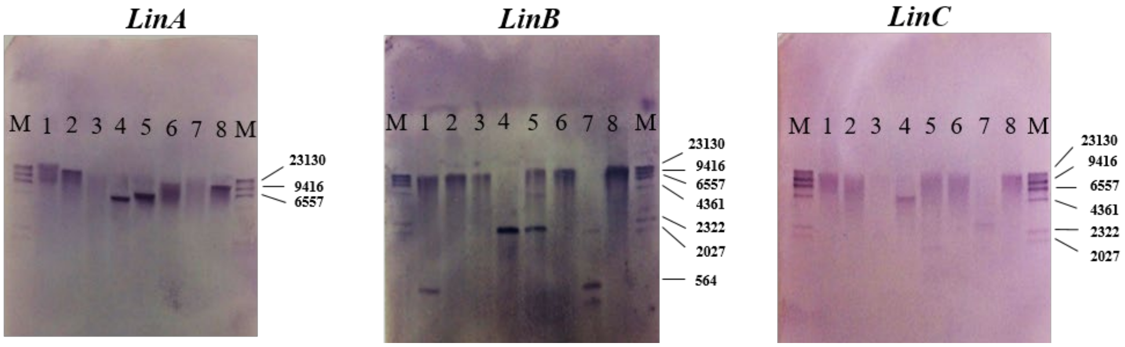


Figure 7. Southern blot analysis of genomic DNA of *Sphingobium* sp. S6 showing nitrocellulose membrane probed with DIG-labeled *LinA*, *LinB,* and *LinC* DNA probes, respectively. M: DIG-labeled Mol. Wt. marker, *LinA*–*LinC* Lanes DNA digested with 1. *Bam*HI, 2. *Eco*RI, 3. *Eco*RV, 4. *Hind*III, 5. *Pst*I, 6. *Sac*I, 7. *Sal*I, 8. *Xho*I.

Table 4. Approximate sizes (bp) of the hybridizing bands for restriction enzymes-digested *Lin* genes from *Sphingobium* sp. S6

***Lin* gene**	**Approximate size of the band(s) (bp) generated by each restriction enzyme**							
	Lane 1	**Lane 2**	**Lane 3**	**Lane 4**	**Lane 5**	**Lane 6**	**Lane 7**	**Lane 8**
LinA	6557	9416	–	4361	6557	9416	9416	9416
LinB	9416, 564	9416	9416	2027	9416, 4361, 2027	9416	2027, 564	9416
LinC	9416	6557	–	4361	9416, 2027	9416	4361	9416
LinD	9416	9416	6557	4361	9416	9416	9416	6557
LinE	9416, 564	6557, 564	6557, 125	4361, 564	4361, 564	6557, 564	9416, 564	9416, 564
LinR	23130, 564	9416, 2322, 564	9416, 6557, 4361, 564	9416, 4361, 2322, 564	9416, 4361, 2322, 2027, 564, 125	9416, 2322, 564	9416, 4361, 2322, 2027, 564	9416, 2322, 564
LinX	6557, 564	9416, 564	9416, 564	4361, 564	4361, 564	6557, 564	9416, 564	9416, 564
IS*6100*	23130	9416, 2322	9416, 6557, 4361	9416, 4361, 2322, 564	9416, 4361, 2322, 2027, 564	9416, 2322	9416, 4361, 2322, 2027, 564	9416, 2322

NB: (-): No bands were observed.

Nevertheless, *LinA* and *LinB* gene coding sequences (CDS) do not contain any restriction sites for all the restriction enzymes used except *Bam*HI and *Sal*I, which have one internal restriction site in *LinB* gene sequence (cut at positions 736 and 602, respectively). Similarly, the *LinC* gene CDS contains one internal restriction site for *Eco*RV and *Pst*I only (cut at positions 330 and 578, respectively) while the rest of the enzymes do not. Accordingly, one copy of *LinA*, *LinB*, and *LinC* gene were present in the genome of *Sphingobium* sp. S6.

4.1.2.2 Hybridization of *LinD*, *LinE,* and *LinR*

Southern blot hybridization of *LinD*, *LinE*, and *LinR* gene from *Sphingobium* sp. S6 with their respective DNA probes produced positive hybridization bands (**Fig. 8**). One band of sizes around 9416, 6557, and 4361 bp was observed in the S6 DNA probed with the *LinD* gene in all set of restriction-enzyme digests. As for S6 DNA probed with the *LinE* gene, all restriction enzymes generated two fragments of sizes around 9416, 6557, 4361, 564 and 125 bp. On the other hand, multiple bands were observed in the S6 DNA probed with the *LinR* gene in all restriction enzyme digests. With the exception of *Bam*HI and *Sal*I (cut at positions 413 and 394, respectively), all the restriction enzymes used do not contain any restriction sites internal to the *LinD* gene.

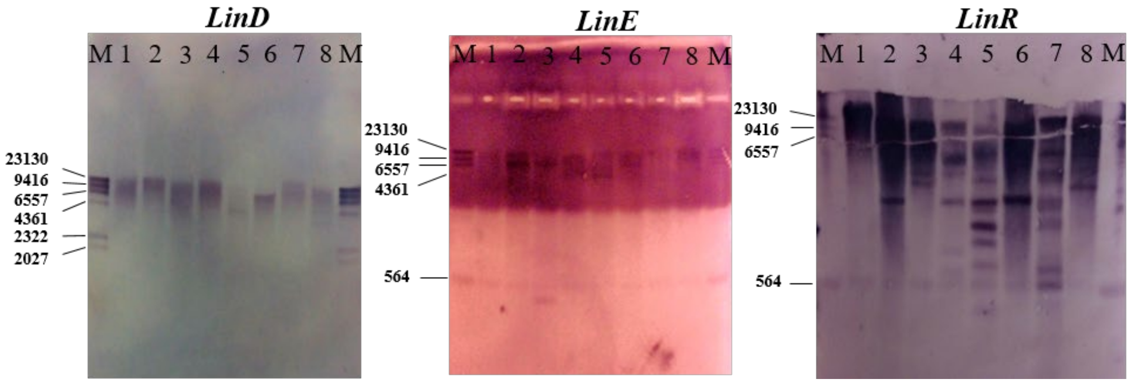


Figure 8. Southern blot analysis of genomic DNA from *Sphingobium* sp. S6 showing nitrocellulose membranes hybridized with *LinD*, *LinE,* and *LinR* DNA probes, respectively. M: DIG-labeled Mol. Wt. Marker, *LinD–LinR* Lanes 1 to 8, genomic DNA digested with *Bam*HI, *Eco*RI, *Eco*RV, *Hind*III, *Pst*I, *Sac*I, *Sal*I, and *Xho*I, respectively.

On the other hand, the *LinE* gene contains internal restriction sites for *BamH*I and *Eco*RV (cut at positions 67 and 263, respectively) only while the rest of the enzymes used do not. In the case of *LinR* gene, only *Eco*RI, *Eco*RV, and *Xho*I (cut at positions 290, 442, and 422, respectively) contain restriction sites internal to the gene while the rest of the restriction enzymes do not. Thus, one copy of *LinD*, two copies of *LinE*, and multiple copies of *LinR* gene occur in the genome of *Sphingobium* sp. S6. The approximate sizes of the bands were as shown (**Table 4**).

4.1.3.3 Hybridization of *LinX* and IS*6100*

Southern blot hybridization of the *LinX* gene and IS*6100* from *Sphingobium* sp. S6 with their respective probes produced positive hybridization bands (**Fig. 9**). Two bands of sizes around 9416, 6557, 4361, and 564 bp were observed in the *LinX* gene blot for all sets of restriction enzymes. However, in the IS*6100* gene blot, multiple bands were present. All of the restriction enzymes used in the blot do not contain any internal restriction sites to *LinX* gene whereas for *IS*6100, only *Hind*III and *Sal*I (cutting at nucleotide positions 729 and 409/566, respectively) have an internal restriction sites. Hence, two copies of *LinX* gene and multiple copies of IS*6100* occur in the genome of *Sphingobium* sp. S6. The approximate sizes of the bands were as shown (**Table 4)**.

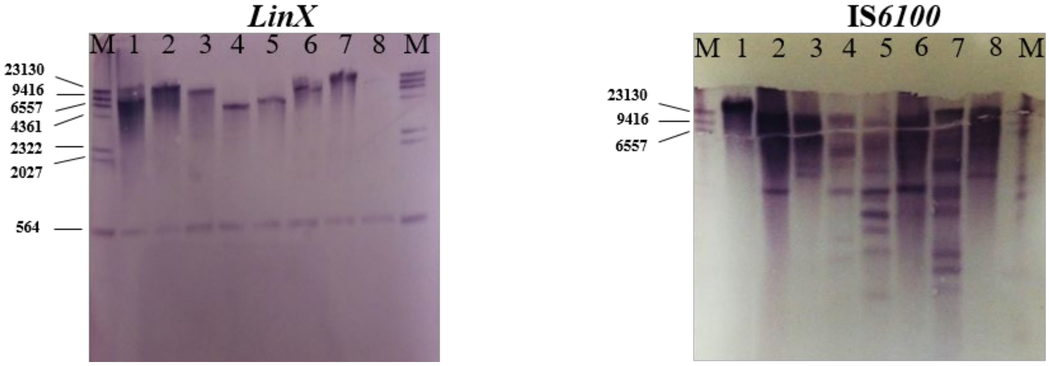


Figure 9. Southern blot hybridization of genomic DNA of *Sphingobium* sp. S6 showing nitrocellulose membrane hybridized with *LinX* and IS*6100* DNA probes. M: DIG-labeled Mol. wt marker, *LinX*–IS*6100* Lanes 1 to 8, genomic DNA digested with *Bam*HI, *Eco*RI, *Eco*RV, *Hind*III, *Pst*I, *Sac*I, *Sal*I, and *Xho*I, respectively.

4.2 Evaluating the genetic variability of *Lin* genes in *Sphingobium* sp. S6 and *Sphingobium* sp. S8 by *Lin* gene sequencing, sequence analysis, and phylogenetic investigation

4.2.1 Amplification of *Lin* genes by PCR

Lin genes present in *Sphingobium* sp. S6 and *Sphingobium* sp. S8 were amplified by PCR following detection by southern blot hybridization. Agarose gel electrophoresis of PCR amplicons from *Sphingobium* sp. S6 and *Sphingobium* sp. S8 showed identical *Lin* genes (based on the sizes) to be present. The approximate sizes of bands of *LinA*, *LinB*, *LinC*, *LinD*, *LinE*, *LinR*, *LinX*, and IS*6100* were 500, 900, 700, 1100, 1000, 900, 700, and 700 bp, respectively in both *Sphingobium* sp. S6 (**Fig. 10A**) and *Sphingobium* sp. S8 (**Fig. 10B**).

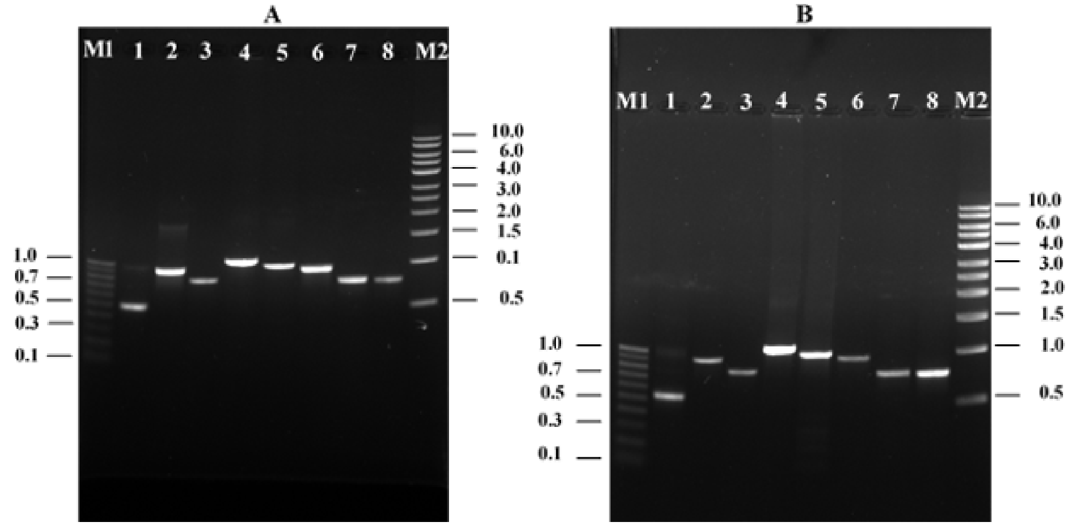


Figure 10. Agarose gel electrophoretic profile of PCR amplicons. **A**) DNA of *Sphingobium* sp. S6. M1: PCR 100 bp Low Ladder, M2: 1 kbp DNA Ladder; Lane 1: *LinA*, Lane 2: *LinB*, Lane 3: *LinC*, Lane 4: *LinD*, Lane 5: *LinE*, Lane 6: *LinR*, Lane 7: *LinX*, and Lane 8: IS*6100*. **B**) DNA of *Sphingobium* sp. S8. M1: PCR 100 bp Low Ladder, M2: 1 kbp DNA Ladder; Lane 1: *LinA*, Lane 2: *LinB*, Lane 3: *LinC*, Lane 4: *LinD*, Lane 5: *LinE*, Lane 6: *LinR*, Lane 7: *LinX*, and Lane 8: IS*6100.*

4.2.2 Sequencing and DNA sequence analysis

DNA sequence analysis of *Lin* genes from *Sphingobium* sp. S6 revealed the expected sizes 468, 888, 807, 1050, 975, 922, and 750 bp of *LinA*, *LinB*, *LinC*, *LinD*, *LinE*, *LinR*, and *LinX*, respectively. On the other hand, the expected sizes of *LinA*, *LinB*, *LinC*, *LinD*, *LinE*, *LinR*, and *LinX* genes from *Sphingobium* sp. S8 were 450, 872, 750, 1033, 951, 883, and 749 bp, respectively. The *Lin* genes of *Sphingobium* sp. S6 and *Sphingobium* sp. S8 were compared and shown to be identical at 93%, 96%, 76%, 75%, 87%, 95%, and 97% for *LinA*, *LinB*, *LinC*, *LinD*, *LinE*, *LinR*, and *LinX*, respectively.

Analysis of protein sequences of the respective *Lin* genes from *Sphingobium* sp. S6 showed they were closely similar to those from other HCH-degrading *Sphingobium* and/or *Sphingomonas* species at 98.7%, 99.3%, 100%, 100%, 99.6%, 99%, and 100% sequence similarity for LinA, LinB, LinC, LinD*,* LinE, LinR, and LinX, respectively. Similarly, those of *Sphingobium* sp. S8 showed highest sequence similarities at 94.2%, 96.9%, 99.6%, 99.4%, 98.4%, 97.9%, and 100% for LinA, LinB, LinC, LinD, LinE, LinR, and LinX, respectively. All the *Lin* genes (except *LinB*) of *Sphingobium* strains S6 and S8 were closely related to *Sphingobium japonicum* UT26S at the percentage sequence similarities shown (**Table 5**).

Table 5. Comparisons of translated protein sequences of the respective *Lin* genes from *Sphingobium* sp. S6 and *Sphingobium* sp. S8 with their closest match in UniprotKB/TrEMBL and PDB databases

Lin Gene	Nucleotides (aa) bp		Highest percent (%) sequence identity / similarity (aa)		Most closely related bacterial species (based on % protein sequence identity/similarity)	
	S6	S8	S6	S8	Bacteria	Accession No.
LinA	468	450	98.7/98.7	93.5/94.2	*S. japonicum* UT26S	BAI96690.1
	(156)	(150)	"	"	*S. indicum* B90A	APL95055.1
			"	"	*S. francense* Sp+	AAU11089.2
LinB	888	872	98.6/99.3	95.9/96.9	*Sphingobium* sp. TKS	AMK21182.1
	(296)	(291)	"	"	*P. aeruginosa* ITRC-5	ABP93361.1
LinC	807	750	100/100	99.2/99.6	*S. japonicum* UT26S	BAI95393.1
	(250)	(250)	"	"	*Sphingomonas* sp. NM05	ABG77568.1
LinD	1050	1033	99.4/100	99.1/99.4	*S. japonicum* UT26S	sp|D4Z909.1
	(346)	(344)				
LinE	975	951	99.6/99.6	98.1/98.4	*S. japonicum* UT26S	Q9WXE6.1
	(321)	(317)				
LinR	922	883	98.6/99.0	97.2/97.9	*S. japonicum* UT26S	Q9ZN79.3
	(301)	(293)	"	"	*S. baderi* LL03	EQA99717.1
LinX	750	749	100/100	100/100	*S. japonicum* UT26S	BAI96692.1
	(250)	(250)				

NB: S6 – *Sphingobium* sp. S6; S8 – *Sphingobium* sp. S8

4.2.3 Nucleotide sequence accession numbers

Nucleotide sequences of *Lin* genes (*LinA–LinX*) from *Sphingobium* sp. S6 were submitted to the GenBank via NCBI's BankIt submission portal available at (https://submit.ncbi.nlm.nih.gov/about/bankit/) and assigned the accession numbers MN649851–MN649857, respectively. Similarly, nucleotide sequences of *Lin* genes from *Sphingobium* sp. S8 (*LinA–LinX*) were deposited to be GenBank under the accession numbers MN649844–MN649850, respectively. In addition, the corresponding protein sequences (LinA–LinX) of the respective *Lin* genes from *Sphingobium* sp. S6 and *Sphingobium* sp. S8 were assigned the protein IDs QGJ16213−QGJ16219 and QGJ16206−QGJ16212, respectively.

4.2.4 Multiple sequence alignment and phylogenetic analysis

4.2.4.1 HCH dehydrochlorinase (LinA)

Translated HCH dehydrochlorinase (LinA) protein sequences from *Sphingobium* sp. S6 (Protein ID: QGJ16213.1) and *Sphingobium* sp. S8 (Protein ID: QGJ16206.1) contained 156 and 150 amino acid residues, respectively (**Table 5)**. From the MSA using seventeen HCH dehydrochlorinase (LinA) sequences (**Fig. 11**), most of the residues appeared to be conserved. However, LinA from *Sphingobium* sp. S6 (QGJ16213.1) and *Sphingobium* sp. S8 (QGJ16207.1) possessed four amino acid substitutions that were unique to the two dehydrochlorinases. These included Phe4, Gly10, Ser11, and Asn13 at the N-terminus of LinA from *Sphingobium* sp. S8 (QGJ16206.1), and Arg144 and Thr145 at the C-terminus of LinA from *Sphingobium* sp. S6 (QGJ16213.1).

Some mutated residues were also evident among the rest of the dehydrochlorinases in the alignment including Gln20, Gly23, Val35, Ile64, Tyr68, Thr71, Gln78, Cys96, Thr110, Cys111, Tyr113, Asn115, Leu126, Leu129, Gly131, Met133, Ala148, Leu149, Leu151, Gln (Thr)152, Lys153, and Thr154. Besides, LinA from *Sphingobium* sp. S6 (QGJ16213.1) and *Sphingobium* sp. S8 (QGJ16206.1) both belonged to the nuclear transport factor 2-like (NTF2) protein superfamily containing SnoaL-like domain similar to LinA from UT26S. This domain is highly conserved and spans between amino acid residues 5 and 130 and contains 34 highly conserved amino acid residues.

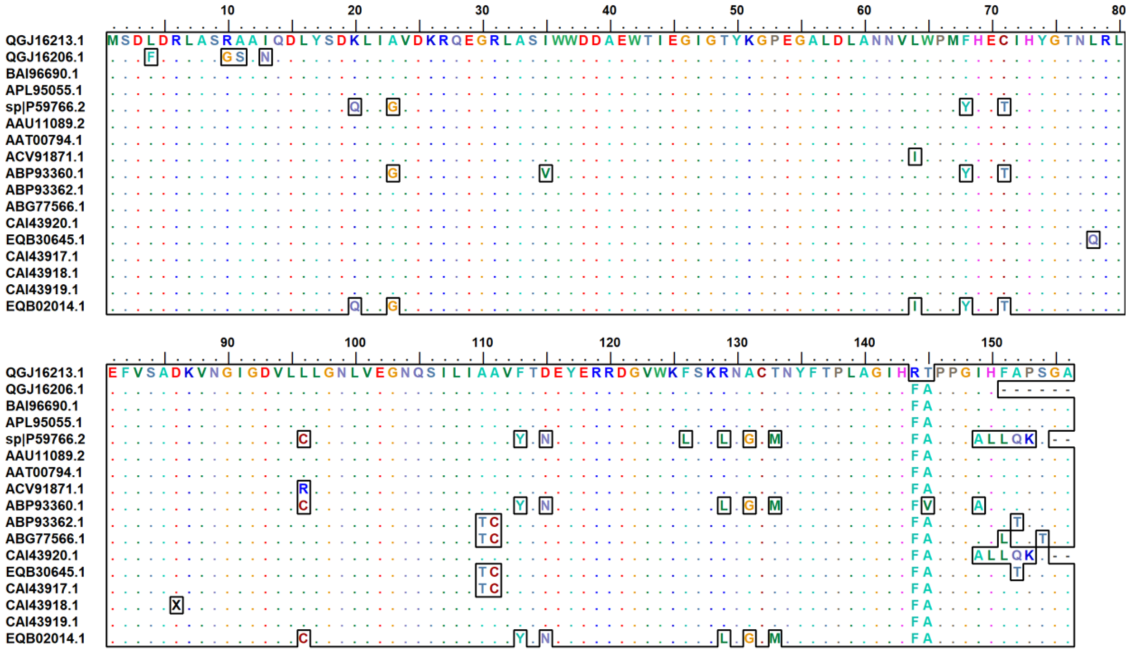


Figure 11. Multiple alignments of HCH dehydrochlorinase (LinA) protein sequences homologous to LinA from *Sphingobium* sp. S6 (QGJ16213.1) and *Sphingobium* sp. S8 (QGJ16206.1). The conservation pattern for identical residues is represented by dots and mutated residues are boxed. The alignment graphics output was generated by the use of BioEdit v7.0.5.3.

LinA sequences from both *Sphingobium* sp. S6 (QGJ16213.1) and *Sphingobium* sp. S8 (QGJ16206.1) were identical to each other at 92.3% sequence identity and 92.9% sequence similarity. They were closely similar to dehydrochlorinases from *S. japonicum* UT26S (BAI9660.1), *S. indicum* B90A (APL95055.1), *Sphingobium francense* Sp+ (AAU11089.2), *Rhodanobacter lindaniclasticus* (AAT00794.1), and *Sphingomonas* sp. γ16-1 (CAI43919.1). LinA sequence from *Sphingobium* sp. S6 (QGJ16213.1) was highly identical to the four dehydrochlorinases at 98.7% sequence identity and similarity, whereas LinA from *Sphingobium* sp. S8 (QGJ16206.1) showed the highest percent sequence identity and similarity to the four dehydrochlorinases at 93.5% and 94.2%, respectively (**Appendix 1A**).

From the phylogenetic tree constructed using twenty-six HCH dehydrochlorinase (LinA) sequences (**Fig. 12**), three distinct clusters were evident (**I**, **II** & **II**). The largest cluster **I** comprised of dehydrochlorinases from HCH-degrading *Sphingobium* and/or *Sphingomonas* species (78 to 100% support values). HCH-dehydrochlorinases (LinA) of bacteria strains S6 and S8 (accession numbers QGJ16213.1 and QGJ16206.1, respectively) formed a sister group relationship with dehydrochlorinases in this cluster, including the well-known and most closely related strains UT26S, B90A, and Sp+. They were characterized by short internal branches of approximately zero branch lengths, evidence of high sequence conservation. The second cluster **II** comprised mainly type 1 (LinA1 and LinAa) dehydrogenases from *S. indicum* B90A (sp|P59766.2) and *P. aeruginosa* (ABP93360.1), respectively, that are more diverged from LinA (56 to 57% support values). The third cluster **III**, on the other hand, consisted of dehydrochlorinases that are less similar to dehydrochlorinases in clusters **I** and **II**, mainly from two outgroup genera *Novosphingobium* and *Sphingopyxis*. They were characterized by long internal branches of considerable branch lengths and very identical to each other (99 to 100% support values).

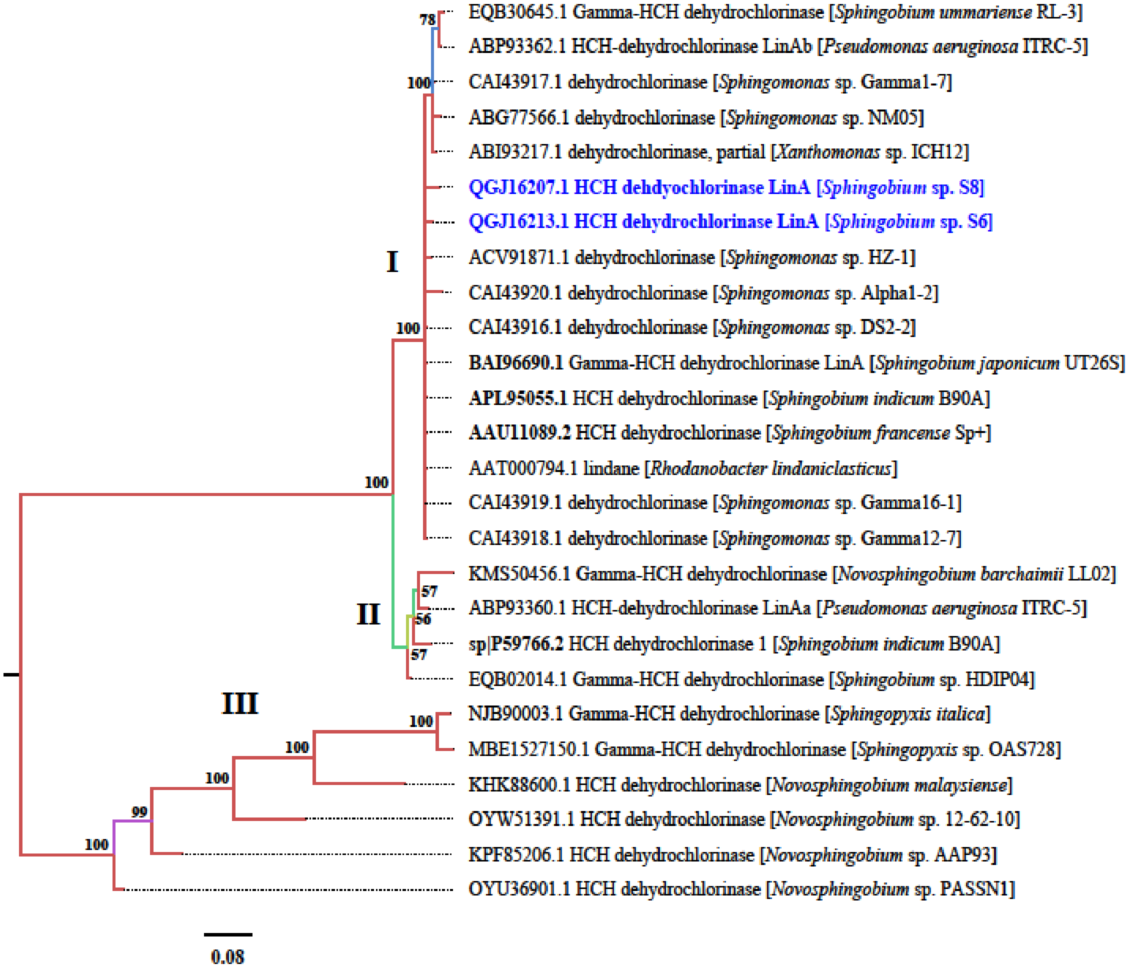

78
EQB30645.1 Gamma-HCH dehydrochlorinase [Sphingobium ummariense RL-3]
ABP93362.1 HCH-dehydrochlorinase LinAb [Pseudomonas aeruginosa ITRC-5]
100
CAI43917.1 dehydrochlorinase [Sphingomonas sp. Gamma1-7]
ABG77566.1 dehydrochlorinase [Sphingomonas sp. NM05]
ABI93217.1 dehydrochlorinase, partial [Xanthomonas sp. ICH12]
QGJ16207.1 HCH dehdyochlorinase LinA [Sphingobium sp. S8]
QGJ16213.1 HCH dehydrochlorinase LinA [Sphingobium sp. S6]
I
ACV91871.1 dehydrochlorinase [Sphingomonas sp. HZ-1]
CAI43920.1 dehydrochlorinase [Sphingomonas sp. Alpha1-2]
100
CAI43916.1 dehydrochlorinase [Sphingomonas sp. DS2-2]
BAI96690.1 Gamma-HCH dehydrochlorinase LinA [Sphingobium japonicum UT26S]
APL95055.1 HCH dehydrochlorinase [Sphingobium indicum B90A]
AAU11089.2 HCH dehydrochlorinase [Sphingobium francense Sp+]
AAT000794.1 lindane [Rhodanobacter lindaniclasticus]
100
CAI43919.1 dehydrochlorinase [Sphingomonas sp. Gamma16-1]
CAI43918.1 dehydrochlorinase [Sphingomonas sp. Gamma12-7]
KMS50456.1 Gamma-HCH dehydrochlorinase [Novosphingobium barchaimii LL02]
57
II
ABP93360.1 HCH-dehydrochlorinase LinAa [Pseudomonas aeruginosa ITRC-5]
56
sp|P59766.2 HCH dehydrochlorinase 1 [Sphingobium indicum B90A]
57
EQB02014.1 Gamma-HCH dehydrochlorinase [Sphingobium sp. HDIP04]
III
100
NJB90003.1 Gamma-HCH dehydrochlorinase [Sphingopyxis italica]
100
MBE1527150.1 Gamma-HCH dehydrochlorinase [Sphingopyxis sp. OAS728]
100
KHK88600.1 HCH dehydrochlorinase [Novosphingobium malaysiense]
99
OYW51391.1 HCH dehydrochlorinase [Novosphingobium sp. 12-62-10]
100
KPF85206.1 HCH dehydrochlorinase [Novosphingobium sp. AAP93]
OYU36901.1 HCH dehydrochlorinase [Novosphingobium sp. PASSN1]
0.08

Figure 12. Phylogenetic tree by MrBayes (v3.2.7) for HCH-dehydrochlorinase (LinA) protein sequences. LinA sequences from *Sphingobium* strains S6 and S8 are shown in blue. Numbers indicated on the nodes are percent posterior probabilities showing statistical support for each node. The scale bar below the tree shows the number of expected changes (or substitutions) for each site. Because there was no particular outgroup, the root was placed at the midpoint between the most similar and least similar sequences. The references strains used included UT26S, B90A and Sp+ with accession numbers BAI96690.1, APL95055.1, and AAU11089.2, respectively.

4.2.4.2 Haloalkane dehalogenase (LinB)

Translated LinB protein sequences from *Sphingobium* sp. S6 (Protein ID: QGJ16214.1) and *Sphingobium* sp. S8 (Protein ID: QGJ16207.1) contained 296 and 291 amino acid residues, respectively (**Table 5**). From the MSA using sixteen haloalkane dehalogenase sequences (**Fig. 13**), almost all residues appear to be conserved among dehalogenases in the alignment except for the presence of a few amino acid substitutions. Some mutated residues were identified in LinB sequences from *Sphingobium* sp. S6 (QGJ16214.1) and

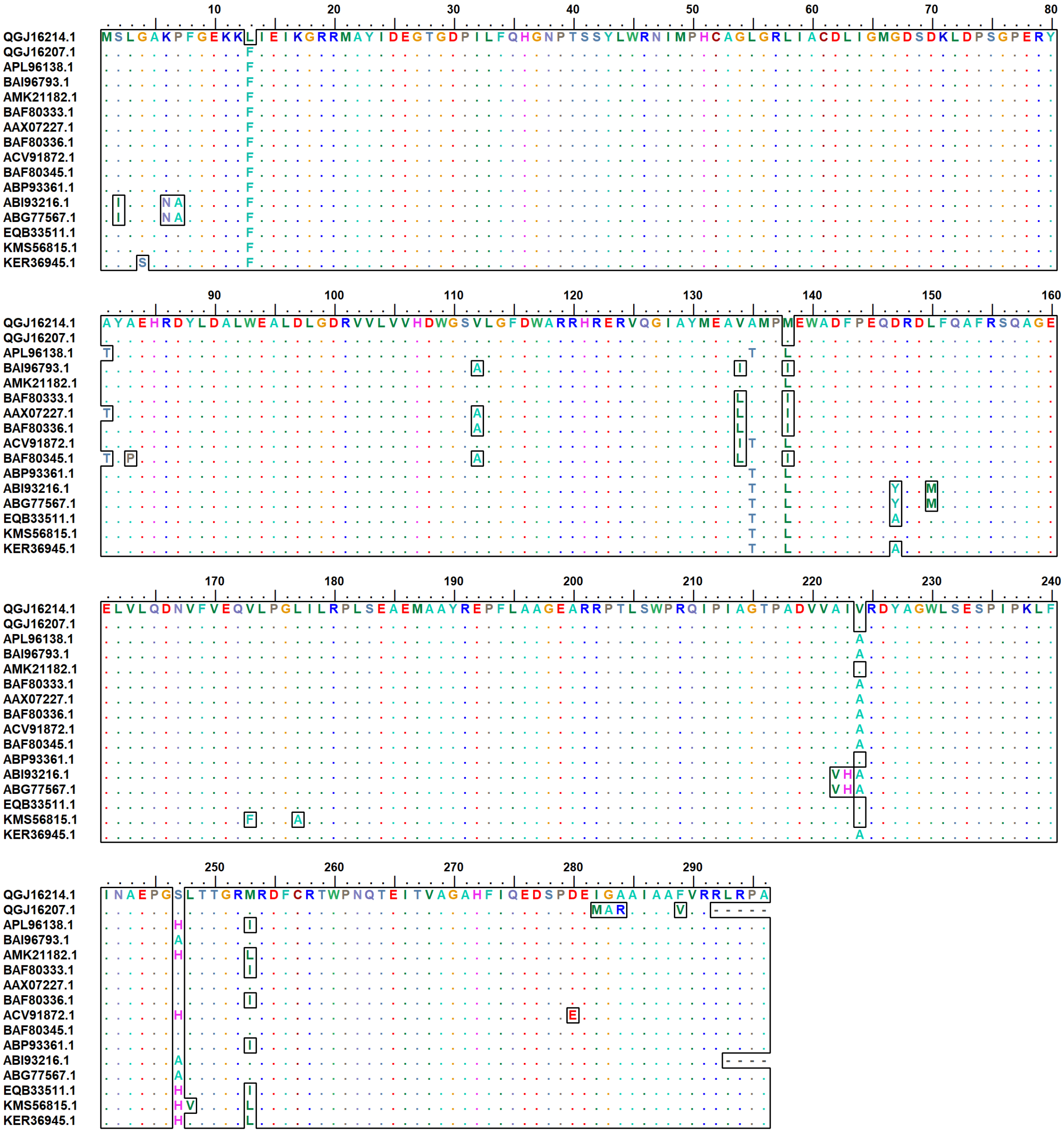


Figure 13. Multiple alignments of haloalkane dehalogenase (LinB) protein sequences homologous to LinB from *Sphingobium* sp. S6 (QGJ16214.1) and *Sphingobium* sp. S8 (QGJ16207.1). The conservation pattern for identical residues is represented by dots and mutated residues are boxed. The alignment graphics output was generated by the use of BioEdit v7.0.5.3.

Sphingobium sp. S8 (QGJ16207.1), which included Leu13 at the N-terminus in QGJ16214.1 and Met282, Ala283, Arg284, and Val289 at the C-terminus in QGJ16207.1. The Met138 substitution was common to the two dehalogenases but all other residues were conserved. Other noticeable substitutions included residues at positions 134, 138, 224, 247, and 253 and these are the most highly variable residues among dehalogenases.

LinB sequences from *Sphingobium* sp. S6 (QGJ16214.1) and *Sphingobium* sp. S8 (QGJ16207.1) were identical to each other at 96.6% sequence identity and 96.9% sequence similarity and were closely similar to haloalkane dehalogenases from *Sphingobium* sp. TKS (AMK2118.2) and *Pseudomonas aeruginosa* (ABP93361.1). LinB sequence from *Sphingobium* sp. S6 (QGJ16214.1) showed the highest percent sequence identity and similarity to AMK2118.2 and ABP93361.1 at 98.6% and 99.3%, respectively, whereas LinB from *Sphingobium* sp. S8 (QGJ16214.1) showed the highest percent sequence identity and similarity at 95.9% and 96.9%, respectively (**Appendix 1B**).

The phylogenetic tree constructed using twenty-five haloalkane dehalogenase sequences revealed two distinct clusters (**Fig. 14**). The larger cluster (**I**) comprised of dehalogenases from HCH- degrading *Sphingobium* and/or *Sphingomonas* species including the well-known and closely similar *Sphingobium* strains UT26S, B90A, and Sp+. Members within this cluster formed sister groups (taxa) and were characterized by short internal branches (56 to 100% support values). The haloalkane dehalogenases from *Sphingobium* strains S6 and S8 (accession numbers QGJ16214.1 and QGJ16207.1, respectively) belonged to this cluster and formed a sister taxon (72% support value). The second cluster (**II**) comprised of haloalkane dehalogenases from the genera *Gammaproteobacteria*, *Alphaproteobacteria*, *Cupriavidus*, *Mycobacteria*, and *Verrucomicrobia* (74–100% support values).

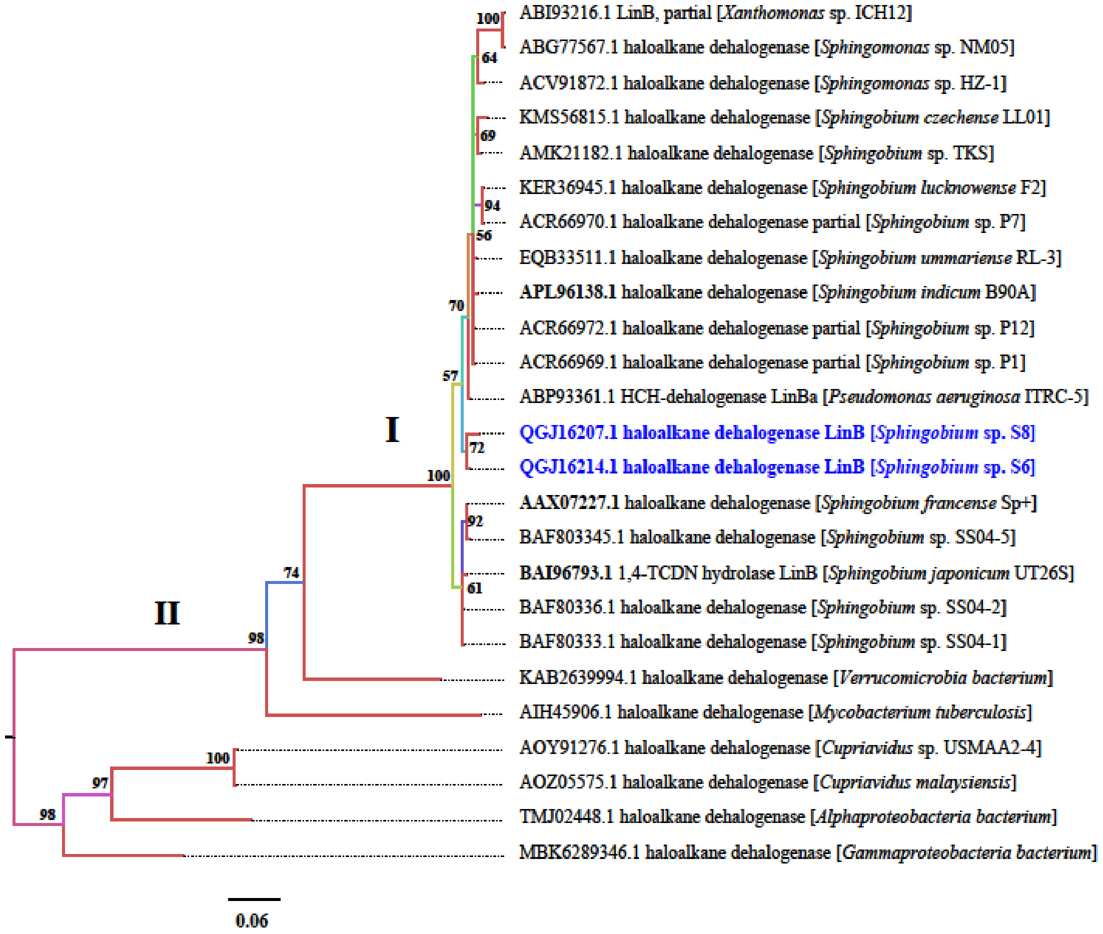

ABI93216.1 LinB, partial [Xanthomonas sp. ICH12]
ABG77567.1 haloalkane dehalogenase [Sphingomonas sp. NM05]
ACV91872.1 haloalkane dehalogenase [Sphingomonas sp. HZ-1]
KMS56815.1 haloalkane dehalogenase [Sphingobium czechense LL01]
AMK21182.1 haloalkane dehalogenase [Sphingobium sp. TKS]
KER36945.1 haloalkane dehalogenase [Sphingobium lucknowense F2]
ACR66970.1 haloalkane dehalogenase partial [Sphingobium sp. P7]
EQB33511.1 haloalkane dehalogenase [Sphingobium ummariense RL-3]
APL96138.1 haloalkane dehalogenase [Sphingobium indicum B90A]
ACR66972.1 haloalkane dehalogenase partial [Sphingobium sp. P12]
ACR66969.1 haloalkane dehalogenase partial [Sphingobium sp. P1]
ABP93361.1 HCH-dehalogenase LinBa [Pseudomonas aeruginosa ITRC-5]
QGJ16207.1 haloalkane dehalogenase LinB [Sphingobium sp. S8]
QGJ16214.1 haloalkane dehalogenase LinB [Sphingobium sp. S6]
AAX07227.1 haloalkane dehalogenase [Sphingobium francense Sp+]
BAF803345.1 haloalkane dehalogenase [Sphingobium sp. SS04-5]
BAI96793.1 1,4-TCDN hydrolase LinB [Sphingobium japonicum UT26S]
BAF80336.1 haloalkane dehalogenase [Sphingobium sp. SS04-2]
BAF80333.1 haloalkane dehalogenase [Sphingobium sp. SS04-1]
KAB2639994.1 haloalkane dehalogenase [Verrucomicrobia bacterium]
AIH45906.1 haloalkane dehalogenase [Mycobacterium tuberculosis]
AOY91276.1 haloalkane dehalogenase [Cupriavidus sp. USMAA2-4]
AOZ05575.1 haloalkane dehalogenase [Cupriavidus malaysiensis]
TMJ02448.1 haloalkane dehalogenase [Alphaproteobacteria bacterium]
MBK6289346.1 haloalkane dehalogenase [Gammaproteobacteria bacterium]
100
64
69
94
56
70
57
72
100
92
61
74
98
97
98
100
I
II
0.06

Figure 14. Phylogenetic tree by MrBayes (v3.2.7) for haloalkane dehalogenase (LinB) protein sequences. LinB sequences from *Sphingobium* strains S6 and S8 are shown in blue. Numbers indicated on the nodes are percent posterior probabilities showing statistical support for each node. The scale bar below the tree shows the number of expected changes (or substitutions) for each site. Because there was no particular outgroup, the root was placed at the midpoint between the most similar and least similar sequences. The accession numbers of the reference strains (UT26S, B90A, and Sp+) were BAI96793.1, APL96138.1, and AAX07227.1, respectively.

4.2.4.3 2,5-DDOL dehydrogenase (LinC)

Translated 2,5-DDOL dehydrogenase (LinC) protein sequences from *Sphingobium* sp. S6 (Protein ID: QGJ16215.1) and *Sphingobium* sp. S8 (Protein ID: QGJ16208.2) each comprised of 250 amino acid residues, respectively (**Table 5**). From the MSA using ten 2,5-DDOL dehydrochlorinase sequences (**Fig. 15**), all residues were conserved in the

LinC sequence from *Sphingobium* sp. S6 (QGJ16215.1). On the contrary, the LinC sequence from *Sphingobium* sp. S8 (QGJ16208.2) contained two unique amino acid substitutions of Thr243 and Asn244 at the C-terminus.

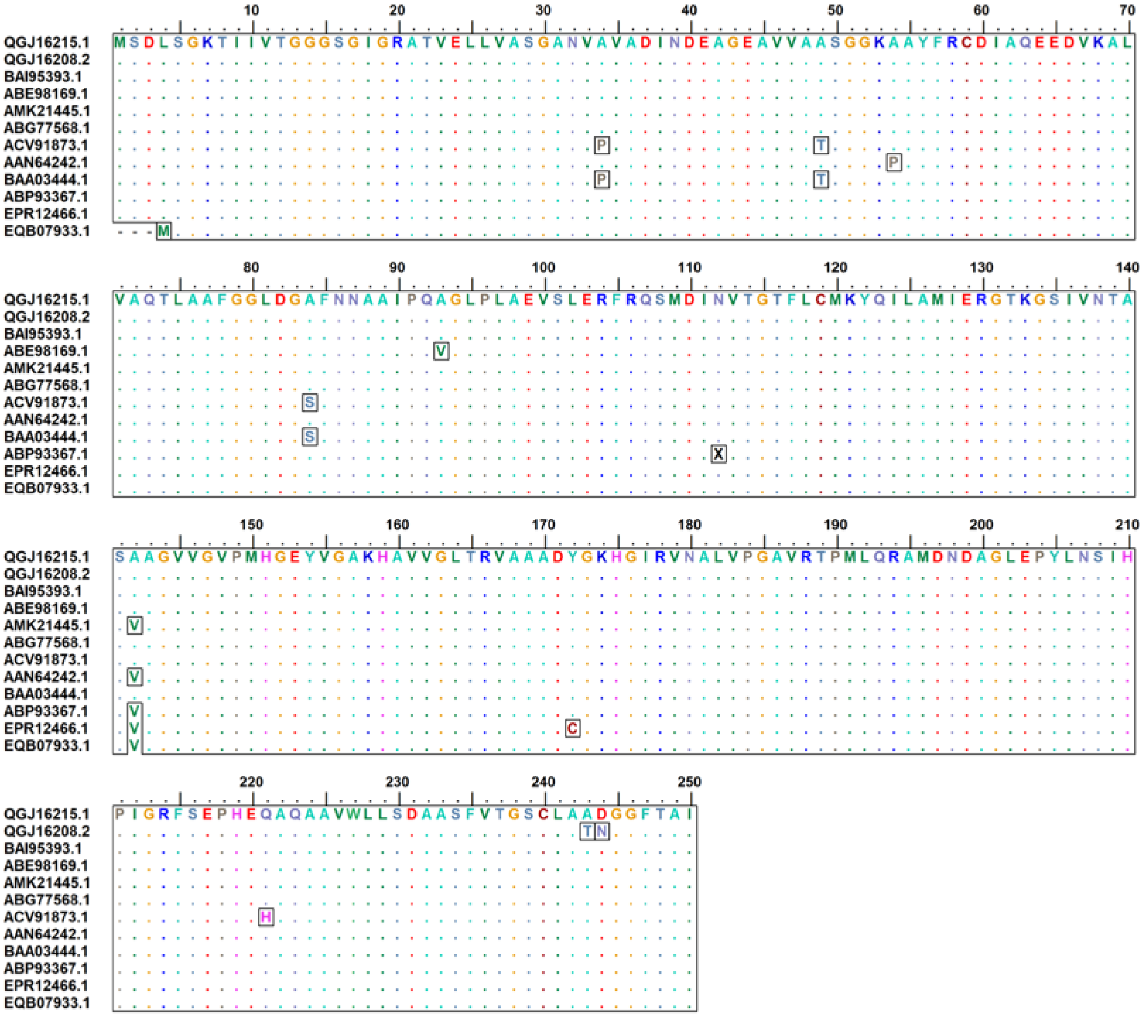


Figure 15. Multiple alignments of 2,5-DDOL dehydrogenase (LinC) protein sequences homologous to LinC from *Sphingobium* sp. S6 (QGJ16215.1) and *Sphingobium* sp. S8 (QGJ16208.2). The conservation pattern for identical residues is represented by dots and mutated residues are boxed. The alignment graphics output was generated by the use of BioEdit v7.0.5.3.

A conserved domains search by ScanProsite (de Castro *et al.*, 2006) available at (https://prosite.expasy.org/scanprosite/) revealed LinC sequences from both *Sphingobium* sp. S6 (QGJ16215.1) and *Sphingobium* sp. S8 (QGJ16208.2) possess the short-chain dehydrogenases/reductases (SDR) family signature (ADH_SHORT) that spans 141–169 (S$_{141}$AAGVVGVPMHGEYVGAKHAVVGLTRVAA$_{169}$) residues and contained the active site residue tyrosine at position 154. LinC sequences from both *Sphingobium* sp. S6 (QGJ16215.1) and *Sphingobium* sp. S8 (QGJ16208.2) were identical to each other at 99.2% sequence identity and 99.6% sequence similarity. LinC sequence from *Sphingobium* sp. S6 (QGJ16215.1) was very identical to the 2,5-DDOL dehydrogenase from *S. japonicum* UT26S (BAI95393.1) and the short-chain alcohol dehydrogenase from *Sphingomonas* sp. NM05 (ABG77568.1) at 100% sequence identity and similarity, respectively. On the contrary, the LinC sequence from *Sphingobium* sp. S8 (QGJ16208.2) was also closely similar to the two dehydrogenases (BAI95393.1 and ABG77568.1) at 99.2% sequence identity and 99.6% sequence similarity, respectively (**Appendix 1C**).

The phylogenetic tree constructed using twenty-four 2,5-DDOL (short-chain alcohol) dehydrogenase sequences revealed two separate clusters (**I & II**) as shown (**Fig. 16**). The first cluster (**I**) comprised of 2,5-DDOL dehydrogenases from HCH-degrading *Sphingobium* and/or *Sphingomonas* species, including the well-known and most highly similar strains UT26S, BHC-A and NM05. This cluster was characterized by short internal branches of zero branch lengths (92 to 100% support values). The 2,5-DDOL dehydrogenases of bacteria strains S6 and S8 (accession numbers QGJ16215.1 and QGJ16208.2, respectively) belonged to this cluster. The second cluster (**II**) comprised of 2,5-DDOL dehydrogenase sequences from the genera *Pseudomonas*, *Curvibacter*, *Massilia*, *Noviherbaspirillum*, and *Crocosphaera* (100% support values). They were less similar to those of cluster **I** and were characterized by long internal branches.

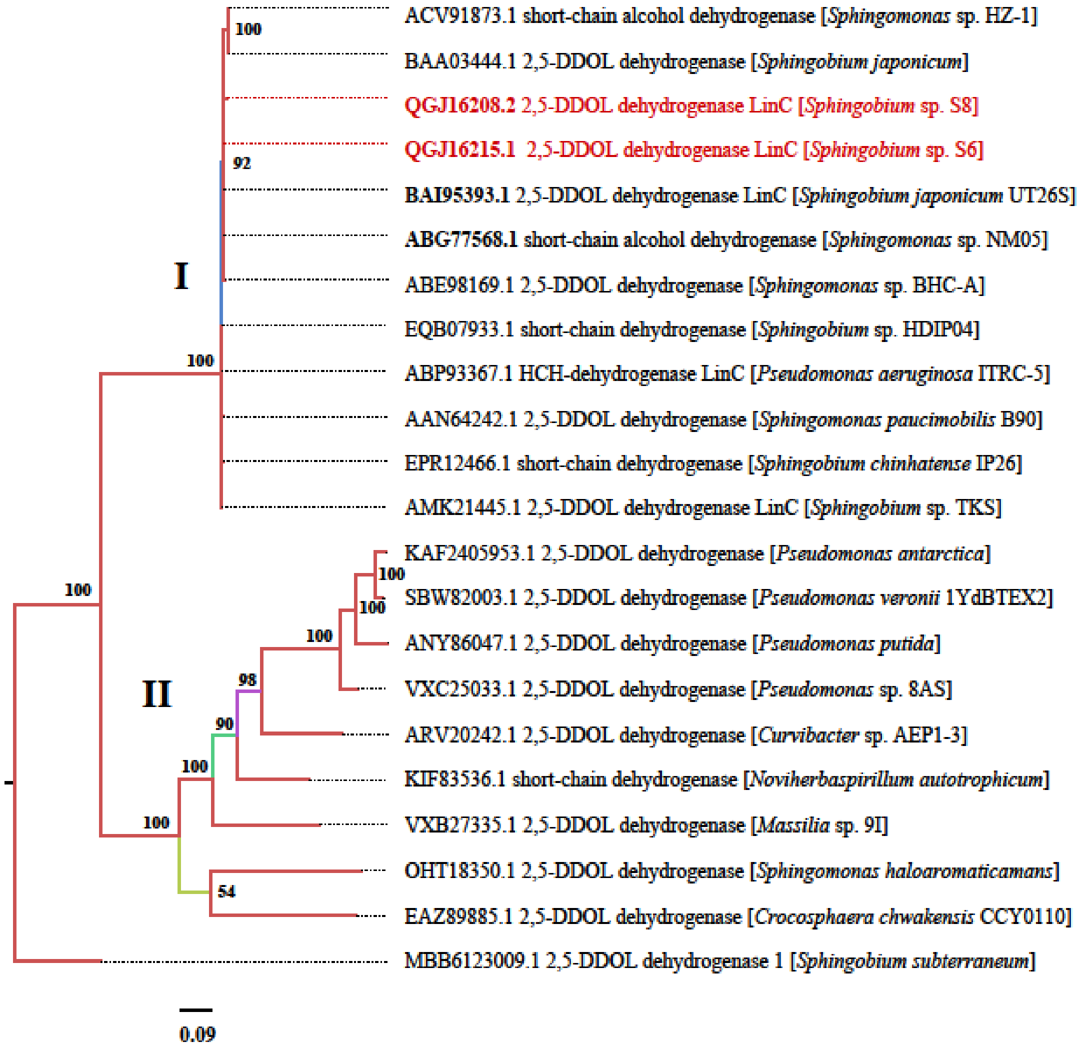

ACV91873.1 short-chain alcohol dehydrogenase [Sphingomonas sp. HZ-1]
BAA03444.1 2,5-DDOL dehydrogenase [Sphingobium japonicum]
QGJ16208.2 2,5-DDOL dehydrogenase LinC [Sphingobium sp. S8]
QGJ16215.1 2,5-DDOL dehydrogenase LinC [Sphingobium sp. S6]
BAI95393.1 2,5-DDOL dehydrogenase LinC [Sphingobium japonicum UT26S]
ABG77568.1 short-chain alcohol dehydrogenase [Sphingomonas sp. NM05]
ABE98169.1 2,5-DDOL dehydrogenase [Sphingomonas sp. BHC-A]
EQB07933.1 short-chain dehydrogenase [Sphingobium sp. HDIP04]
ABP93367.1 HCH-dehydrogenase LinC [Pseudomonas aeruginosa ITRC-5]
AAN64242.1 2,5-DDOL dehydrogenase [Sphingomonas paucimobilis B90]
EPR12466.1 short-chain dehydrogenase [Sphingobium chinhatense IP26]
AMK21445.1 2,5-DDOL dehydrogenase LinC [Sphingobium sp. TKS]
KAF2405953.1 2,5-DDOL dehydrogenase [Pseudomonas antarctica]
SBW82003.1 2,5-DDOL dehydrogenase [Pseudomonas veronii 1YdBTEX2]
ANY86047.1 2,5-DDOL dehydrogenase [Pseudomonas putida]
VXC25033.1 2,5-DDOL dehydrogenase [Pseudomonas sp. 8AS]
ARV20242.1 2,5-DDOL dehydrogenase [Curvibacter sp. AEP1-3]
KIF83536.1 short-chain dehydrogenase [Noviherbaspirillum autotrophicum]
VXB27335.1 2,5-DDOL dehydrogenase [Massilia sp. 9I]
OHT18350.1 2,5-DDOL dehydrogenase [Sphingomonas haloaromaticamans]
EAZ89885.1 2,5-DDOL dehydrogenase [Crocosphaera chwakensis CCY0110]
MBB6123009.1 2,5-DDOL dehydrogenase 1 [Sphingobium subterraneum]
100
92
I
100
100
100
100
100
98
II
90
100
100
54
0.09

Figure 16. Phylogenetic tree by MrBayes (v3.2.7) for 2,5-DDOL dehydrogenase (LinC) protein sequences. LinC sequences from *Sphingobium* strains S6 and S8 are shown in red. **MBB6123009.1**: 2,5-DDOL dehydrogenase 1 (*Sphingobium subterraneum*) was used as the outgroup in rerooting the tree. Numbers indicated on the nodes are percent posterior probabilities showing statistical support for each node. The scale bar below the tree shows the number of expected changes (or substitutions) for each site. The accession number of the reference strains (UT26S, BHC-A, and NM05) were BAI95393.1, ABE98169.1, and ABG77568.1, respectively.

4.2.4.4 2,5-DCHQ reductive dechlorinase (LinD)

Translated 2,5-dichloro-2,5-hydroquinone (2,5-DCHQ) reductive dechlorinase (LinD) protein sequences from *Sphingobium* sp. S6 (Protein ID: QGJ16216.1) and *Sphingobium* sp. S8 (Protein ID: QGJ16209.1) contained 346 and 344 amino acid residues, respectively (**Table 5**). From the MSA using seven 2,5-DCHQ reductive dechlorinase sequences (**Fig. 17**), all residues appeared to be conserved in all the LinD sequences. However, some amino acid substitutions were evident, including Met24 at the N-terminal end in LinD sequence from *Sphingobium* sp. S8 (QGJ16209.1), and Asn336 and Gln337 at the C-terminus in LinD sequence from *Sphingobium* sp. S6 (QGJ16216.1). Other notable amino acid substitutions included Leu64, Pro69, Ser82, Pro116, Arg144, Ala175, Gly194, Arg196, Ala198, Arg206, Arg230, Lys247, Ala275, Leu292, Leu298, Ile300, and Arg301 among the rest of the dechlorinase sequences in the alignment. The amino acid substitutions in LinD sequences from both *Sphingobium* sp. S6 (QGJ16216.1) and *Sphingobium* sp. S8 (QGJ16209.1) was unique to the two dechlorinases.

A conserved domain (CD) search by ScanProsite (https://prosite.expasy.org/scanprosite/) (de Castro *et al*., 2006) revealed LinD sequences from both *Sphingobium* sp. S6 (QGJ16216.1) and *Sphingobium* sp. S8 (QGJ162109.1) contained the soluble glutathione *S*-transferase (GST) domains at the N- and C-terminal ends. For LinD sequence from *Sphingobium* sp. S6 (QGJ16216.1), the GST N-terminal (GST_NTER) and GST C-terminal (GST_CTER) domains spanned residues 43–154 and 189–342, respectively. Similarly, the GST N-terminal and GST C-terminal domains spanned residues 43–154 and 189–335, respectively in the LinD sequence from *Sphingobium* sp. S8 (QGJJ16209.1). LinD sequences from both *Sphingobium* sp.S6 (QGJ16216.1) and *Sphingobium* sp. S8 (QGJ16209.1) were identical to each other at 98.5% sequence identity and 99.4% sequence similarity. They were closely similar to 2,5-DCHQ reductive dechlorinase (LinD) from *Sphingobium japonicum* UT26S (sp|D4Z909.1). LinD sequence from *Sphingobium* sp. S6 (QGJ16216.1) showed the highest sequence identity and similarity at 99.1% and 100%, respectively, whereas LinD sequence from *Sphingobium* sp. S8 (QGJ16209.1) showed the highest sequence identity and similarity 99.1% and 99.4%, respectively (**Appendix 1D**).

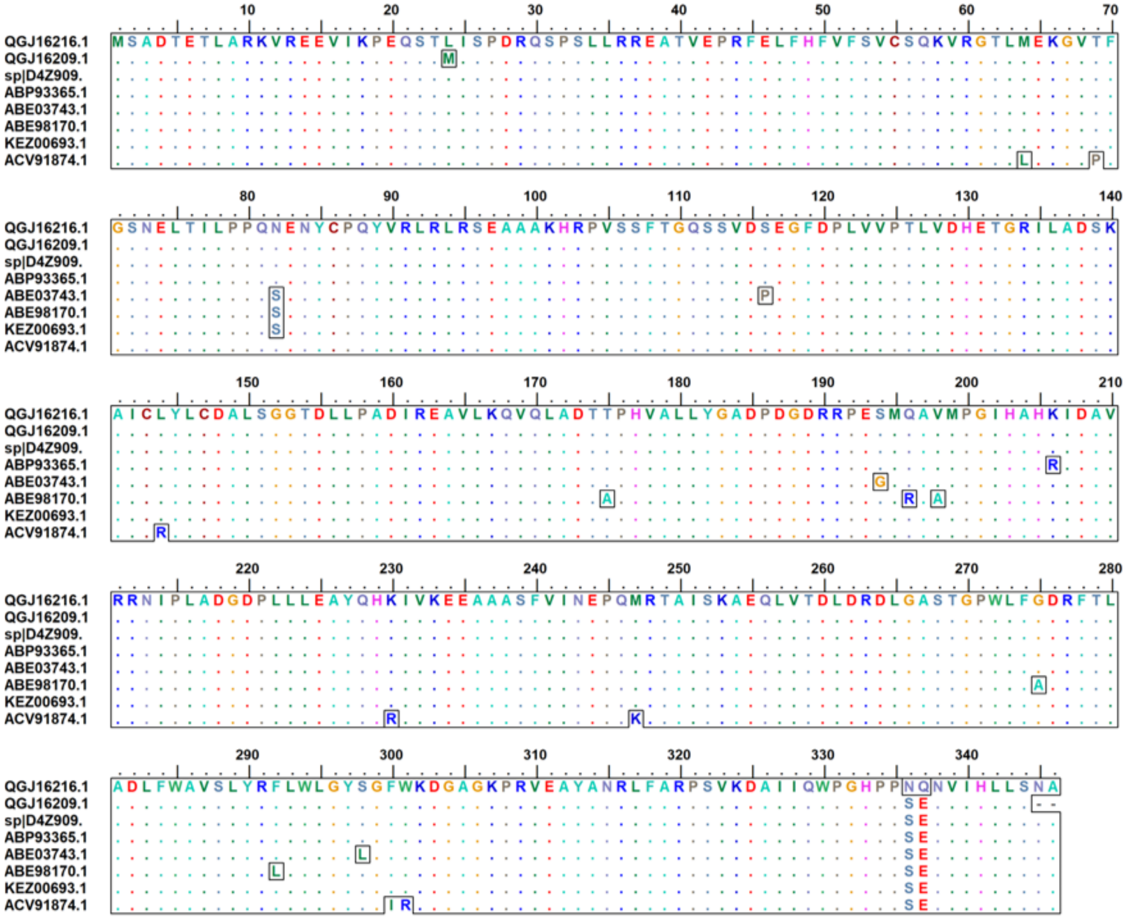


Figure 17. Multiple alignments of 2,5-DCHQ reductive dechlorinase (LinD) protein sequences homologous to LinD from *Sphingobium* sp. S6 (QGJ16216.1) and *Sphingobium* sp. S8 (QGJ16209.1). The conservation pattern for identical residues is represented by dots and mutated residues are boxed. The alignment graphics output was generated by the use of BioEdit v7.0.5.3.

From the phylogenetic tree constructed using fifteen reductive dechlorinase sequences (**Fig. 18**), two distinct clusters (**I & II**) were evident. The first cluster (**I**) comprised of reductive dechlorinases from HCH-degrading *Sphingobium* and/or *Sphingomonas* bacteria. The second cluster (**II**) comprised of reductive dechlorinases that are less similar to those of cluster **I** and were mainly from the outgroup genera *Variovorax*, *Rhizobium*, and *Pseudomonas*. LinD sequences of both *Sphingobium* species strains S6 and S8 (accession numbers QGJ16216.1 and QGJ16209.1, respectively) clustered with the well-known and most closely similar *S. japonicum* UT26S, in cluster **I**. Dechlorinases in cluster **I** were characterized by short internal branches of approximately zero branch lengths and posterior probabilities ranging from 85−100%. The data was insufficient to give an unambiguous branching pattern within the tree.

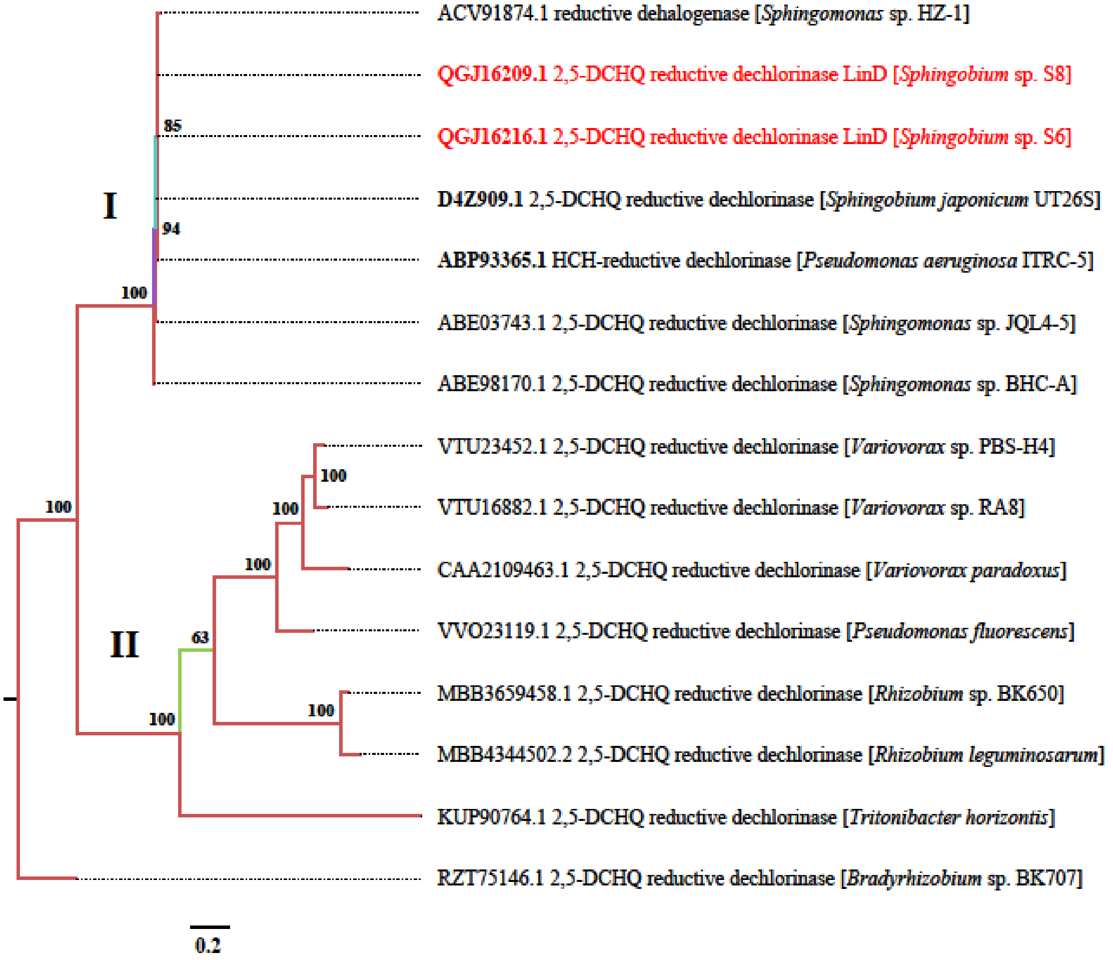

ACV91874.1 reductive dehalogenase [*Sphingomonas* sp. HZ-1]
QGJ16209.1 2,5-DCHQ reductive dechlorinase LinD [*Sphingobium* sp. S8]
85
QGJ16216.1 2,5-DCHQ reductive dechlorinase LinD [*Sphingobium* sp. S6]
I
D4Z909.1 2,5-DCHQ reductive dechlorinase [*Sphingobium japonicum* UT26S]
94
ABP93365.1 HCH-reductive dechlorinase [*Pseudomonas aeruginosa* ITRC-5]
100
ABE03743.1 2,5-DCHQ reductive dechlorinase [*Sphingomonas* sp. JQL4-5]
ABE98170.1 2,5-DCHQ reductive dechlorinase [*Sphingomonas* sp. BHC-A]
VTU23452.1 2,5-DCHQ reductive dechlorinase [*Variovorax* sp. PBS-H4]
100
100
100
VTU16882.1 2,5-DCHQ reductive dechlorinase [*Variovorax* sp. RA8]
100
CAA2109463.1 2,5-DCHQ reductive dechlorinase [*Variovorax paradoxus*]
VVO23119.1 2,5-DCHQ reductive dechlorinase [*Pseudomonas fluorescens*]
II
63
MBB3659458.1 2,5-DCHQ reductive dechlorinase [*Rhizobium* sp. BK650]
100
100
MBB4344502.2 2,5-DCHQ reductive dechlorinase [*Rhizobium leguminosarum*]
KUP90764.1 2,5-DCHQ reductive dechlorinase [*Tritonibacter horizontis*]
RZT75146.1 2,5-DCHQ reductive dechlorinase [*Bradyrhizobium* sp. BK707]
0.2

Figure 18. Phylogenetic tree by MrBayes (v3.2.7) for 2,5-DCHQ reductive dechlorinase (LinD) protein sequences. LinD sequences from *Sphingobium* strains S6 and S8 are shown in red. **RZT75146.1**: 2,5-DCHQ reductive dechlorinase (*Bradyrhizobium* sp. BK707) was used as the outgroup in rerooting the tree. Numbers indicated on the nodes are percent posterior probabilities showing statistical support for each node. The scale bar below the tree shows the number of expected changes (substitutions) for each site. The reference strains included UT26S and ITRC-5 with accession numbers D4Z909.1 and ABP93365.1, respectively.

4.2.4.5 Chloro/hydroquinone 1,2-dioxygenase (LinE)

Translated (chloro) hydroquinone 1,2-dioxygenase (LinE) protein sequences from *Sphingobium* sp. S6 (Protein ID: QGJ16217.1) and *Sphingobium* sp. S8 (Protein ID: QGJ16210.1) is composed of 321 and 317 amino acid residues, respectively (**Table 5**). From the MSA using seven dioxygenase sequences (**Fig. 19**), all residues were conserved except for a few substituted amino acid residues, including Phe11, Gln13, Phe14, Arg19, Trp35, Arg158, Thr163, Glu208, Leu209, Asp233, Asp265, Cys268, Pro313, and Leu321.

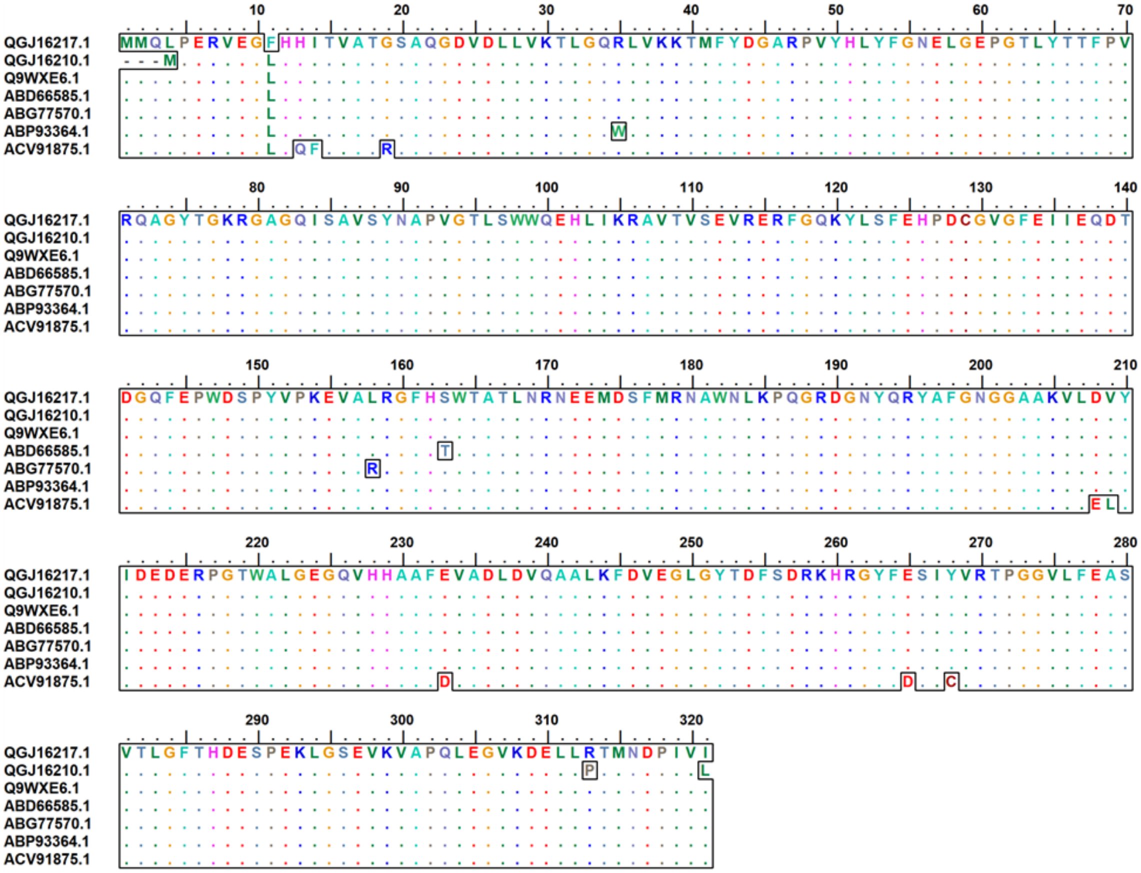


Figure 19. Multiple alignments of (chloro) hydroquinone 1,2-dioxygenase (LinE) protein sequences homologous to LinE from *Sphingobium* sp. S6 (QGJ16217.1) and *Sphingobium* sp. S8 (QGJ16210.1). The conservation pattern for identical residues is represented by dots and mutated residues are boxed. The alignment graphics output was generated by the use of BioEdit v7.0.5.3.

The substitutions of Phe11 at the N-terminus in LinE sequence from *Sphingobium* sp. S6 (QGJ16217.1), and Met4 and Pro313 at the N- and C-termini, respectively in LinE sequence from *Sphingobium* sp. S8 (QGJ16210.1), were unique to the two meta-cleavage dioxygenases. In addition, LinE from *Sphingobium* sp. S8 (QGJ16210.1) contained a deletion of three residues at the N-terminal end and one residue at the C-terminal end.

Based on ScanProsite (https://prosite.expasy.org/scanprosite/) search for conserved domains (de Castro *et al*., 2006), the LinE sequence from both *Sphingobium* sp. S6 (QGJ16217.1) and *Sphingobium* sp. S8 (QGJ16210.1) was found to contain the vicinal oxygen chelate (VOC) domain that is typical of all meta-cleavage dioxygenases. This domain spanned residues 10–138 and 160–282 at the N- and C-termini, respectively in LinE sequence from *Sphingobium* sp. S6 (QGJ16217.1) and contained the active site residue glutamate at positions 134 and 278. In LinE sequence from *Sphingobium* sp. S8 (QGJ16210.1), on other hand, this domain spanned residues 7–135 and 157–279 at the N- and C-terminal ends, respectively, and contained the active site residue glutamate at positions 131 and 275.

LinE sequences from both *Sphingobium* sp. S6 (QGJ16217.1) and *Sphingobium* sp. S8 (QGJ16210.1) were identical to each other at 97.8% sequence identity and 98.1% sequence similarity. They were closely similar to (chloro) hydroquinone 1,2-dioxygenase (LinE) from *S. japonicum* UT26S (Q9WXE6.1). LinE sequence from *Sphingobium* sp. S6 (QGJ16217.1) showed the highest sequence identity and similarity to Q9WXE6.1 at 99.6% and 99.6%, respectively. LinE sequence from *Sphingobium* sp. S8 (QGJ16210.1) showed the highest sequence identity and similarity to Q9WXE6.1 at 98.1% and 98.4%, respectively (**Appendix 1E**).

The phylogenetic tree constructed using eighteen dioxygenase sequences (**Fig. 20**), revealed two separate clusters (**I & II**). The first cluster (**I**) comprised of dioxygenases from HCH-degrading *Sphingobium* or *Sphingomonas* bacteria whereas the second cluster (**II**) consisted of dioxygenases from outgroup genera *Caballeronia*, *Paraburkholderia*, *Burkholderia*, *Rhodospirillaceae*, *Sphingopyxis,* and *Pigmentiphaga*. LinE sequences of

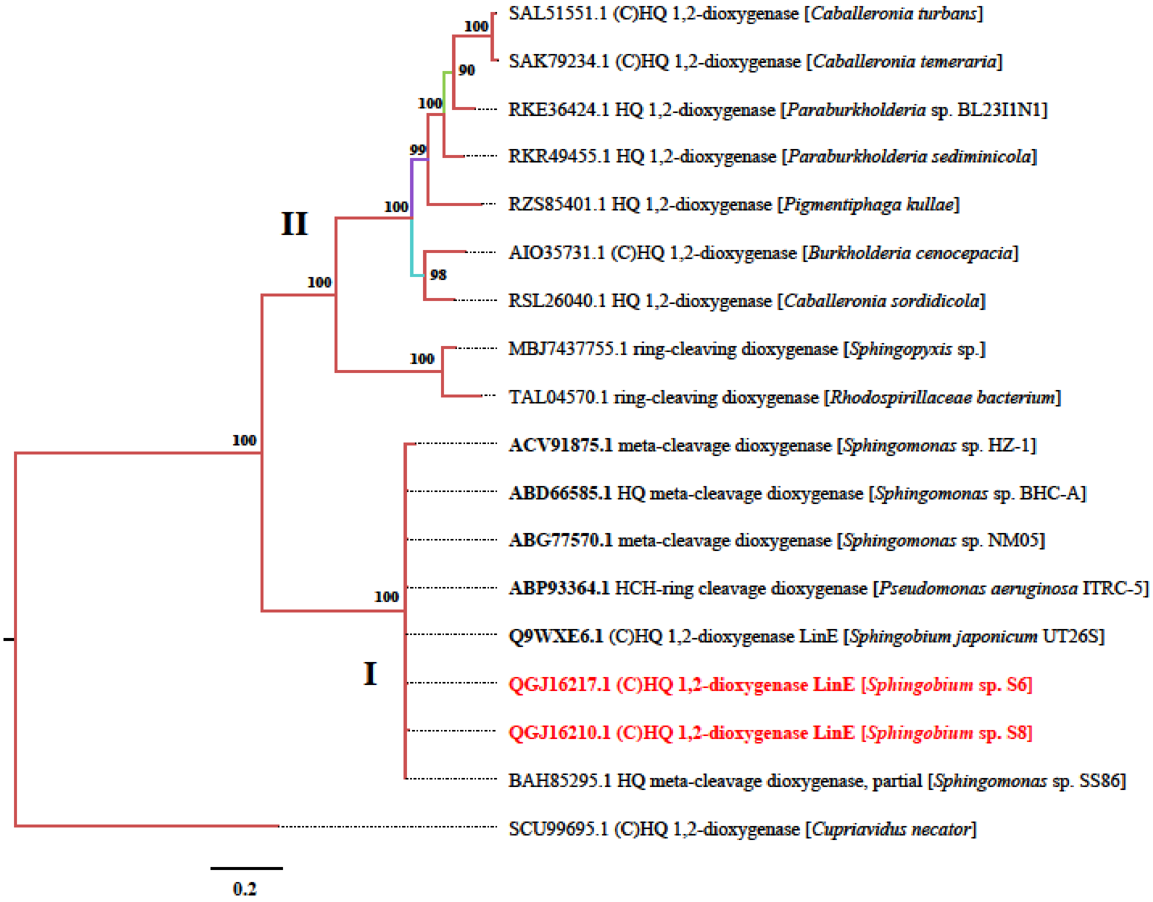

SAL51551.1 (C)HQ 1,2-dioxygenase [Caballeronia turbans]
SAK79234.1 (C)HQ 1,2-dioxygenase [Caballeronia temeraria]
RKE36424.1 HQ 1,2-dioxygenase [Paraburkholderia sp. BL23I1N1]
RKR49455.1 HQ 1,2-dioxygenase [Paraburkholderia sediminicola]
RZS85401.1 HQ 1,2-dioxygenase [Pigmentiphaga kullae]
AIO35731.1 (C)HQ 1,2-dioxygenase [Burkholderia cenocepacia]
RSL26040.1 HQ 1,2-dioxygenase [Caballeronia sordidicola]
MBJ7437755.1 ring-cleaving dioxygenase [Sphingopyxis sp.]
TAL04570.1 ring-cleaving dioxygenase [Rhodospirillaceae bacterium]
ACV91875.1 meta-cleavage dioxygenase [Sphingomonas sp. HZ-1]
ABD66585.1 HQ meta-cleavage dioxygenase [Sphingomonas sp. BHC-A]
ABG77570.1 meta-cleavage dioxygenase [Sphingomonas sp. NM05]
ABP93364.1 HCH-ring cleavage dioxygenase [Pseudomonas aeruginosa ITRC-5]
Q9WXE6.1 (C)HQ 1,2-dioxygenase LinE [Sphingobium japonicum UT26S]
QGJ16217.1 (C)HQ 1,2-dioxygenase LinE [Sphingobium sp. S6]
QGJ16210.1 (C)HQ 1,2-dioxygenase LinE [Sphingobium sp. S8]
BAH85295.1 HQ meta-cleavage dioxygenase, partial [Sphingomonas sp. SS86]
SCU99695.1 (C)HQ 1,2-dioxygenase [Cupriavidus necator]
100
90
100
99
100
98
100
100
100
100
II
I
0.2

Figure 20. Phylogenetic tree by MrBayes (v3.2.7) for (C)HQ 1,2-dioxygenase [(chloro) hydroquinone 1,2-dioxygenase] (LinE) protein sequences, also called meta-cleavage dioxygenases or ring-cleavage dioxygenases. LinE sequences from *Sphingobium* strains S6 and S8 are shown in red. **SCU99695.1**: (chloro) hydroquinone 1,2-dioxygenase (*Cupriavidus necator*) was used as the outgroup in rerooting the tree. Numbers indicated on the nodes are percentage posterior probabilities showing statistical support for each node. The scale bar below the tree shows the number of expected changes (or substitutions) for each site. The accession numbers of the reference strains (UT26S, BHC-A, NM05, ITRC-5, and HZ-1) were Q9WXE6.1, ABD66585.1, ABG77570.1, ABP93364.1, and ACV91875.1, respectively.

both *Sphingobium* strains S6 and S8 (accession numbers QGJ16217.1 and QGJ16210.1, respectively) clustered together with dioxygenases from HCH-degrading *Sphingomonas*/*Sphingobium* species, including the most closely similar *S. japonicum* UT26S. Short internal branches of approximately zero branch lengths and posterior probability of 100% were characteristic of the *Sphingobium*/*Sphingomonas* cluster **I**, a likely indication that LinE was highly conserved within this group.

4.2.4.6 LysR-type transcriptional regulator (LinR)

Translated LysR-type transcriptional regulator (LTTR), LinR protein sequences from *Sphingobium* sp. S6 (Protein ID: QGJ16218.1) and *Sphingobium* sp. S8 (Protein ID: QGJ16211.2) contained 301 and 293 amino acid residues, respectively (**Table 5**). From the MSA using eight transcriptional regulator sequences (**Fig. 21**), all transcriptional regulators have a highly conserved region spanning many residues with only a few notable variations at the N- and C-termini. The N-terminal end appeared to be highly variable among all the transcriptional regulators in the alignment, especially the region spanning residues 1 to 15. LinR sequence from *Sphingobium* sp. S6 (QGJ16218.1) was much more conserved than LinR sequence from *Sphingobium* sp. S8 (QGJ16211.2). Notable amino acid substitutions included Pro10 and Gln306 in the LinR sequence from *Sphingobium* sp. S6 (QGJ16218.1). Ser18 at the N-terminus and Cys295, Asp296, Arg297, Ser298, Gly299, and Thr300 at the C-terminus, were the notable amino acid substitutions in the LinR sequence from *Sphingobium* sp. S8 (QGJ16211.2). Other amino acid substitutions included Val163, Asp174, Thr195, Arg251, Asp252, and Thr255 in the rest of the transcriptional regulators.

LinR sequences from both *Sphingobium* sp. S6 (QGJ16218.1) and *Sphingobium* sp. S8 (QGJ16211.2) contained the DNA-binding LysR-type HTH domain (HTH_LYSR) spanning residues 10–67 in QGJ16218.1 and residues 1–53 in QGJ16211.2 at the N-terminal region, typical of all LTTRs. Moreover, this domain contained the characteristic

DNA-binding HTH-motif (VSAAARELDLPQPTASHGLA) spanned residues 27–46 and 13–32 in QGJ16218.1 and QGJ16211.2, respectively, and was conserved in all the transcriptional regulators in the alignment.

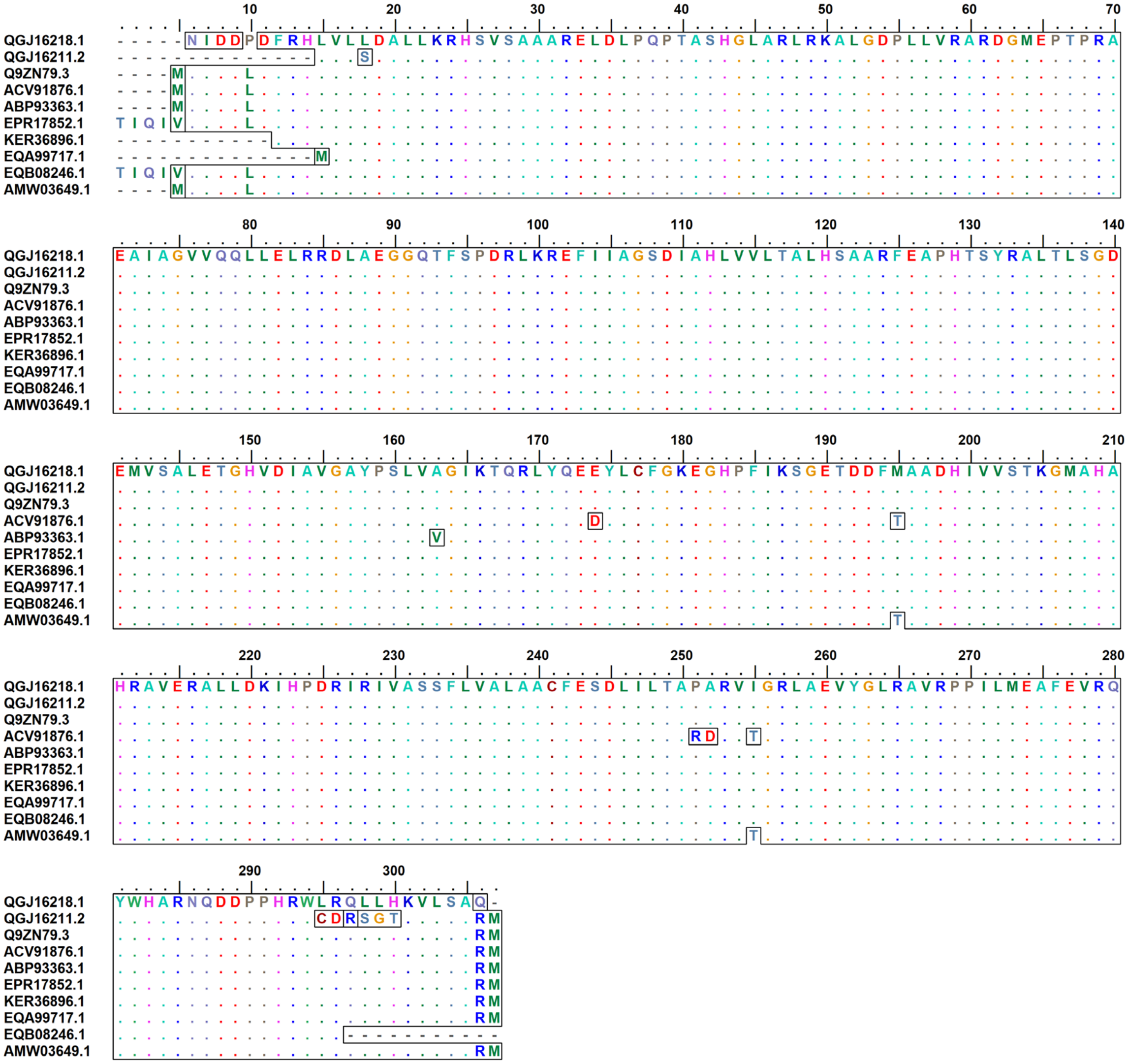


Figure 21. Multiple alignments of LysR-type transcriptional regulator (LinR) protein sequences homologous to LinR from *Sphingobium* sp. S6 (QGJ16218.1) and *Sphingobium* sp. S8 (QGJ16211.2). The conservation pattern for identical residues is represented by dots and mutated residues are boxed. The alignment graphics output was generated by the use of BioEdit v7.0.5.3.

Besides, LinR sequences from both *Sphingobium* sp. S6 (QGJ16218.1) and *Sphingobium* sp. S8 (QGJ16211.2) were identical to each other at 94.0% sequence identity and 94.7% sequence similarity. LinR sequence from *Sphingobium* sp. S6 (QGJ16218.1) was closely similar to the HTH-type transcriptional regulator (LinR) from *S. japonicum* UT26S (Q9ZN79.3) at 98.6% sequence identity and 99.0% sequence similarity. On the other hand, LinR sequence from *Sphingobium* sp. S8 (QGJ16211.2) was closely similar to the transcriptional regulator from *S. baderi* LL03 (EQA99717.1) at 97.2% sequence identity and 97.9% sequence similarity, respectively (**Appendix 1F**).

From the phylogenetic tree constructed using twenty (LysR family) transcriptional regulators, two separate clusters (**I & II**) were evident (**Fig. 22**). The first cluster (**I**) comprised of transcriptional regulators from HCH-degrading *Sphingobium* and/or *Sphingomonas* species. The second cluster (**II**) consisted of LysR family transcriptional regulators (LTTRs) sequences less similar to those of cluster **I** and included sequences from the genera *Sphingopyxis*, *Novosphingobium*, and *Pseudomonas*. LinR sequences of both *Sphingobium* strains S6 and S8 (accession numbers QGJ16218.1 and QGJ16211.2) clustered with those from the most closely similar strains *S. japonicum* UT26S (Q9ZN79.3), *S. chinhatense* IP26 (EPR17852.1), and *S. baderi* LL03 (EQA99717.1) in cluster **I**. Short internal branches of approximately zero branch lengths were characteristic of this cluster **I** (59 to 100% support values).

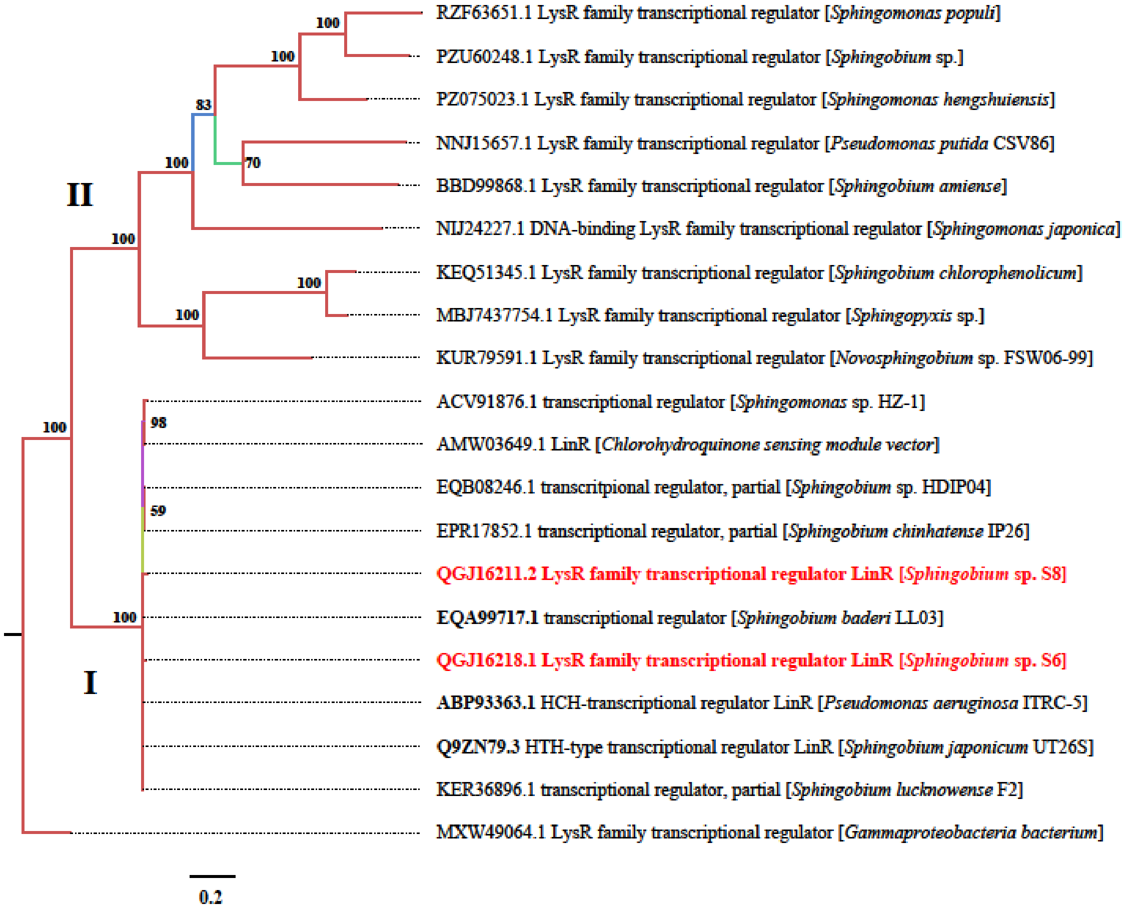

100
100
83
100
70
II
100
100
100
100
98
59
100
I
RZF63651.1 LysR family transcriptional regulator [Sphingomonas populi]
PZU60248.1 LysR family transcriptional regulator [Sphingobium sp.]
PZ075023.1 LysR family transcriptional regulator [Sphingomonas hengshuiensis]
NNJ15657.1 LysR family transcriptional regulator [Pseudomonas putida CSV86]
BBD99868.1 LysR family transcriptional regulator [Sphingobium amiense]
NIJ24227.1 DNA-binding LysR family transcriptional regulator [Sphingomonas japonica]
KEQ51345.1 LysR family transcriptional regulator [Sphingobium chlorophenolicum]
MBJ7437754.1 LysR family transcriptional regulator [Sphingopyxis sp.]
KUR79591.1 LysR family transcriptional regulator [Novosphingobium sp. FSW06-99]
ACV91876.1 transcriptional regulator [Sphingomonas sp. HZ-1]
AMW03649.1 LinR [Chlorohydroquinone sensing module vector]
EQB08246.1 transcritpional regulator, partial [Sphingobium sp. HDIP04]
EPR17852.1 transcriptional regulator, partial [Sphingobium chinhatense IP26]
QGJ16211.2 LysR family transcriptional regulator LinR [Sphingobium sp. S8]
EQA99717.1 transcriptional regulator [Sphingobium baderi LL03]
QGJ16218.1 LysR family transcriptional regulator LinR [Sphingobium sp. S6]
ABP93363.1 HCH-transcriptional regulator LinR [Pseudomonas aeruginosa ITRC-5]
Q9ZN79.3 HTH-type transcriptional regulator LinR [Sphingobium japonicum UT26S]
KER36896.1 transcriptional regulator, partial [Sphingobium lucknowense F2]
MXW49064.1 LysR family transcriptional regulator [Gammaproteobacteria bacterium]
0.2

Figure 22. Phylogenetic tree by MrBayes (v3.2.7) for LysR family transcriptional regulator (LinR) protein sequences. LinR sequences from *Sphingobium* strains S6 and S8 are shown in blue. **MXW49064.1**: LysR family transcriptional regulator (*Gammaproteobacteria bacterium*) was used as outgroup in rerooting the tree. Numbers indicated on the nodes are the percent posterior probabilities showing statistical support for each node. The scale bar below the tree shows the number of expected changes (or substitutions) for each site. The accession numbers of the reference strains (UT26S, LL03, and ITRC-5) were Q9ZN79.3, EQA99717.1, and ABP93363.1, respectively.

4.2.4.7 2,5-DDOL dehydrogenase (LinX)

Translated protein sequences of the LinC-like 2,5-DDOL dehydrogenase (LinX) from *Sphingobium* sp. S6 (Protein ID: QGJ16219.1) and *Sphingobium* sp. S8 (Protein ID: QGJ16212.1) each consisted of 250 amino acid residues (**Table 5**). From the MSA using eight dehydrogenase sequences (**Fig. 23**), all residues were conserved, except for a few mutated residues including Val12, Asp80, Gly81, and Ala238. No apparent amino acid substitutions in the LinX sequences from both *Sphingobium* sp. S6 (QGJ16219.1) and *Sphingobium* sp. S8 (QGJ16212.1).

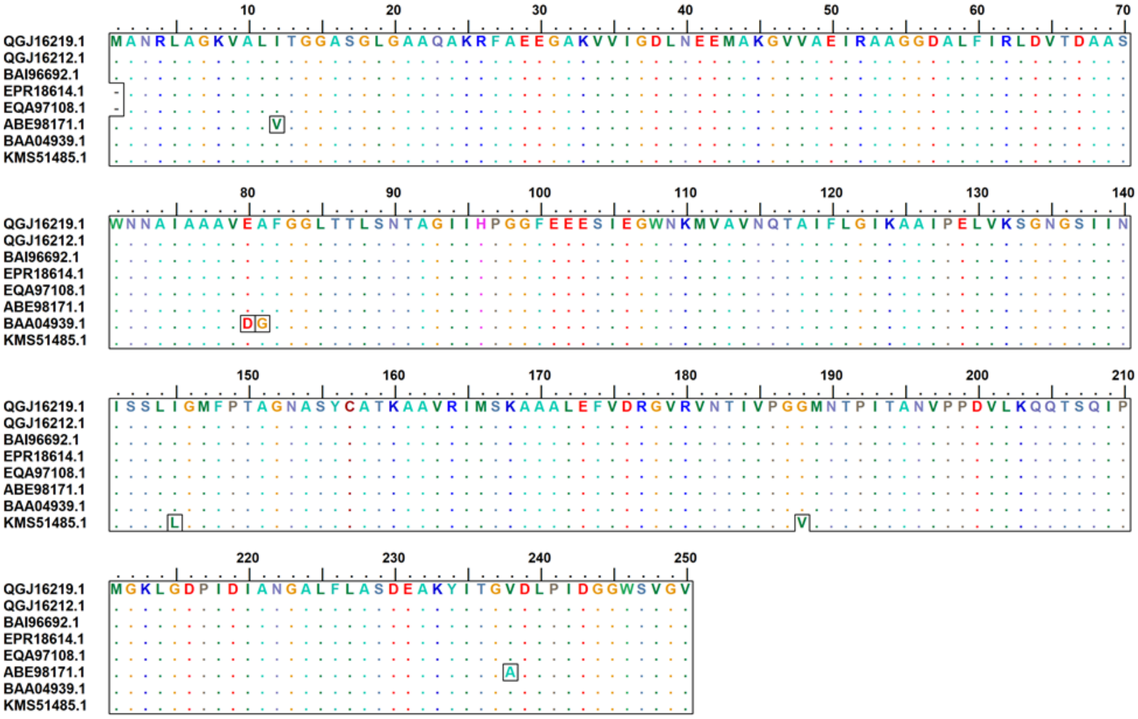

10
20
30
40
50
60
70
QGJ16219.1
QGJ16212.1
BAI96692.1
EPR18614.1
EQA97108.1
ABE98171.1
BAA04939.1
KMS51485.1
MANRLAGKVALITGGASGLGAAQAKRFAEEGAKVVIGDLNEEMAKGVVAEIRAAGGDALFIRLDVTDAAS
-
-
V
80
90
100
110
120
130
140
WNNAIAAAVEAFGGLTTLSNTAGIIHPGGFEEESIEGWNKMVAVNQTAIFLGIKAAIPELVKSGNGSIIN
D G
150
160
170
180
190
200
210
ISSLIGMFPTAGNASYCATKAAVRIMSKAAALEFVDRGVRVNTIVPGGMNTPITANVPPDVLKQQTSQIP
L
V
220
230
240
250
MGKLGDPIDIANGALFLASDEAKYITGVDLPIDGGWSVGV
A

Figure 23. Multiple alignments of 2,5-DDOL dehydrogenase (LinX) protein sequences homologous to LinX from *Sphingobium* sp. S6 (QGJ16219.1) and *Sphingobium* sp. S8 (QGJ16212.1). The conservation pattern for identical residues is represented by dots and mutated residues are boxed. The alignment graphics output was generated by the use of BioEdit v7.0.5.3.

Based on ScanProsite (https://prosite.expasy.org/scanprosite/) search for conserved domains (CDs) {Citation}, both LinX sequences (QGJ16219.1 and QGJ16212.1) contained the short-chain dehydrogenases/reductases (SDR) family signature (ADH_SHORT) spanning residues 53–81 in QGJ16219.1 and residues 143–171 in QGJ16212.1. They also contained active site residues threonine and tyrosine at positions 66 and 156, respectively. LinX sequences from both *Sphingobium* sp. S6 (QGJ16219.1) and *Sphingobium* sp. S8 (QGJ16212.1) were identical to each other (100% sequence identity and similarity). They were very similar (100% sequence identity and similarity) to 2,5-DDOL dehydrogenase (LinX) from *S. japonicum* UT26S (BAI96692.1. Moreover, the LinC-like (LinX) sequences from both *Sphingobium* sp. S6 (QGJ162191.1) and *Sphingobium* sp. S8 (QGJ16212.1) were identical to the LinC sequences from *Sphingobium* sp. S6 (QGJ16215.1) and Sphingobium sp. S8 (QGJ16208.2) at 31 % sequence identity and 50% sequence similarity (**Appendix 1F)**.

From the phylogenetic tree constructed using twenty-three dehydrogenase sequences, three distinct clusters were evident (**Fig. 24**). Cluster (**I**) comprised of dehydrogenase sequences from HCH-degrading *Sphingomonas* or *Sphingobium* species whose LinX was very much conserved as shown by the short internal branches of zero branch lengths. LinX sequences from *Sphingobium* strains S6 and S8 (accession numbers QGJ16219.1 and QGJ16212.1, respectively) belonged to this cluster, together with the most closely similar *S. japonicum* UT26S (BAI96692.1). The second cluster (**II**) is composed of dehydrogenases that are less similar but closely related to those from cluster **I**, including the second type of dehydrogenase (LinX2) from *S. indicum* B90A. The third cluster (**III**) consisted of dehydrogenases mainly from the genera *Sphingopyxis* and *Novosphingobium*, which are less similar and more separate from those of clusters **I** and **II**.

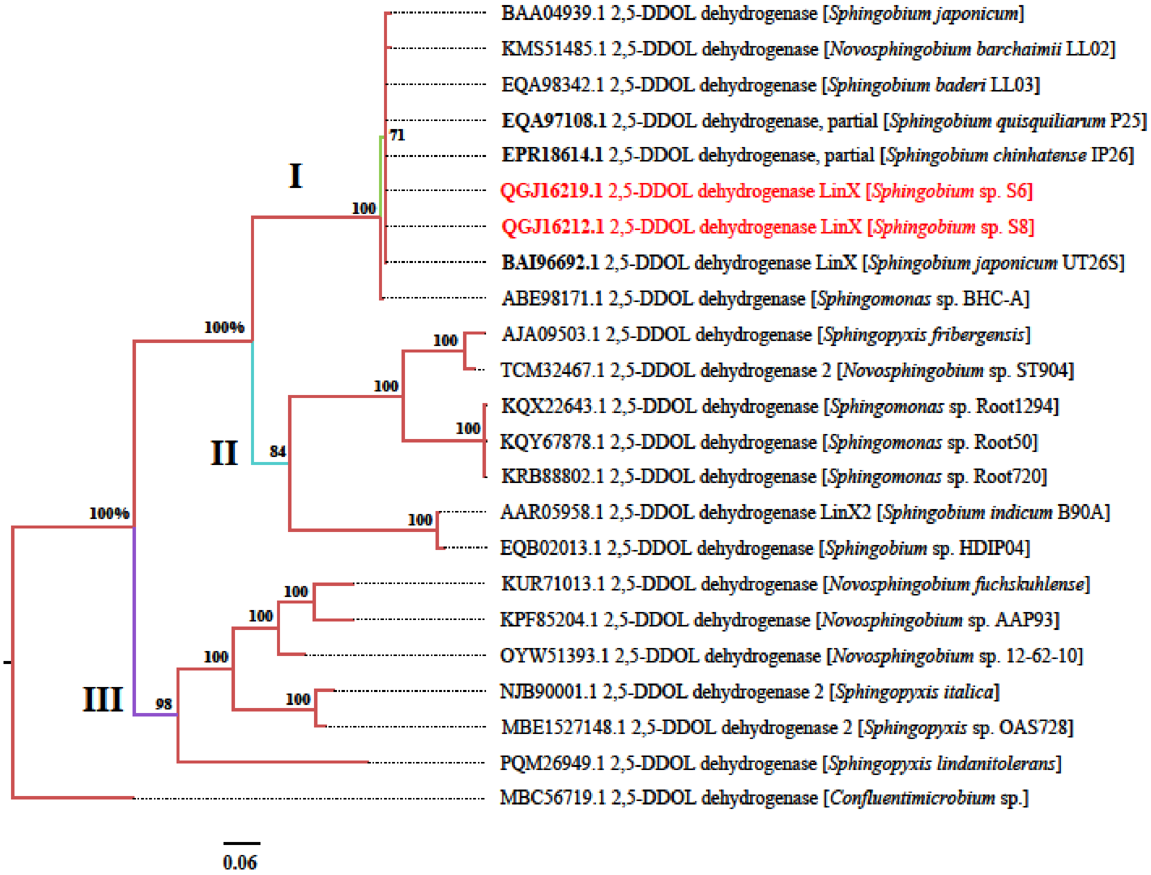

BAA04939.1 2,5-DDOL dehydrogenase [Sphingobium japonicum]
KMS51485.1 2,5-DDOL dehydrogenase [Novosphingobium barchaimii LL02]
EQA98342.1 2,5-DDOL dehydrogenase [Sphingobium baderi LL03]
EQA97108.1 2,5-DDOL dehydrogenase, partial [Sphingobium quisquiliarum P25]
EPR18614.1 2,5-DDOL dehydrogenase, partial [Sphingobium chinhatense IP26]
QGJ16219.1 2,5-DDOL dehydrogenase LinX [Sphingobium sp. S6]
QGJ16212.1 2,5-DDOL dehydrogenase LinX [Sphingobium sp. S8]
BAI96692.1 2,5-DDOL dehydrogenase LinX [Sphingobium japonicum UT26S]
ABE98171.1 2,5-DDOL dehydrgenase [Sphingomonas sp. BHC-A]
AJA09503.1 2,5-DDOL dehydrogenase [Sphingopyxis fribergensis]
TCM32467.1 2,5-DDOL dehydrogenase 2 [Novosphingobium sp. ST904]
KQX22643.1 2,5-DDOL dehydrogenase [Sphingomonas sp. Root1294]
KQY67878.1 2,5-DDOL dehydrogenase [Sphingomonas sp. Root50]
KRB88802.1 2,5-DDOL dehydrogenase [Sphingomonas sp. Root720]
AAR05958.1 2,5-DDOL dehydrogenase LinX2 [Sphingobium indicum B90A]
EQB02013.1 2,5-DDOL dehydrogenase [Sphingobium sp. HDIP04]
KUR71013.1 2,5-DDOL dehydrogenase [Novosphingobium fuchskuhlense]
KPF85204.1 2,5-DDOL dehydrogenase [Novosphingobium sp. AAP93]
OYW51393.1 2,5-DDOL dehydrogenase [Novosphingobium sp. 12-62-10]
NJB90001.1 2,5-DDOL dehydrogenase 2 [Sphingopyxis italica]
MBE1527148.1 2,5-DDOL dehydrogenase 2 [Sphingopyxis sp. OAS728]
PQM26949.1 2,5-DDOL dehydrogenase [Sphingopyxis lindanitolerans]
MBC56719.1 2,5-DDOL dehydrogenase [Confluentimicrobium sp.]
I
II
III
71
100
100%
100
100
100
84
100
100%
100
100
100
100
98
100
0.06

Figure 24. Phylogenetic tree by MrBayes (v3.2.7) for 2,5-DDOL dehydrogenase (LinX) protein sequences. LinX sequences from *Sphingobium* strains S6 and S8 are shown in red. **MBC56719.1**: 2,5-DDOL dehydrogenase (*Confluentimicrobium sp.*) was used as outgroup. Numbers indicated on the nodes are the percent posterior probabilities showing statistical support for each node. The scale bar below the tree shows the number of expected changes (or substitutions) for each site. The accession numbers of the reference strains UT26S, IP26, and P25 were BAI96692.1, EPR18614.1, and EQA97108.1, respectively.

4.3 Determining the three-dimensional (3D) structures of key enzymes in the HCH degradation pathway via comparative modeling

4.3.1 HCH-dehydrogenase (LinA)

4.3.1.1 Homology modeling of LinA

The 3D structures of target LinA from *Sphingobium* sp. S6 (protein ID: QGJ16213.1) and *Sphingobium* sp. S8 (protein ID: QGJ16206.1) were modeled by using 3a76 (crystal structure of LinA from *Sphingobium japonicum* UT26) as a template. The template shared the highest sequence identity with the target LinA from *Sphingobium* sp. S6 (98.72%) and *Sphingobium* sp. S8 (97.33%) among existing structures. Further, both LinA sequences matched along the entire chain length and there are no gaps in the alignment. Also, the target LinA proteins (QGJ16213.1 and QGJ16206.1) and template belonged in the same nuclear transport factor 2 (NTF2)-like) protein superfamily. The template structure had been elucidated at high resolution (2.25 Å) and shows good quality parameters (e.g. $R_{free} = 0.276$). Thus, the LinAS6 and LinAS8 models (**Fig. 25**), each comprising of three chains A, B, and C were significantly similar to the template (RMSDs of 0.069 Å and 0.074 Å, respectively).

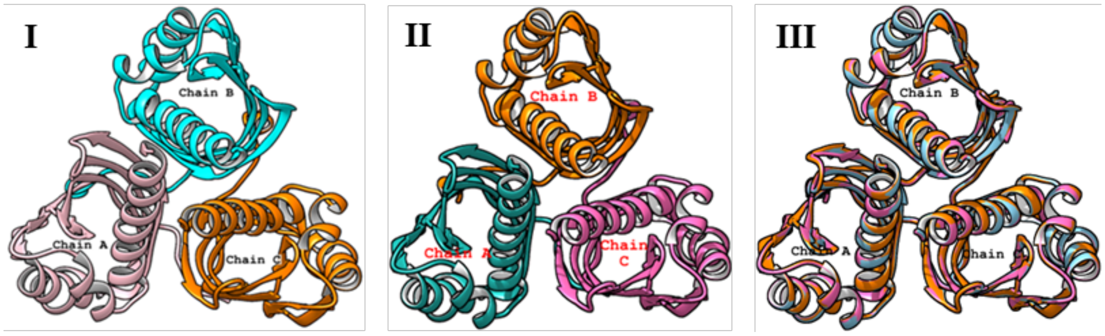


Figure 25. Theoretical 3D structures (top view) of LinA from *Sphingobium* sp. S6 and *Sphingobium* sp. S8. **I**) LinAS6 homotrimer. Each subunit is rendered in a separate color with chain A in pink, chain B in cyan, and chain C in orange. **II**) LinAS8 homotrimer. Each subunit is rendered in a separate color with chain A in light sea green, chain B in orange, and chain C in hot pink. **III**) LinAS6 (orange) and LinAS8 (hot pink) model superimposed with the template, 3A76 (cyan). The figures were rendered using UCSF Chimera v1.15.

Moreover, there was no significant deviation in the backbone conformation of the LinA protomers of both LinAS6 and LinAS8 models, with RMSD <0.5 Å for 150 Cα atoms. The LinAS6 and LinAS8 models are considered rather reliable, as indicated by the quality estimates (**Table 6**). Since the template matched along with the entire length of each target LinA and exhibited high sequence identity, the scaffolding and conserved parts are likely of high quality. The structures were refined by energy minimization to solve clashes and non-favorable stereochemistry (Vihinen, 2021).

Table 6. Quality estimate parameters used to assess the reliability of homology models generated by SWSS-MODEL, including coverage and percent (%) sequence identity and similarity of the target–template alignment

3D Model	Oligo-State	Template (PDB ID)	Coverage (%)	% Seq Identity/ Similarity	QSGE /GMQE	QMEAN (Z-score)	RMSD (Å)
LinAS6	Trimer	3a76	100	98.72/62	0.53/0.90	-1.46	0.069
LinAS8	Trimer	3a76	100	97.33/61	0.58/0.91	-1.83	0.074
LinBS6	Monomer	1mj5	100	97.97/62	0.00/0.99	0.42	0.062
LinBS8	Monomer	1mj5	100	96.96/61	0.00/0.99	0.09	0.074
LinCS6	Tetramer	5x8h	98	40/40	0.84/0.77	-0.48	0.664
LinCS8	Tetramer	5x8h	98	39/40	0.88/0.74	-0.55	0.664
LinDS6	Dimer	7aia	74	25/32	0.32/0.41	-5.33	3.073
LinDS8	Dimer	7aia	74	24.90/32	0.33/0.41	-4.65	3.110
LinES6	Dimer	4huz	98	52.08/47	0.56/0.87	-1.03	0.349
LinES8	Dimer	4huz	100	51.58/47	0.57/0.88	-1.25	0.538

Each protomer of LinAS6 and LinAS8 models comprise of a conically shaped α+β barrel fold consisting of a curved mixture of six β-sheets of the strand order β2–β1–β6–β5– β4–β3 and four α-helices on each side, respectively as previously reported in the crystal structure of LinA by Okai *et al*. (2010). However, an helix (η2) spanning residues 139 to 153 at the C-terminal portion and directly involved in the interaction with the β-strand (β6) of the adjacent subunit reported in the crystal structure of LinA, was present only in chain C of both LinAS6 and LinAS8 models and spanned residues 139-152. Interactions between subunits, especially helix α1, strands β3-β4-β6, and helix η2, are thought to stabilize the LinA homotrimer whereas interactions among the hydrophobic surfaces of the protomers stabilize the core region of LinA trimer (Okai *et al*., 2010). Further, interactions between Ly26 and Asp93′, Asp19 and Arg79′ (where prime indicates a different subunit) forms two salt bridges that also stabilize the LinA trimer and are conserved in proteins belonging to the alpha/beta barrel fold (Okai *et al*., 2010).

4.3.1.2 Validation of predicted LinA models

Ramachandran plot analyses of predicted LinA models revealed 369 (93.7%) residues occur in the most favored regions, 25 (6.3%) residues in the additionally allowed regions and none in the generously allowed and disallowed regions of the LinAS6 model. On the other hand, the LinAS8 model contained 361 (93.5%) residues in the most favored regions, 25 (6.5%) residues in the additionally allowed regions and none in the generously allowed and disallowed regions. Ramachandran plots of LinAS6 and LinAS8 models (**Fig. 26**) and the evaluation statistics of each model (**Table 7**) were as shown.

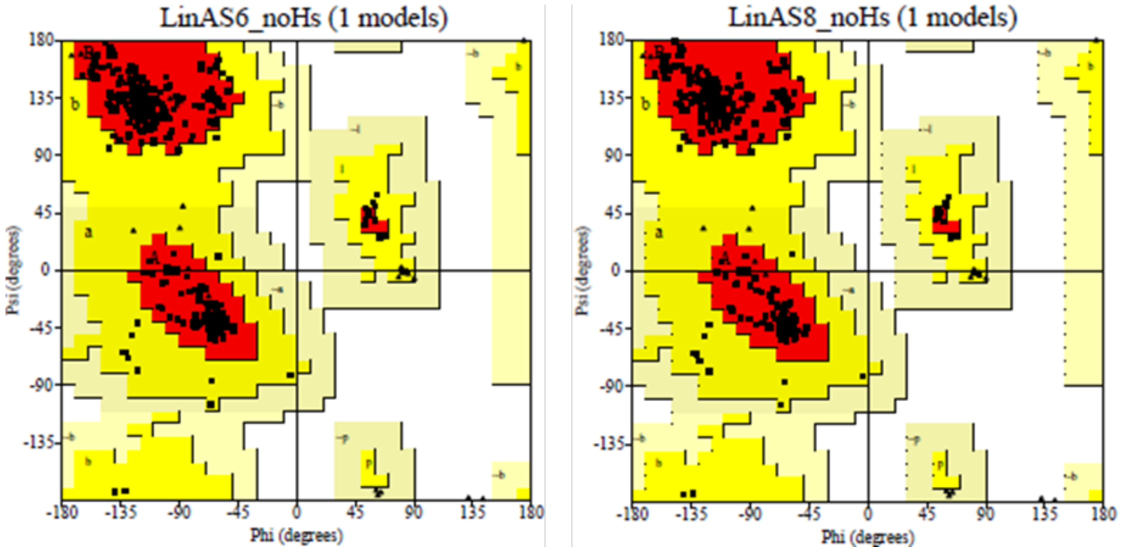


Figure 26. Ramachandran plot by PROCHECK for LinAS6 and LinAS8 models showing the φ–ψ distribution for the different regions. The core regions (marked A, B, L), additionally allowed regions (marked a, b, l, p), generously allowed regions (marked ~a, ~b, ~l, ~p), and disallowed regions are shown in red, yellow, grey and white colors, respectively. Non-proline and non-glycine residues are represented by black squares and glycine residues by black triangles.

Table 7. Validation parameters by PROCHECK, MOLPROBITY, VERIFY3D, PROSAII, and ERRAT for evaluating the structural quality of the 3D homology models generated by SWISS-MODEL

Validation statistic	**Parameters used in the evaluation of the model**	**Theoretical three-dimensional (3D) model**			
		LinAS6	**LinAS8**	**LinBS6**	**LinBS8**
PROCHECK	Residues in most favoured regions [A, B, L], %	369, 93.7%	361, 93.5%	226, 90.0%	222, 90.6%
	Residues in additionally allowed regions [a, b, l, p], %	25, 6.3%	25, 6.5%	24, 9.6%	22, 9.0%
	Residues in generously allowed regions [~a, ~b, ~l, ~p], %	0, 0.0%	0, 0.0%	1, 0.4%	1, 0.4%
	Residues in disallowed regions, %	0, 0.0%	0, 0.0%	0, 0.0%	0, 0.0%
	Procheck G-factor[a] (φ/ψ) Z-score[g]	0.04	0.04	-0.04	-0.04
	Procheck G-factor[a] (all dihedral angles) Z-score[g]	-0.59	-0.59	-0.00	-0.00
	Overall Procheck G-factor	-0.22	-0.17	-0.09	-0.13
MOL-PROBITY	MolProbity score^	1.66	1.62	1.08	0.96
	MolProbity Clashscore	0.77	0.90	0.96	0.99
VERIFY3D	3D–1D score ≥ 0.2 (%), Z-score[g]	73.67%, -3.05	69.13%, -2.73	100%, -2.89	100%, -2.89
PROSAII	ProSAII (-ve) Z-score[g]	-1.12	-1.57	-0.04	0.00
ERRAT	Overall quality factor (%)	90.02	90.38	94.74	99.28

[a] Residues selected using the parameter; S(phi)+S(psi)>=1.8, for dihedral angle order. [g] Determined by comparing the standard deviation and mean of sets of 252 X-ray crystal structures of <500 residues; a positive value indicates a 'better' score.
Generated using PSVS (Protein Structure Validation Suite) v1.5 and SAVES v6.

The distribution of torsional phi and psi (φ and ψ) and dihedral angles within the LinAS6 and LinAS8 models was the same with normalized of Z-scores of 0.04 and -0.59, respectively. However, the overall G-factor of LinAS6 and LinAS8 models was -0.22 and -0.17, respectively all within acceptable range (≤ -0.50). From the 3D profile analysis, only 73.67% of residues in the LinAS6 model had averaged 3D-1D scores of 0.2 or higher, all normalized to a Z-score of -3.05 and the 3D profile was as shown (**Fig. 27A**). Similarly, only 69.13% of residues had averaged 3D-1D scores of 0.2 or higher in the LinAS8 model, with a normalized Z-score of -2.73 and the 3D profile was as shown (**Fig. 27B**). The normalized energy Z-scores by ProSAII for LinAS6 and LinAS8 models were determined to be -1.12 and -1.57, respectively. Their ProSAII (-ve) scores were plotted over a window average of seven residues and the energy graphs were as shown (**Fig. 28A & B**).

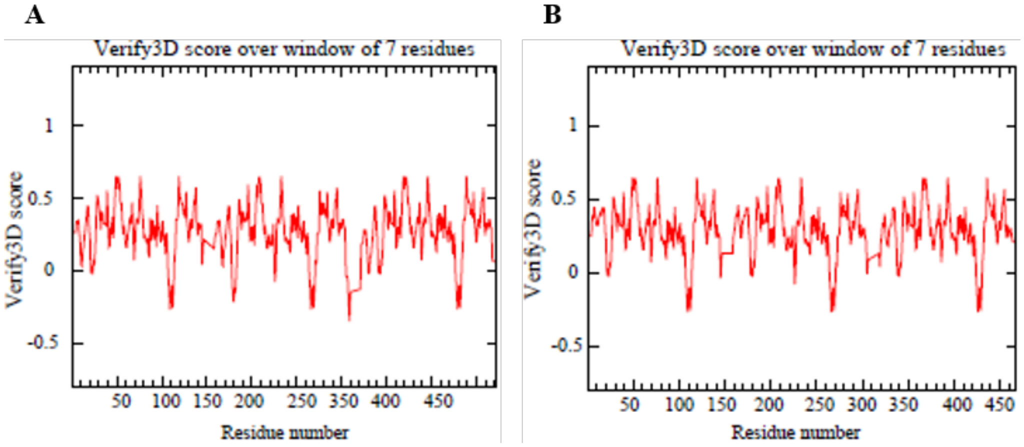


Figure 27. Verify3D profiles of the LinAS6 and LinBS8 models. **A**) 3D profile of LinAS6 model showing the Verify3D scores plotted against residue number in a window average of 7 residues. **B**) 3D profile of LinAS8 model showing the Verify3D scores plotted against residue number in a window average of 7 residues. Positive scores indicate correctly modeled segments of the 3D structure.

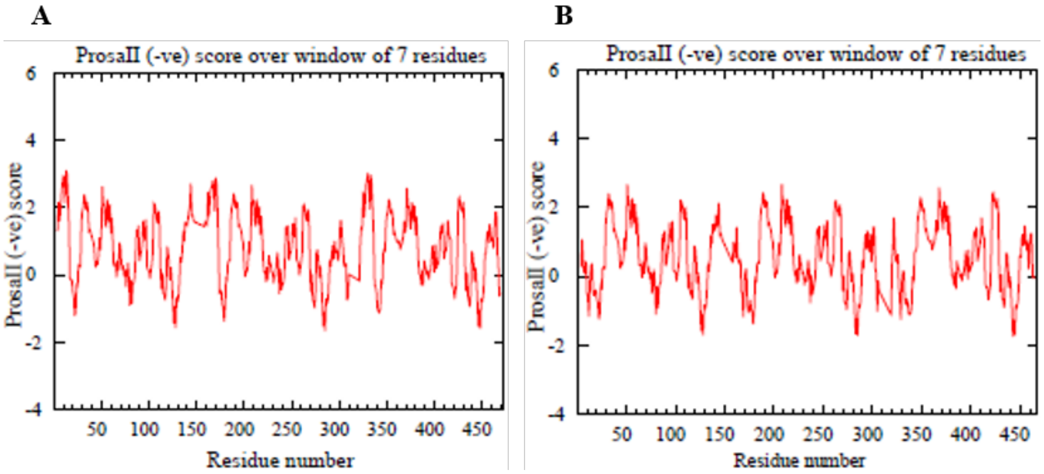


Figure 28. ProSAII profiles of LinAS6 and LinAS8 models. **A**) Energy graph of LinAS6 model showing ProSAII (-ve) scores plotted against residue number over a window average of 7 residues. **B**) Energy graph of LinAS6 model showing ProSAII (-ve) scores plotted against residue number over a window average of 7 residues. Positive scores indicate correctly modeled segments of the 3D structure.

The overall quality factor of the LinA protomers of both LinAS6 and LinAS8 models was the same (i.e. 90%). However, suspected regions of error due to mistraced and/or misregistered atoms (exceeding the 95% and 99% confidence limits) were apparent in the protomers of both LinAS6 and LinAS8 models and were highlighted in yellow and red, respectively. The ERRAT (v2.0) plot for chain A of LinAS6 and LinAS8 models and its overall quality factor was as shown (**Fig. 29A & B**).

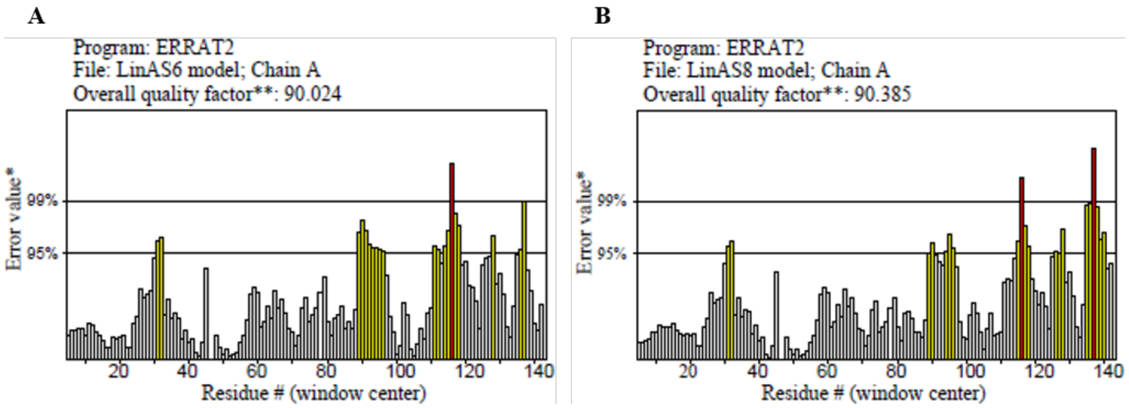


Figure 29. ERRAT profile plots for the LinA protomer (chain A) of LinAS6 and LinAS8 models. **A**) ERRAT profile plot of LinAS6 (chain A). **B**) ERRAT profile plot of LinAS8 (chain A). The overall quality factor (%) for the LinA protomer of each LinA model is indicated above the plot. Regions of error exceeding the 95% confidence limit for rejection are highlighted yellow and those exceeding the 99% confidence are highlighted red. *Two lines drawn on the error axis show the confidence for rejection of regions exceeding the error value. **Overall quality factor of the model is expressed as the percentage of the protein whose calculated error value is below the 95% limit for rejection. Average overall quality factor for good structures of high resolution is 95% or more whereas that of lower resolution structures (2.5–3Å) is around 91%.

4.3.1.3 Solvent accessibility and secondary structure of LinA

Secondary structural elements of both LinAS6 and LinAS8 protomers consisted of six beta-strands, four alpha-helices, and one 3_{10} helix spanning residues in the following order: α1(2–26), α2(29–33), β1(36–44), β2(48–51), α3(52–62), α4(64–67), β3(68–82), β4(87–100), η1(102–104), β5(105–120), and β6(123–138) as shown (**Fig. 30**). A structural alignment of LinAS6 and LinAS8 models with the crystal structures of three LinA proteins (PDB IDs: 3A76, 3S5C, and 5KVB) most homologous to LinA from both

Sphingobium sp. S6 and *Sphingobium* sp. S8 showed their secondary structures were conserved.

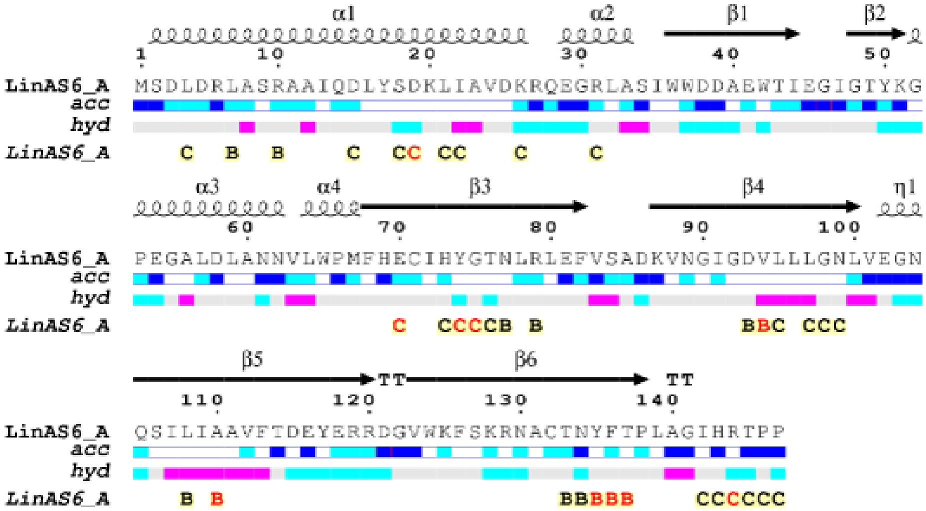


Figure 30. The flat figure for LinA sequence from *Sphingobium* sp. S6 adorned with secondary structural elements (β-strands, helices, and turns represented by arrows, squiggles, and TT letters, respectively). The first and second bar underneath the sequence indicates solvent accessibility (blue for accessible, cyan for intermediate, and white for buried) and hydropathy (cyan for hydrophilic, grey for neutral, and pink for hydrophobic), respectively. Letters at the bottom represent crystallographic, protein-protein, and protein-ligand contacts. The output was generated by ENDscript 2.0 (https://endscript.ibcp.fr/ESPript/ENDscript/).

4.3.2 Active binding pocket and 3D protein fold of LinA

4.3.2.1 Modeled LinA from *Sphingobium* sp. S6

Both LinAS6 model and the template (3a76) adopt the same 3D protein fold and have similar binding sites for each of the LinA protomers (chains A, B, and C) (**Fig. 31**). The catalytic residues in the active site of LinA, also called "catalytic triad" consisting of Lys20, Asp25, and His73, and the hydrophobic residues (Trp42, Leu64, Leu96, and Phe113) surrounding them were found to be conserved in the LinA from S6. The observed mutations (Arg144 and Thr145) in the LinA sequence from *Sphingobium* sp. S6 were localized at the C-terminal region (**Fig. 32**).

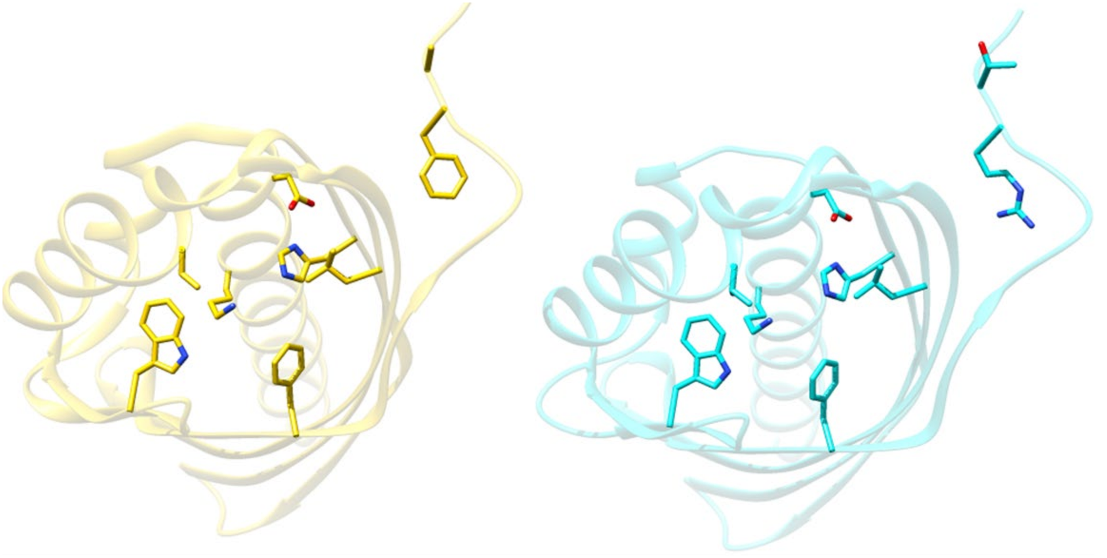


Figure 31. 3D protein folds and active binding pocktes of LinA (chain A) protomers of LinAS6 and template (3a76). The template (gold) and model (cyan) are shown side by side including the active site residues (depicted as colored sticks in gold and cyan, respectively).

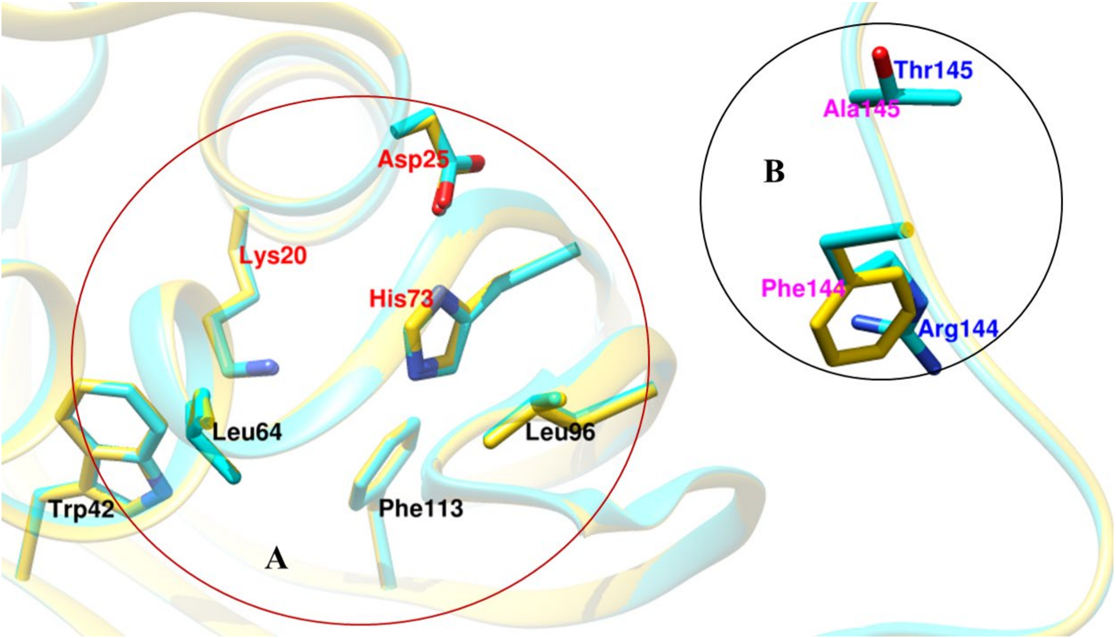


Figure 32. Substrate binding pocket of LinA (chain A) protomer of LinAS6. A) Active binding site showing the catalytic residues (red) and the surrounding hydrophobic residues (black). B) Mutated residues between the template and LinA model (shown in pink and blue, respectively) are outside the active site. The template (gold) and LinA model of *Sphingobium* sp. S6 (cyan) have been superimposed and the residues are represented as colored sticks (gold and cyan, respectively).

4.3.2.2 Modeled LinA from *Sphingobium* sp. S8

Both the template (3a76) and LinAS8 model adopt the same 3D protein fold and have similar binding sites in each of the LinA protomers (chains A, B, and C) (**Fig. 33**). Moreover, the catalytic residues in the active site of LinA, otherwise called "catalytic triad" consisting of Lys20, Asp25, and His73, and the hydrophobic residues (Trp42, Leu64, Leu96, and Phe113) surrounding them were found to be conserved in the LinA from S8. The observed mutations (Phe4, Gly10, Ser11 and Asn13) in the LinA sequence from *Sphingobium* sp. S8 were located at the N-terminal region (**Fig. 34**).

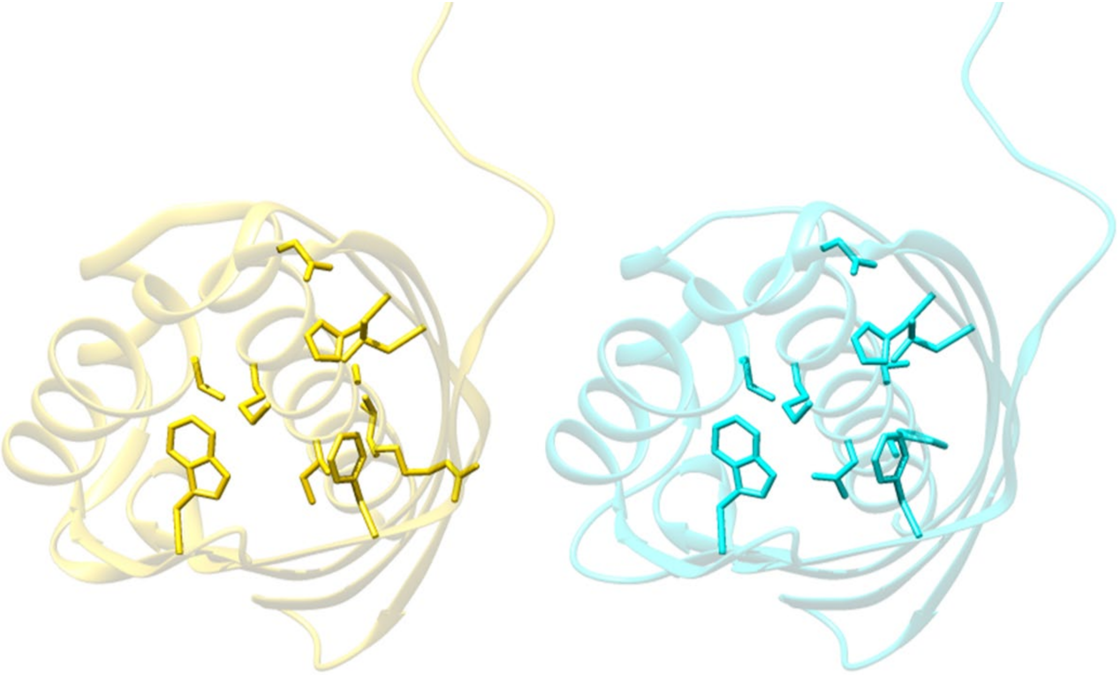


Figure 33. 3D protein folds and active binding pockets of LinA (chain A) protomers of LinAS8 and template (3a76). The template (gold) and model (cyan) are shown side by side including the active site residues (depicted as yellow and blue sticks, respectively).

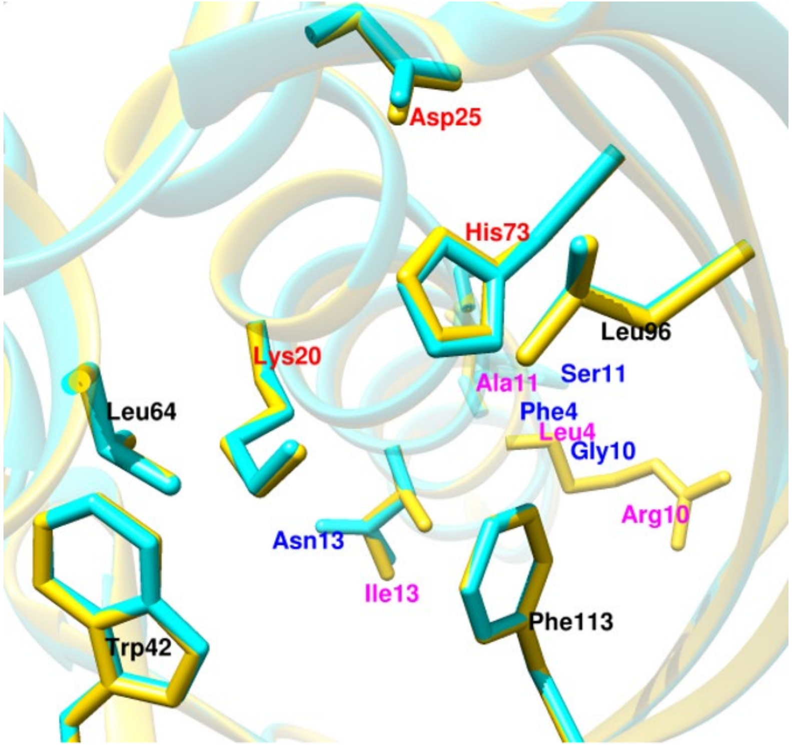


Figure 34. Substrate binding pocket of LinA (chain A) protomer of LinAS8. Catalytic residues in the active site and the surrounding hydrophobic residues are labeled in red and black, respectively. Some of the mutated residues between the template (pink) and LinA model (blue) are shown. The template (gold) and LinA model of *Sphingobium* sp. S6 (cyan) have been superimposed and the residues are represented as colored sticks (gold and cyan, respectively).

4.3.3 Haloalkane dehalogenase (LinB)

4.3.3.1 Homology modeling of LinB

The 3D structures of LinB from *Sphingobium* sp. S6 (protein id: QGJ16214.1) and *Sphingobium* sp. S8 (protein ID: QGJ16207.1) were constructed using 1mj5 [crystal structure of 1,3,4,6-tetrachloro-1,4-cyclohexadiene (1,4-TCDN) hydrolase (LinB) from *Sphingobium japonicum* UT26] as a template. The template shared the highest sequence identity and similarity with the target LinB from *Sphingobium* sp. S6 (97.97% and 62%, respectively) and *Sphingobium* sp. S8 (96.96% and 61%, respectively) among existing structures. There were no gaps in the alignment and the sequences matched along the entire chain length. The template had been determined at high resolution (0.95 Å) and showed good quality parameters (e.g. $R_{free} = 0.141$). Further, the target LinB protein and the template belonged in the same α/β hydrolase family. The LinBS6 and LinBS8 models were significantly similar to the template (RMSDs of 0.062 Å and 0.074 Å, respectively) (**Fig. 35**). The LinB models were considered to be rather reliable, as indicated by the quality estimates (**Table 6**). Because the template matched along with the entire LinB sequences and had high sequence identity, the scaffolding and conserved parts are likely of high quality. The structures were refined by energy minimization to solve clashes and non-favorable stereochemistry (Vihinen, 2021).

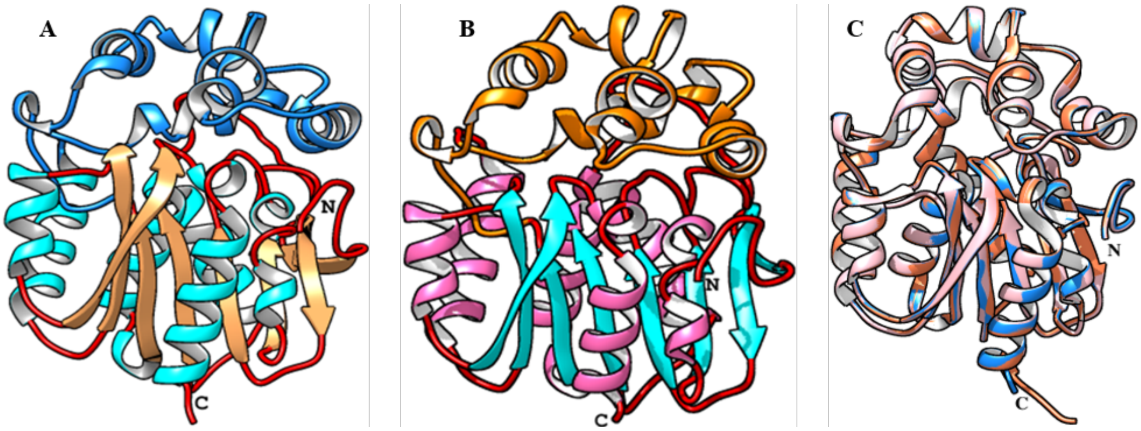


Figure 35. Theoretical 3D structures of LinB from *Sphingobium* sp. S6 and *Sphingobium* sp. S8. A) Monomeric LinBS6 structure. Helices and loops of the core domain are shown in cyan and red, respectively whereas the helices and loops of the cap domain are in dodger blue. Beta strands are shown in tan. B) Monomeric LinBS8 structure. Helices and loops of the core domain are shown in hot pink and red, respectively whereas the helices and loops of the cap domain are in orange. Beta strands are shown in cyan. C) LinBS6 (dodger blue) and LinBS8 (pink) models superimposed with the template, 1mj5 (coral). Letters N and C show the N- and C-terminal ends, respectively. Rendering of the structures was done using UCSF Chimera v1.15.

Both LinBS6 and LinBS8 models adopt a typical alpha/beta-hydrolase fold that consists of core (main) and cap domains, similar to other structurally known dehalogenases (Damborský & Koča, 1999; Marek *et al*., 2000). The core (main) domain spanned residues 3–133 and 214–296 and is conserved in all alpha/beta-hydrolase fold proteins (Marek *et al*., 2000; Novak *et al*., 2014). It comprised of eight β-strands flanked by six alpha helices, two on one side and four on the opposite side of the beta sheet as previously reported in

S. japonicum UT26. The beta sheet is a mixture with topology +1, +2, -1x, 2x, +1x, +1x, +1x, and directionality +, –, +, +, +, +, +, +, similar to HanR reported by Novak *et al.* (2014). The cap domain, on the other hand, spanned residues 134–213 and consisted of six alpha helices (α3–α8) and is the region reported to contribute to variability among HLD enzymes. The two domains form a cavity between them that is predominantly hydrophobic and is where the active site is located (Oakley *et al.*, 2004).

4.3.3.2 Validation of predicted LinB models

Ramachandran plot analyses of predicted LinB models showed 226 (90.0%) residues occur in the most favored regions, 24 (9.6%) residues in the additionally allowed regions, 1 (0.4%) residue in the generously allowed regions, and none in the disallowed regions of LinBS6 model. On the other hand, the LinBS8 model contained 222 (90.6%) residues in the most favored regions, 22 (9.0%) in the additionally allowed regions, 1 (0.4%) residue in the generously allowed regions and none in the disallowed regions (**Table 7**). Ramachandran plots of the LinBS6 and LinBS8 models were as shown (**Fig. 36)**.

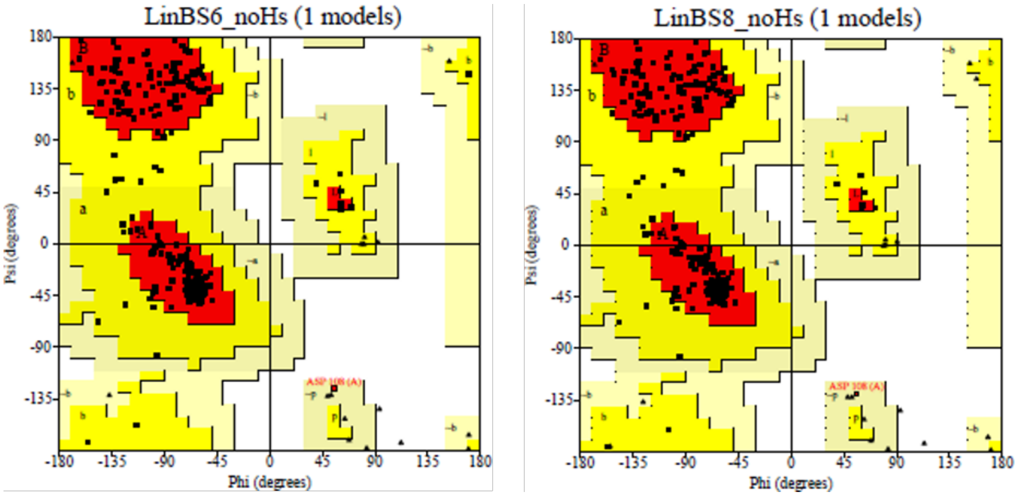


Figure 36. Ramachandran plot by PROCHECK for LinBS6 and LinBS8 models showing the φ–ψ distribution for the different regions. The core regions (marked A, B, L), additionally allowed regions (marked a, b, l, p), generously allowed regions (marked ~a, ~b, ~l, ~p), and disallowed regions are shown in red, yellow, grey and white colors, respectively. Non-proline and non-glycine residues are represented by black squares and glycine residues by black triangles. Residues in disallowed regions are shown in red.

The distribution of torsional (φ and ψ) and dihedral angles in the LinBS6 and LinBS8 models was the same, with normalized G-factor Z-scores of -0.04 and -0.00, respectively. However, the overall G-factors of LinBS6 and LinBS8 models were slightly different (-0.09 and -0.13, respectively) but within acceptable range (≤ -0.50) (**Table 7**). From the 3D profile assessment, all (100%) residues in both LinBS6 and LinBS8 models had averaged 3D/1D scores of 0.2 or higher, with a normalized Z-score of -2.89. In addition, the PROSAII (-ve) Z-scores of both LinBS6 and LinBS8 models were comparable (-0.04 and 0.00, respectively). The Verify3D score and ProSAII (-ve) score profiles of LinBS6 and LinBS8 models plotted over a window average of 7 residues were as shown (**Fig. 37**).

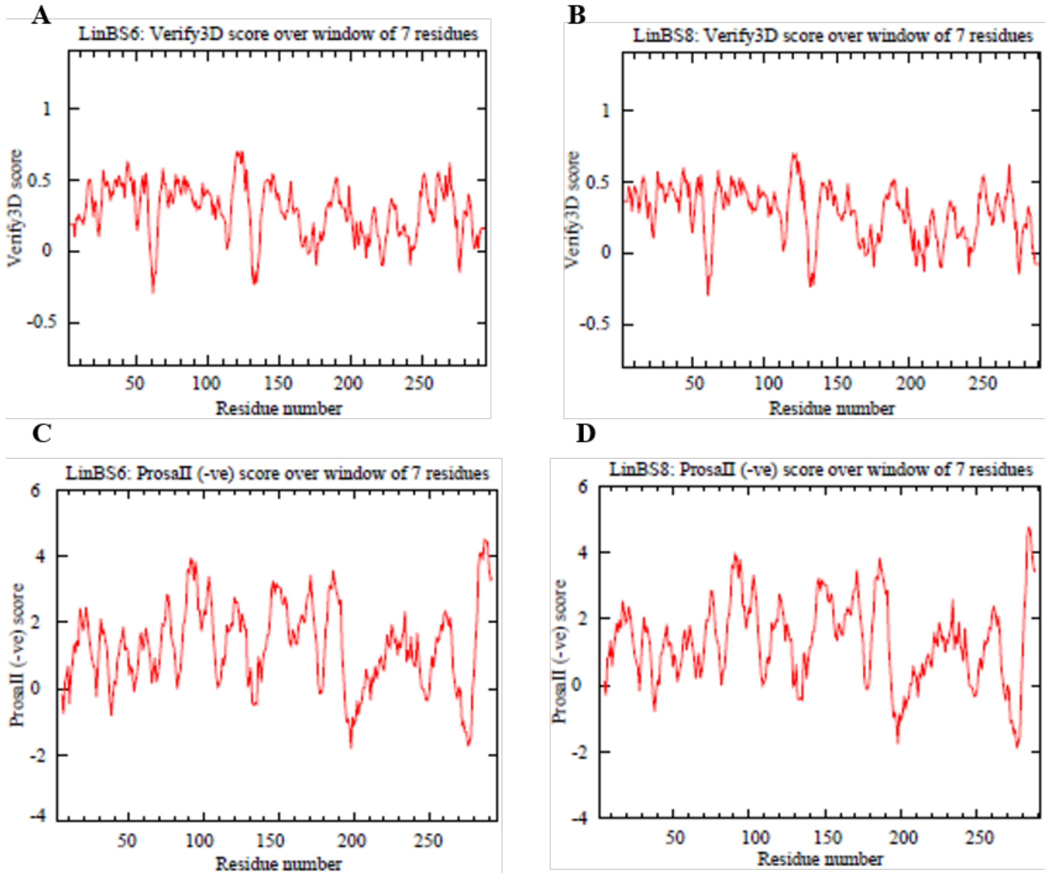


Figure 37. Verify3D and ProSAII profile plots of LinB models. **A & C**) Verify3D and ProSAII profiles, respectively of LinBS6 model plotted over a window average of 7 residues. **B & D**) Verify3D and ProSAII profile plots, respectively LinBS8 model over a window average of 7 residues. Positive scores indicate correctly modeled segments of the 3D structure.

The overall quality factor of LinBS6 model was 94.74% (close to the 95% limit for good high-resolution structures) while that of LinBS8 model was 99.29% (comparable to high resolution X-ray structures). Regions of error comprising residues with grossly mistraced and/or misregistered atoms and exceeding the 95% and 99% rejection limits (highlighted in yellow and red, respectively) were apparent in both LinBS6 model. These included atoms of residues at positions 7 and 36–41 (highlighted red) and 9–11, 42–44, and 51–52 (highlighted yellow). However, LinBS8 model contained only two residues (51 and 52) with mistraced atoms (highlighted yellow). The ERRAT v2.0 plots of the LinBS6 and LinBS8 models were as shown (**Fig. 38**).

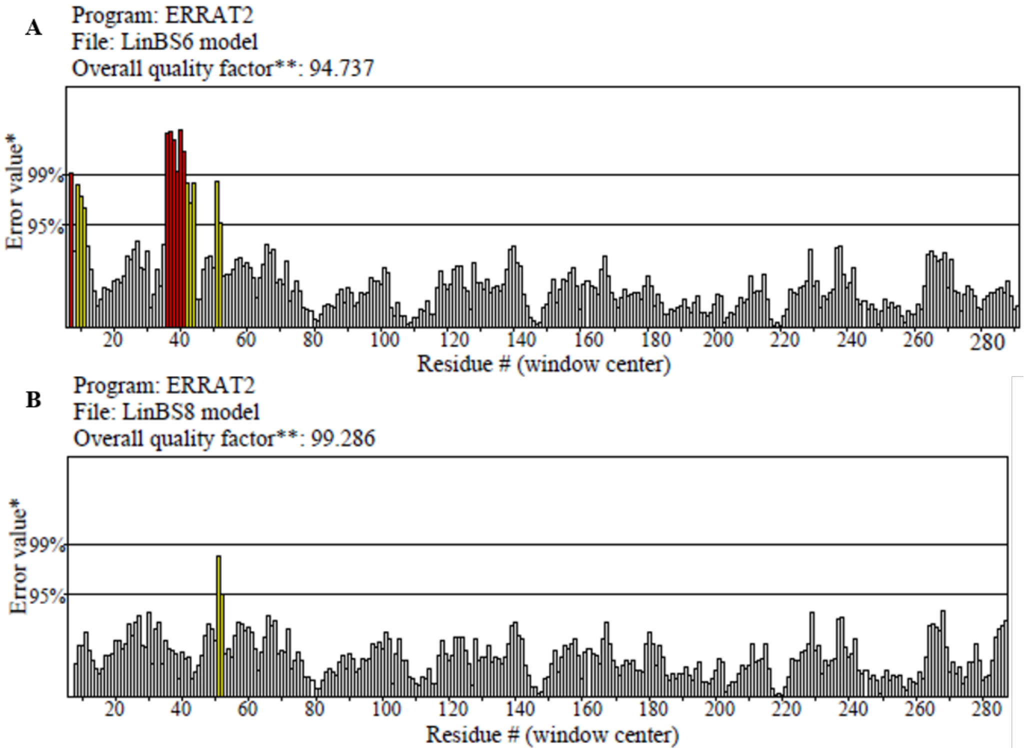


Figure 38. ERRAT profile plots of LinBS6 and LinBS8 models. **A**) ERRAT profile plot of LinBS6 model with overall quality (%) indicated above the plot. **B**) ERRAT profile plot of LinBS8 model with overall quality factor (%) shown above the plot. Regions of error exceeding the 95% confidence limit for rejection are highlighted yellow and those exceeding the 99% confidence are highlighted red. *Two lines drawn on the error axis show the confidence for rejection of regions exceeding the error value. **Overall quality factor of the model is expressed as the percentage of the protein whose calculated error value is below the 95% limit for rejection. The average overall quality factor for good structures of high resolution is 95% or more whereas that of lower resolution structures (2.5–3Å) is around 91%.

4.3.3.3 Solvent accessibility and secondary structure of LinB

Secondary structural elements of the LinBS6 and LinBS8 models consisted of eight beta strands, eleven alpha helices and nine 3_{10} helices spanning residues in the following order: β1(12–16), β2(19–26), β3(31–35), η1(42–45), η2(49–52), β4 (57–61), α1(82–95), β5(102–108), α2(109–120), η3(122–124), β6(125–133), η4(140–142), η5(145–149), α3(150–155), α4(159–164), α5(168–171), α6(173–176), α7(184–191), η6(192–194), η7(199–201), α8(202–206), η8(208–210), α9(218–231), β7(238–245), α10(251–257), β8(263–270), η9(274–276), and α11(279–293) as shown (**Fig. 39**).

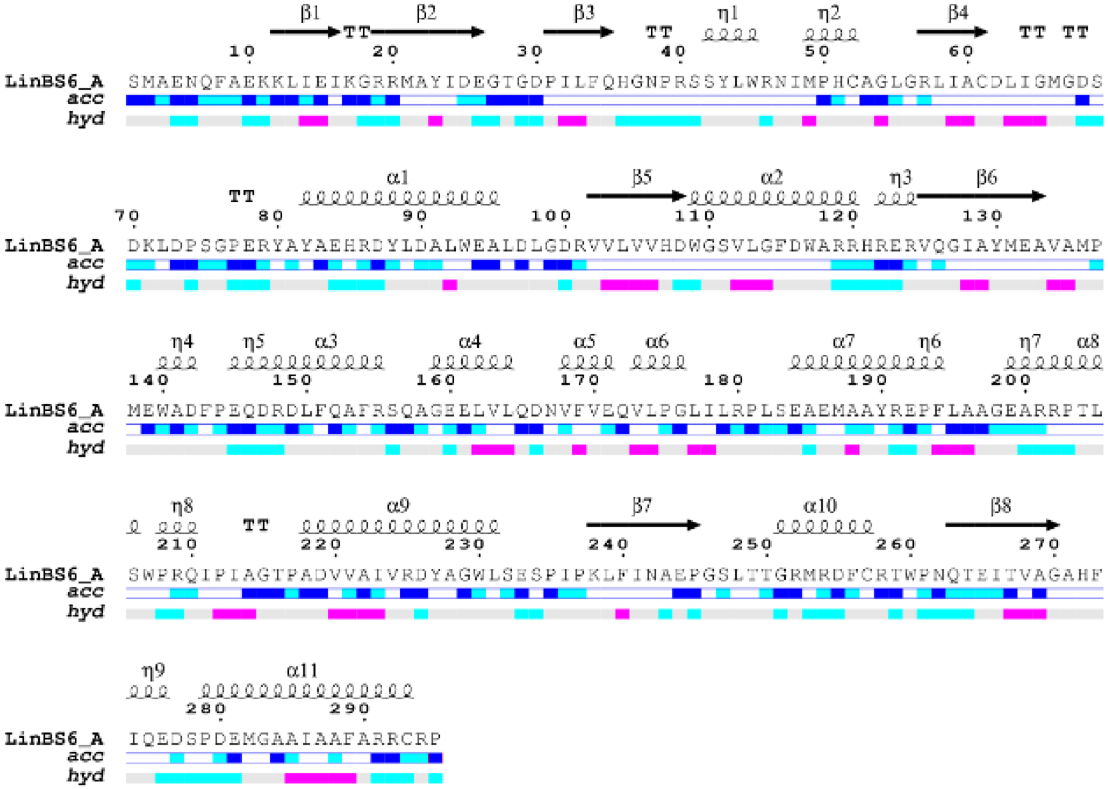

β1
β2
β3
η1
η2
β4
TT
TT
TT
TT
10
20
30
40
50
60
LinBS6_A
SMAENQFAEKKLIEIKGRRMAYIDEGTGDPILFQHGNPRSSYLWRNIMPHCAGLGRLIACDLIGMGDS
acc
hyd
α1
β5
α2
η3
β6
TT
70
80
90
100
110
120
130
LinBS6_A
DKLDPSGPERYAYAEHRDYLDALWEALDLGDRVVLVVHDWGSVLGFDWARRHRERVQGIAYMEAVAMP
acc
hyd
η4
η5
α3
α4
α5
α6
α7
η6
η7
α8
140
150
160
170
180
190
200
LinBS6_A
MEWADFPEQDRDLFQAFRSQAGEELVLQDNVFVEQVLPGLILRPLSEAEMAAYREPFLAAGEARRPTL
acc
hyd
η8
TT
α9
β7
α10
β8
210
220
230
240
250
260
270
LinBS6_A
SWFRQIPIAGTPADVVAIVRDYAGWLSESPIPKLFINAEPGSLTTGRMRDFCRTWPNQTEITVAGAHF
acc
hyd
η9
α11
280
290
LinBS6_A
IQEDSPDEMGAAIAAFARRCRP
acc
hyd

Figure 39. The flat figure for LinB sequence from *Sphingobium* sp. S6 adorned with secondary structural elements (β-strands, helices, and turns represented by arrows, squiggles, and TT letters, respectively). The first and second bar underneath the sequence indicates solvent accessibility (blue for accessible, cyan for intermediate, and white for buried) and hydropathy (cyan for hydrophilic, grey for neutral, and pink for hydrophobic), respectively. Letters at the bottom represent crystallographic, protein-protein, and protein-ligand contacts. The output was generated by ENDscript 2.0 (https://endscript.ibcp.fr/ESPript/ENDscript/).

A structural alignment of LinBS6 and LinBS8 models with the crystal structures of four LinB proteins (PDB IDs: 1MJ5, 1CV2, 1IZ7, and 4H7K) most homologous to LinB from *Sphingobium* sp. S6 and *Sphingobium* sp. S8 showed the secondary structure elements were conserved. However, subtle differences occur in the arrangement of β-strands (β1, β2, and β5) and α-helices (α3, α4, and α9) among the LinB proteins. In addition, the residues at positions 3-7, 9, 40, 81, 112, 134-135, 138, 224, 247, 282, 290, and 293 were also seen to differ between the dehalogenases.

4.3.4 Active binding pocket and 3D protein fold of LinB

4.3.4.1 Modeled LinB from *Sphingobium* sp. S6

The LinBS6 model and template adopt the same 3D protein fold and possess similar binding sites (**Fig. 40**). The active site is localized within the hydrophobic cavity formed by the main (core) domain and the cap domain and consists of a catalytic trio comprising of the nucleophile Asp108, catalytic base His272 and catalytic acid Glu132 (Novak *et al*., 2014; Oakley *et al*., 2004). The arrangement of the catalytic triad within the LinBS6 model is similar to the template (1mj5). Of the two mutations (Leu13 and Met138) identified in the LinB sequence from *Sphingobium* sp. S6, only Met138 substitution occur within the active binding pocket of the LinBS6 model whereas Leu13 lies outside the binding region (**Fig. 41**).

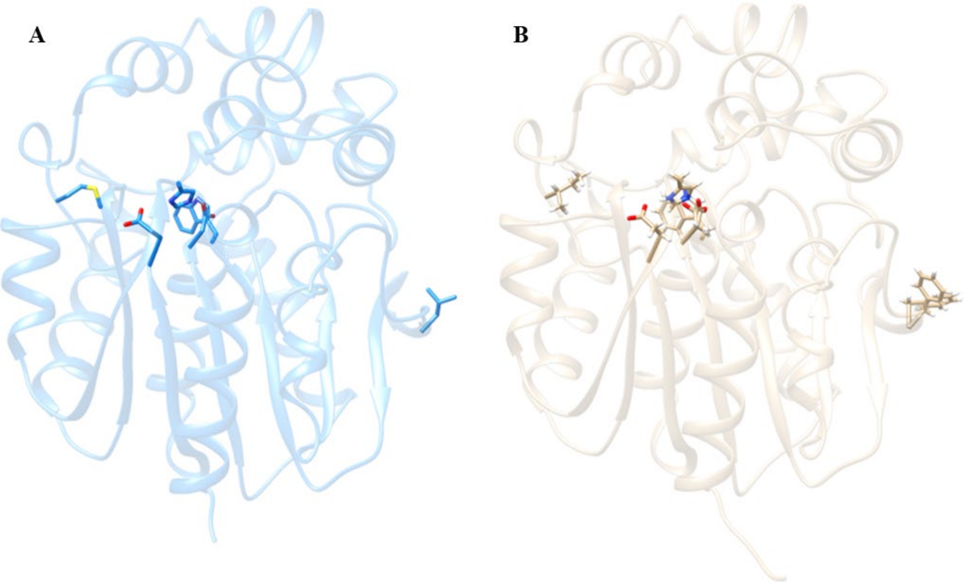


Figure 40. 3D protein folds of LinBS6 model and template (1mj5) . A) LinBS6 model (blue) from *Sphingobium* sp. S6 and B) template , 1mj5 (tan) shown side by side. Residues (including mutations) in the binding sites are displayed as colored sticks in blue and tan, respectively.

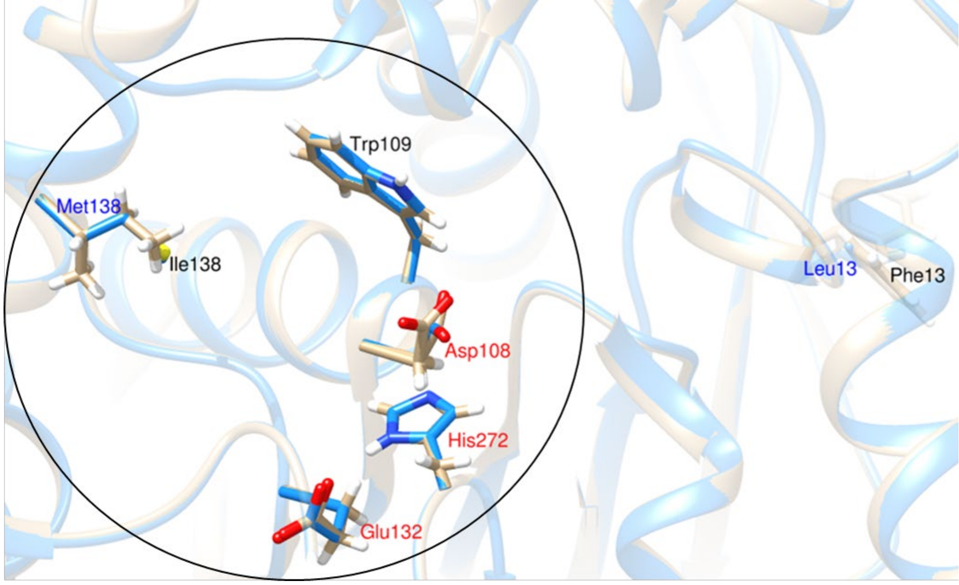


Figure 41. Substrate binding pocket of LinB from *Sphingobium* sp. S6 (LinBS6). The circle (not drawn to scale) represents the probable active site containing the catalytic residues (red). The mutated residues (Met138 and Leu13) identified in the LinB sequence (blue) and their location within the 3D protein fold is also shown. The LinB model (blue) and template (tan) have been superimposed and the residues displayed as colored sticks in blue and tan, respectively.

4.3.4.2 Modeled LinB from *Sphingobium* sp. S8

The LinBS8 model and template adopt the same 3D protein fold and have similar binding sites (**Fig. 42**). The active site is localized within the hydrophobic cavity formed by the core (main) domain and cap domain and consists of a catalytic trio comprising of the nucleophilic-Asp108, catalytic base-His272 and catalytic acid-Glu132 (Novak *et al*., 2014; Oakley *et al*., 2004). The arrangement of the catalytic trio (Asp108-His272-Glu132) within the LinBS8 model is same as the template (1mj5). Among the amino acid substitutions (mutations) identified in the LinB sequence from *Sphingobium* sp. S8 (Val134, Met138, Met282, Ala283, Arg284, and Val289), only Val134 and Met138 substitutions are located within the active binding region whereas the rest lie outside the substrate binding pocket (at the C-terminus) (**Fig. 43**)

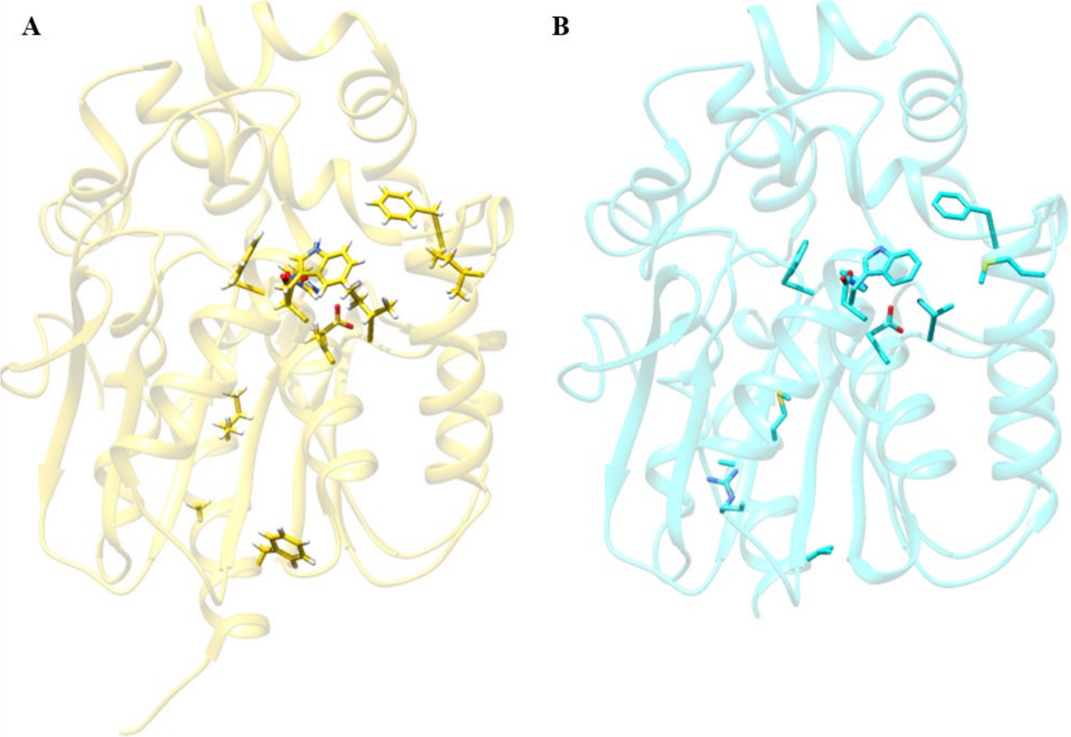


Figure 42. 3D protein folds of LinBS8 model and template (1mj5). A) The template , 1mj5 (gold) and B) LinBS8 model (cyan) from *Sphingobium* sp. S8 shown side by side. Residues (including mutations) in the binding sites are displayed as colored sticks in gold and cyan, respectively.

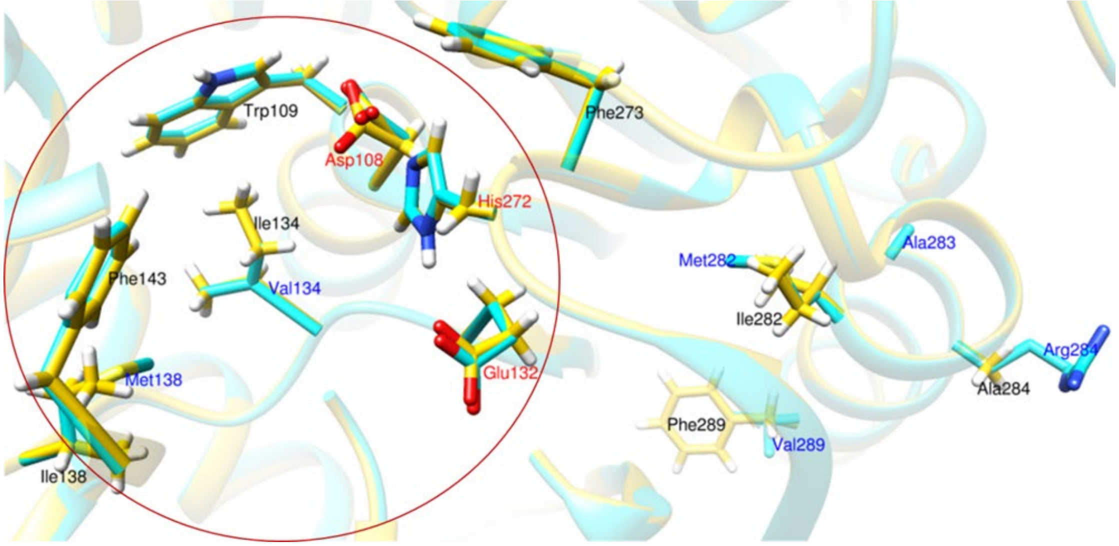


Figure 43. Substrate binding pocket of LinB from *Sphingobium* sp. S8 (LinBS8). The circle (not drawn to scale) represents the probable active site containing the catalytic residues (red). The mutated residues identified in the LinB sequence (blue) and their location within the 3D protein fold is also shown. The LinB model (cyan) and template (gold) have been superimposed and the residues displayed as colored sticks in cyan and gold, respectively.

Theoretical 3D structures of LinC, LinD, and LinE – the other key Lin pathway enzymes, were modeled and validated in the same manner as LinA and LinB. The modeled 3D structures of LinC from *Sphingobium* sp. S6 and *Sphingobium* sp. S8 (LinCS6 and LinCS8 models, respectively) were homo-tetramers composed of chain A, chain B, chain C, and chain D. On the other hand, the 3D structures of LinD from *Sphingobium* sp. S6 and *Sphingobium* sp. S8 (i.e. LinDS6 and LinDS8 models, respectively) comprised of two protomers (chain A and chain B) that were not true homodimers. However, the 3D structures of LinE from *Sphingobium* sp. S6 and *Sphingobium* sp. S8 (LinES6 and LinES8 models, respectively) comprising of chains A and B were actual homodimers (**Appendix 2A−4B**). These models were also evaluated in the same way as LinA and LinB models and the data presented as shown (**Appendix 5**).

CHAPTER FIVE

5.0 DISCUSSION

5.1 Presence and copy numbers of *Lin* genes and the IS*6100* in the *Sphingobium* sp. S6

Southern blot analyses of seven *Lin* genes and IS*6100* in *Sphingobium* sp. S6 showed the presence of one copy of *LinA*, *LinB*, *LinC*, and *LinD*, two copies for *LinE* and *LinX*, and multiple copies of *LinR* and IS*6100*. The number of copies of *LinA* varies among well-known HCH-degrading *Sphingobium* strains B90A, UT26 and Sp+. Strain B90A contains three copies whereas UT26 and Sp+ have one copy each, similar to strain S6. In addition, B90A also contains three copies of *LinX* compared to two copies in strain S6 and one copy in strains UT26 and Sp+. Also, because *LinA* and *LinX* yielded hybridizing fragments of same size in the blot they may be located very closely in strain S6. Similarly, *LinD* and *LinE* blots generated fragments of same sizes and are therefore likely to occur together in strain S6. In fact, *LinD*, *LinE*, and *LinR* to form a single operon with *LinR* being located upstream of *LinE* in strains B90A and UT26 (Dogra *et al*., 2004). The *LinR* is a transcriptional regulator controlling *LinD* and *LinE* expression (Lal *et al*., 2006) hence probably occur together.

On the other hand, single copies of *LinA*, *LinB*, *LinC*, and *LinD* in strain S6 suggest that these *Lin* genes may be located close together. Even though the number of copies of the IS*6100* could not be ascertained with confidence, the presence of multiple copies is consistent with the numbers determined in other HCH-degrading *Sphingobium* strains B90A, UT26 and Sp+, which contained 15, 5, and 6 copies, respectively (Verma *et al*., 2017). These data indicate the vital role that the IS*6100* plays in the dissemination of *Lin* genes among bacterial species from remote geographical locations. Nevertheless, HCH-degrading *Sphingobium* strains S6 and S8 possess identical *Lin* genes although *Lin* gene organization may not be similar to other *Sphingobium* strains. Moreover, the number of copies of *Lin* genes and the IS*6100* tends to differ from strain to strain and the *Lin* genes within the same strain may not have same copy numbers (Dogra *et al*., 2004)

5.2 Genetic variability of *Lin* genes in the *Sphingobium* sp. S6 and *Sphingobium* sp. S8

The *Lin* genes (*LinA*, *LinB*, *LinC*, *LinD*, *LinE*, *LinR*, and *LinX*) and their corresponding protein sequences (LinA, LinB, LinC, LinD, LinE, LinR, and LinX) contained the expected number of base pairs and amino acid residues, respectively in both *Sphingobium* sp. S6 and *Sphingobium* sp. S8. The Lin protein sequences (LinA-LinX) of *Sphingobium* sp. S6 and *Sphingobium* sp. S8 were 92-100% identical and were highly similar to those of other HCH-degrading bacteria, including the well-known *Sphingobium* strains UT26S, B90A, and Sp+. The sequence identities and similarities of LinA, LinB, LinC, LinD, LinE, LinR, and LinX from *Sphingobium* sp. S6 to those from other HCH-degrading bacteria were 98-100%. On the other hand, the sequence identities and similarities of LinA, LinB, LinC, LinD, LinE, LinR, and LinX from *Sphingobium* sp. S8 to other HCH-degrading bacteria were 93-100%. Thus, *Sphingobium* sp. S8 has accumulated more strain-specific mutations than *Sphingobium* sp. S6, even though *Lin* genes in the two *Sphingobium* strains are much conserved at the amino acid level. Percentage sequence identity is defined by the number of characters (residues) that match between two sequences in the alignment while percentage sequence similarity is characterized by the number of residues that resemble between two sequences, based on their physicochemical properties (Rost, 1999).

HCH-dehydrochlorinase (LinA) of *Sphingobium* sp. S6 and *Sphingobium* sp. S8 belonged to the nuclear transport factor 2-like (NTF2-like) protein superfamily. Both LinA are NTF2-like proteins containing SnoaL-like polyketide cyclase domain (SnoaL-4) similar to LinA from *Sphingobium japonicum* UT26. This domain is present in several proteins possessing the SnoaL fold, is highly conserved and spans between amino acid residues 5 and 130, and contains 34 highly conserved amino acid residues (Marchler-Bauer *et al*., 2011, 2015, 2017; Marchler-Bauer & Bryant, 2004). The LinA polypeptide of S8 is shorter than that of S6 by six amino acid residues, representing a 0.6% sequence divergence between them. However, both dehydrochlorinases have about 2−6% sequence divergence with LinA of other HCH-degrading *Sphingobium* strains. In evolutionary terms, HCH-dehydrochlorinases (LinA) diverged into three separate clusters. Dehydrochlorinases from HCH-degrading *Sphingobium* and/or *Sphingomonas* species comprise the largest cluster. LinA from *Sphingobium* strains S6 and S8 clustered with the well-known and most closely related *Sphingobium* strains UT26S, B90A, and Sp+, belonging in this cluster. Further, only four variants of LinA have been reported to date (Nagata *et al*., 2007; Nayyar & Lal, 2016), hence LinA is much more conserved.

Haloalkane dehalogenases (HLDs) belong to the alpha/beta-hydrolase superfamily and are characterized by the presence of α/β-hydrolase fingerprint (Damborský & Koča, 1999). Haloalkane dehalogenase (LinB) of *Sphingobium* strains S6 and S8 belonged to the HLD superfamily containing the PRK03592 domain that spanned between amino acid residues 4 and 291 (Marchler-Bauer & Bryant, 2004). The LinB polypeptide of S8 was shorter than that of S6 by five amino acid residues, representing a 0.3% sequence divergence. However, LinB sequence of S6 and S8 showed 2−5% divergence with HLDs of other *Sphingobium* strains and this might account for the difference in the degradation rates of S6 and S8 with other strains against the same HCH isomer (Cao *et al*., 2013). Both dehalogenases were highly homologous to HLDs from *Sphingobium* sp. TKS and *Pseudomonas aeruginosa* ITRC-5 and seemed to have accumulated strain-specific mutations, especially at position 138. Besides, the amino acid residues at positions 134, 138, 224, 247, and 253 in the LinB sequence of most HLDs tended to be vary.

Nevertheless, the amino acid substitutions at these positions are non-synonymous mutations, which may not affect LinB function but may be responsible for the observed variation in activity towards β-HCH isomer (Nagata *et al*., 2007). Because β-HCH is the most persistent isomer, such sites could represent possible targets for generating LinB variants with improved activity towards β-HCH (Boháč *et al.*, 2002; Damborský *et al.*, 1998). Haloalkane dehalogenases (HLDs) showed they have diverged into two separate clusters on the evolutionary tree, with the largest cluster comprising of HLDs from HCH-degrading *Sphingobium* and/or *Sphingomonas* species, of which HLDs from the *Sphingobium* strains S6 and S8 belonged. The HLDs within this cluster are characterized by low sequence divergence and have a sister group relationships with each other, as exemplified by *Sphingobium* strains S6 and S8.

Both 2,5-DDOL dehydrogenase (LinC) and the LinC-like 2,5-DDOL dehydrogenase (LinX) of *Sphingobium* sp. S6 and *Sphingobium* sp. S8 belong to the short-chain alcohol dehydrogenase/reductase (SDR) protein superfamily (Lal *et al.*, 2010). Two highly conserved regions previously reported to occur among the SDR family proteins (Persson *et al.*, 1991) were present in the LinC of *Sphingobium* strains S6 and S8. The first very conserved region located at the N-terminus and containing an alternating pattern of β-sheets and α-helices (β–α–β) is the binding site for NAD$^+$, a co-factor indispensable to the activity of many dehydrogenases (Nagata *et al.*, 1999). The second homologous region spans from position 150 to 154 on the consensus sequence and contains the active site residues tyrosine and lysine at positions 150 and 154, respectively. But whereas tyrosine appears to be conserved amongst dehydrogenases, lysine tends to be variable (Nagata *et al.*, 1999). Even though tyrosine and lysine were conserved in the LinC of *Sphingobium* strains S6 and S8, there was a positional shift by four residues on the consensus sequence with tyrosine occurring at position 154 and lysine at position 158. But this change is not expected to alter LinC activity because Tyr154 is located in a region with well-conserved hydrophilicity. Further, its occurrence within the conserved region suggests that LinC has NAD$^+$-dependent dehydrogenation (Nagata *et al.*, 1994, 1999).

Dehydrogenases (LinC and LinX) of *Sphingobium* sp. S6 and *Sphingobium* sp. S8 share a high sequence homology with those of other HCH-degrading bacteria, especially UT26S. Moreover, only few mutations could be identified in the entire alignment of LinC and LinX and both dehydrogenases from S6 were very identical (100%) to those of UT26S. Both dehydrogenases are cluster together and belong in the same comprising 2,5-DDOL dehydrogenases from HCH-degrading *Sphingobium* and/or *Sphingomonas* species. The 2,5-DDOL dehydrogenases within this cluster have minimal rate of sequence divergence and appear to be much more conserved.

LinD belongs to the theta class of glutathione *S*-transferases (GSTs), which comprise all GSTs from bacteria (Nagata *et al*., 1999). Little is known about this class of GSTs, including solved crystal structures as well as putative residues responsible for catalytic activity, and hence only few sequences with significant similarity to LinD are available in the databases. Nevertheless, LinD of *Sphingobium* sp. S6 and *Sphingobium* sp. S8 show strong homology to other reductive dechlorinases and are more conserved. The few mutations apparent in the dechlorinases of S6 and S8 were localized at the C-terminal end. Whether these mutated residues are likely to affect LinD activity is not clear because the putative residues responsible for the catalytic activity of LinD have not yet been elucidated. LinD proteins of S6 and S8 appear to be more conserved and are clustered together on the evolutionary tree with dechlorinases of other HCH-degrading strains, including UT26S. However, the data was insufficient data to give an unambiguous branching in the tree because LinD exhibited little similarity to other GST family proteins and thus only a few sequences identical to LinD could be obtained.

LinE is categorized as a *meta*-cleavage dioxygenase because of its similarity to proteins within the *meta*-cleavage dioxygenase family. It is also called chloro/hydroquinone 1,2-dioxygenase and catalyzes cleavage of aromatic compounds with two hydroxyl groups at *para*-positions (Miyauchi *et al*., 1999; Nagata *et al*., 1999). LinE is the first ring cleavage enzyme to show a high preference for chloro/hydroquinone over catechol (Miyauchi *et al.*, 1999). Some of the catalytic residues of *meta*-cleavage dioxygenases that are important for iron (II) binding and catalytic activity (i.e. His162, His229, and Glu278) (Miyauchi *et al.*, 1999) were present and conserved in LinE. Thus, providing evidence that LinE could function as a ring-cleavage dioxygenase, as indicated by Miyauchi *et al.* (1999). Moreover, the putative amino acid residues (His162, His229, and Glu278) could present possible targets in the future when designing site-specific mutagenesis experiments. The LinE polypeptide from *Sphingobium* sp. S8 was shorter than that of *Sphingobium* sp. S6 by four amino acid residues (0.3% divergence) and both polypeptides were highly conserved and showed significant similarity to that of UT26S. These dioxygenases and those of other HCH-degrading *Sphingobium* and/or *Sphingomonas* strains formed one distinct cluster on the evolutionary tree and have low rate of sequence divergence.

LinR belongs to the LysR family of transcriptional regulators (i.e. LTTRs), one of the ubiquitous regulators of transcription of certain genes (especially *LinD* and *LinE*) responsible for breaking down aromatic compounds in prokaryotes (Miyauchi *et al.*, 2002). Thus, LinR induces the expression of LinD and LinE (Nagata *et al.*, 1999). LinR and all other LTTRs have been reported to contain a highly conserved putative helix-turn-helix (HTH) motif at the N-terminal region, that directly interacts with DNA (DNA-binding motif) and spans from serine at position 22 to leucine at position 52 (Miyauchi *et al.*, 2002). Many *LinR* genes contain frameshifts at the 5' end and full-length polypeptides are not always obtained and therefore, the location of the HTH motif may however vary from one transcriptional regulator to another. For instance, the HTH motif spanned from serine at position 26 to leucine at position 56 in the LinR sequences from *Sphingobium* sp. S6 and *Sphingobium* sp. S8. Whereas classic helix-turn-helix structures contain a conserved glycine residue in the middle of the structure, LinR and many other LTTRs contain aspartate (D31) at this position. All LTTRs appeared to be more conserved and seem to have evolved from a common ancestral LTTR protein and are clustered together on the evolutionary tree. LinR from *Sphingobium* strains S6 and S8 clustered with those of *Sphingobium* and/or *Sphingomonas* species, including UT26S.

5.3 Comparative modeling of 3-D structures for key enzymes in the HCH-degradation pathway

Structural models of the Lin pathways enzymes are indispensable in the understanding of their reaction mechanisms (Damborský & Koča, 1999). Also, the identification of interactive residues in the active sites of these enzymes with potential applications in protein engineering experiments such as site-directed mutagenesis relies on the existence of reliable models (Hsieh & Vaisvila, 2013). LinAS6 and LinAS8 models have similar binding sites and adopt the same 3D protein fold as the template and superimposition of the LinAS6 and LinAS8 models with the template revealed there is no significant deviation in the conformation. Detailed analysis of the amino acids of LinA from *Sphingobium* sp. S6 and *Sphingobium* sp. S8 revealed that the catalytic residues in the active pocket of LinA, so called "catalytic triad" (Lys20, Asp25, and His73) were conserved and present in all LinA proteins. The hydrophobic residues (Ile44, Ile47, Leu59, Val63, Met67, Leu100, Ile107, Ala131, Thr133, and Phe136) whose side chains surround the substrate-binding pocket of LinA (Okai *et al*., 2010) were also conserved in all LinA proteins. Similarly, the side chains of Lys20, Leu21, Val24, Asp25, Trp42, Leu64, Phe68, Cys71, His73, Val94, Leu96, Ile109, Ala111, Phe113, and Arg129 that line up the active site of LinA were conserved.

According to Okai *et al*. (2010), the hydrophobic residues (Trp42, Leu64, Leu96, and Phe113) localized in the cavity of LinA play a role critical in the dehydrochlorination activity by determining the substrate size to be catalyzed. LinA mutants lacking Leu64 have no dehydrochlorination activity while Tr42, Leu96 and Phe113 mutants have a 95%, 90%, and 62% reduction in the dehydrochlorination activity, respectively. The amino acid differences in the LinA sequences from *Sphingobium* sp. S6 and *Sphingobium* sp. S8 does not involve these residues. Besides, the mutated residues are located outside the active site of LinA. It is therefore likely that this difference may not affect LinA activity. However, any potential implication of these mutations in the catalytic mechanism of LinA can be unraveled through ligand docking and molecular dynamics simulations.

The LinBS6 and LinBS8 models adopt a typical alpha/beta-hydrolase fold that consists of core (main) and cap domains, similar to the template (1mj5) and other structurally known dehalogenases (Damborský & Koča, 1999; Marek *et al*., 2000). The LinB active site is characterized by a conserved catalytic quintet, consisting of a catalytic trio (Asp108–His272–Glu132) plus one pair of halide-stabilizing residues (Trp–Trp or Trp–Asn) (Novak *et al*., 2014; Oakley *et al*., 2004). The arrangement of the catalytic trio is conserved between the template and the LinBS6 and LinBS8 models. It is such that the nucleophilic Asp108 occurs at the turn between beta-strand β5 and alpha helix α2, the catalytic base His272 occurs within the loop adjoining beta-strand β8 and alpha helix α11 whereas the catalytic acid Glu132 occurs in beta-strand β6. The putative residues (Trp109, Val134, Phe143, Pro144, Gln146, Asp147, Phe151, Phe169, Val173, Leu177, Trp207, Pro208, Ile211, Ala247, Leu248, and Phe273) surrounding the active site reported by Marek *et al*. (2000) were also conserved in the LinB proteins.

Structural modeling and site-directed mutagenesis suggest that seven amino acid residue differences in LinB between B90A−MI1205−BHC-A and UT26−Sp+ accounted for the difference in their ability to degrade 2,3,4,5,6-pentachlorocyclohexanol (PCHL). More so, Val134 and His247 are necessary to correctly orientate PCHL for S_N2 attack. The rest of the differences (Thr81, Val112, Thr135, Leu138, and Ile253) likewise occur within the catalytic site (Thr135 and Leu138 occur near Val134 and the catalytic acid Glu132) but have little effect on activity (Cao *et al*., 2013; Lal *et al*., 2010). Detailed amino acid analysis of LinB from *Sphingobium* sp. S6 and *Sphingobium* sp. S8 revealed some amino acid differences regarding these residues in the two proteins. This entailed Thr81−Ala81, Thr135−Ala135, Leu138−Met138, His247−Ser247, and Ile253−Met253 substitutions in the LinB proteins of strain S6 and strain S8.

Further, four extra amino acid differences in the LinB of strain S8 (Met282, Ala283, Arg284, and Val289) are located outside the active site. It has been postulated that differences in LinB may bring about differences in the rate of degradation among different strains towards the same isomer of HCH. Nevertheless, more studies need to be undertaken to confirm this hypothesis (Cao *et al.*, 2013). Despite the amino acid differences observed in the LinA, LinB, LinC, LinD, and LinE sequences, structural modeling revealed their 3D structures are conserved. In addition, the catalytic residues in the active site as well as putative residues surrounding the substrate binding pockets of these enzymes are also conserved. However, further studies need to be done to ascertain whether the identified mutations (especially those occurring within the substrate binding pocket of LinB) alter the ability of the Lin enzymes to bind and degrade lindane.

CHAPTER SIX

6.0 CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

1) The *Sphingobium* sp. S6 and *Sphingobium* sp. S8 characterized in this study possess *Lin* genes. One copy of *LinA*, *LinB*, *LinC*, and *LinD* gene, two copies of *LinE*, and *LinX* gene, and multiple copies of *LinR* gene and IS*6100* were present within the genome of *Sphingobium* sp. S6.

2) There is no variability of *Lin* genes in *Sphingobium* sp. S6 and *Sphingobium* sp. S8. The *Lin* genes in the two *Sphingobium* strains S6 and S8 are also conserved and cluster together on the phylogenetic tree.

3) Target 3D structures of LinA, LinB, LinC, LinD, and LinE from *Sphingobium* sp. S6 and *Sphingobium* sp. S8 are conserved and possess catalytic residues in the active site that are important in the degradation of lindane.

6.2 Recommendations

Theoretical 3D structures of LinA, LinB, LinC, LinD, and LinE can be utilized in designing mutagenesis experiments based on the identified catalytic sites. Because of these catalytic sites, substrate binding properties of Lin proteins need to be obtained, which would form the basis for designing enzyme mutants with improved properties such as kinetics, catalytic efficiency, and substrate range for degradation of lindane. Moreover, unique substitutions (mutations) were identified in the amino acid sequences of LinA and LinB and some of these mutations were localized in the catalytic site. Therefore, the effect of these mutations on the catalytic activity of LinA and LinB to be studied further.

REFERENCES

Adams, G. O., Fufeyin, P. T., Okoro, S. E., & Ehinomen, I. (2015). Bioremediation, biostimulation and bioaugmention: A review. *International Journal of Environmental Bioremediation & Biodegradation*, *3*(1), 28–39.

Altschul, S. F. (2014). BLAST Algorithm. In John Wiley & Sons, Ltd (Ed.), *ELS* (1st Ed.). Wiley.

Aucha, J. K., Wandiga, S. O., Abong'o, D. A., Madadi, V. O., & Osoro, E. M. (2017). *Organochlorine Pesticides Residue Levels in Air and Soil from Nairobi and Mount Kenya regions, Kenya*.

Bachmann, A., Bruin, W. de, Jumelet, J. C., Rijnaarts, H. H., & Zehnder, A. J. (1988). Aerobic biomineralization of alpha-hexachlorocyclohexane in contaminated soil. *Appl. Environ. Microbiol.*, *54*(2), 548–554.

Benkert, P., Biasini, M., & Schwede, T. (2011). Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, *27*(3), 343–350.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T. G., Bertoni, M., Bordoli, L., & Schwede, T. (2014). SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, *42*(W1), W252–W258.

Boháč, M., Nagata, Y., Prokop, Z., Prokop, M., Monincová, M., Tsuda, M., Koča, J., & Damborský, J. (2002). Halide-Stabilizing Residues of Haloalkane Dehalogenases Studied by Quantum Mechanic Calculations and Site-Directed Mutagenesis. *Biochemistry*, *41*(48), 14272–14280.

Böltner, D., Moreno-Morillas, S., & Ramos, J.-L. (2005). 16S rDNA phylogeny and distribution of lin genes in novel hexachlorocyclohexane-degrading *Sphingomonas* strains. *Environmental Microbiology*, *7*(9), 1329–1338.

Bowie, J., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, *253*(5016), 164–170.

Cao, L., Xu, J., Li, M., Wu, G., Wang, J., Guan, Y., He, J., Li, S., & Hong, Q. (2013). Characterization and analysis of three newly isolated hexachlorocyclohexane (HCH)-degrading strains. *International Biodeterioration & Biodegradation*, *85*, 407–412.

Castrignano, T., De Meo, P. D., Cozzetto, D., Talamo, I. G., & Tramontano, A. (2006). The PMDB protein model database. *Nucleic Acids Research*, *34*(suppl_1), D306–D309.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., & Richardson, D. C. (2010). MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, *66*(1), 12–21.

Colovos, C., & Yeates, T. O. (1993). Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Science*, *2*(9), 1511–1519.

Damborský, J., Bohác, M., Prokop, M., Kutý, M., & Koca, J. (1998). Computational site-directed mutagenesis of haloalkane dehalogenase in position 172. *Protein Engineering, Design and Selection*, *11*(10), 901–907.

Damborský, J., & Koča, J. (1999). Analysis of the reaction mechanism and substrate specificity of haloalkane dehalogenases by sequential and structural comparisons. *Protein Engineering, Design and Selection*, *12*(11), 989–998.

de Castro, E., Sigrist, C. J. A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A., & Hulo, N. (2006). ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research*, *34*(suppl_2), W362–W365.

Doesburg, W. van, Eekert, M. H. A. van, Middeldorp, P. J. M., Balk, M., Schraa, G., & Stams, A. J. M. (2005). Reductive dechlorination of β-hexachlorocyclohexane (β-HCH) by a Dehalobacter species in coculture with a Sedimentibacter sp. *FEMS Microbiology Ecology*, *54*(1), 87–95.

Dogra, C., Raina, V., Pal, R., Suar, M., Lal, S., Gartemann, K.-H., Holliger, C., van der Meer, J. R., & Lal, R. (2004). Organization of *Lin* genes and IS*6100* among different strains of hexachlorocyclohexane-degrading *Sphingomonas paucimobilis*: Evidence for horizontal gene transfer. *Journal of Bacteriology*, *186*(8), 2225–2235.

Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(1), 113.

Egorova, D. O., Buzmakov, S. A., Nazarova, E. A., Andreev, D. N., Demakov, V. A., & Plotnikova, E. G. (2017). Bioremediation of Hexachlorocyclohexane-Contaminated Soil by the New *Rhodococcus wratislaviensis* Strain Ch628. *Water, Air, & Soil Pollution*, *228*(5), 183.

Eisenberg, D., Bowie, J. U., Lüthy, R., & Choe, S. (1992). Three-dimensional profiles for analysing protein sequence–structure relationships. *Faraday Discussions*, *93*, 25–34.

Eisenberg, D., Lüthy, R., & Bowie, J. U. (1997). VERIFY3D: Assessment of protein models with three-dimensional profiles. *Methods in Enzymology*, *277*, 396–404.

Endo, R., Kamakura, M., Miyauchi, K., Fukuda, M., Ohtsubo, Y., Tsuda, M., & Nagata, Y. (2005). Identification and characterization of genes involved in the downstream degradation pathway of γ-hexachlorocyclohexane in Sphingomonas paucimobilis UT26. *Journal of Bacteriology*, *187*(3), 847–853.

França, T. C. C. (2015). Homology modeling: An important tool for the drug discovery. *Journal of Biomolecular Structure and Dynamics*, *33*(8), 1780–1793.

Fuentes, M. S., Benimeli, C. S., Cuozzo, S. A., Saez, J. M., & Amoroso, M. J. (2010). Microorganisms capable to degrade organochlorine pesticides. *Current Research, Technology and Education Topics in Applied Microbiology and Microbial Biotechnology*, 1255–1264.

G, M. M., O, L. J., J, M., & W, W. V. (2019). Contamination from Organochlorine Pesticides (OCPs) and other Pesticides in Agricultural Soils of Buuri, Imenti South and Imenti North Sub counties, Meru County Agroecosystem in Kenya. *Journal of Agriculture*, *3*(1), 1–20.

Garg, N., Lata, P., Jit, S., Sangwan, N., Singh, A. K., Dwivedi, V., Niharika, N., Kaur, J., Saxena, A., Dua, A., Nayyar, N., Kohli, P., Geueke, B., Kunz, P., Rentsch, D., Holliger, C., Kohler, H.-P. E., & Lal, R. (2016). Laboratory and field scale bioremediation of hexachlorocyclohexane (HCH) contaminated soils by means of bioaugmentation and biostimulation. *Biodegradation*, *27*(2), 179–193.

Girish, K., & Kunhi, A. A. M. (2013). Microbial degradation of gamma-hexachlorocyclohexane (lindane). *African Journal of Microbiology Research*, *7*(17), 1635–1643.

Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95–98.

Hsieh, P.C., & Vaisvila, R. (2013). Protein Engineering: Single or Multiple Site-Directed Mutagenesis. In J. C. Samuelson (Ed.), *Enzyme Engineering* (Vol. 978, pp. 173–186). Humana Press.

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754–755.

Humphreys, E. H., Janssen, S., Heil, A., Hiatt, P., Solomon, G., & Miller, M. D. (2008). Outcomes of the California Ban on Pharmaceutical Lindane: Clinical and Ecologic Impacts. *Environmental Health Perspectives*, *116*(3), 297–302.

Imai, R., Nagata, Y., Senoo, K., Wada, H., Fukuda, M., Takagi, M., & Yano, K. (1989). Dehydrochlorination of γ-Hexachlorocyclohexane (γ-BHC) by γ-BHC-Assimilating *Pseudomonas paucimobilis*. *Agricultural and Biological Chemistry*, *53*(7), 2015–2017.

Jaroszewski, L. (2009). Protein structure prediction based on sequence similarity. In *Biomedical Informatics* (pp. 129–156).

Kaur, I., Gaur, V. K., Regar, R. K., Roy, A., Srivastava, P. K., Gaur, R., Manickam, N., & Barik, S. K. (2021). Plants exert beneficial influence on soil microbiome in a HCH contaminated soil revealing advantage of microbe-assisted plant-based HCH remediation of a dumpsite. *Chemosphere*, *280*, 130690.

Kumari, R., Subudhi, S., Suar, M., Dhingra, G., Raina, V., Dogra, C., Lal, S., van der Meer, J. R., Holliger, C., & Lal, R. (2002). Cloning and characterization of *Lin* genes responsible for the degradation of hexachlorocyclohexane isomers by *Sphingomonas paucimobilis* strain B90. *Applied and Environmental Microbiology*, *68*(12), 6021–6028.

Lal, R., Dadhwal, M., Kumari, K., Sharma, P., Singh, A., Kumari, H., Jit, S., Gupta, S. K., Nigam, A., Lal, D., Verma, M., Kaur, J., Bala, K., & Jindal, S. (2008). *Pseudomonas* sp. to *Sphingobium indicum*: A journey of microbial degradation and bioremediation of Hexachlorocyclohexane. *Indian Journal of Microbiology*, *48*(1), 3–18.

Lal, R., Dogra, C., Malhotra, S., Sharma, P., & Pal, R. (2006). Diversity, distribution and divergence of *Lin* genes in hexachlorocyclohexane-degrading sphingomonads. *Trends in Biotechnology*, *24*(3), 121–130.

Lal, R., Pandey, G., Sharma, P., Kumari, K., Malhotra, S., Pandey, R., Raina, V., Kohler, H.-P. E., Holliger, C., Jackson, C., & Oakeshott, J. G. (2010). Biochemistry of Microbial Degradation of Hexachlorocyclohexane and Prospects for Bioremediation. *Microbiology and Molecular Biology Reviews*, *74*(1), 58–80.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, *26*(2), 283–291.

Laskowski, R. A., MacArthur, M. W., & Thornton, J. M. (2006). *PROCHECK: Validation of protein-structure coordinates*.

Lüthy, R., Bowie, J. U., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, *356*(6364), 83–85.

Madadi, V. O., Wandiga, S. O., & Mavuti, K. M. (2017). *Organochlorine Pesticides Residues in Lake Naivasha Catchment Water*.

Manickam, N., Misra, R., & Mayilraj, S. (2007). A novel pathway for the biodegradation of γ-hexachlorocyclohexane by a *Xanthomonas* sp. Strain ICH12. *Journal of Applied Microbiology*, *102*(6), 1468–1478.

Manickam, N., Reddy, M. K., Saini, H. S., & Shanker, R. (2008). Isolation of hexachlorocyclohexane-degrading *Sphingomonas* sp. By dehalogenase assay and characterization of genes involved in γ-HCH degradation. *Journal of Applied Microbiology*, *104*(4), 952–960.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Bryant, S. H. (2017). CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research*, *45*(D1), D200–D203.

Marchler-Bauer, A., & Bryant, S. H. (2004). CD-Search: Protein domain annotations on the fly. *Nucleic Acids Research*, *32*(suppl_2), W327–W331.

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C., & Bryant, S. H. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Research*, *43*(D1), D222–D226.

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Lu, F., Marchler, G. H., Mullokandov, M., Omelchenko, M. V., Bryant, S. H. (2011). CDD: A Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research*, *39*(suppl_1), D225–D229.

Marek, J., Vévodová, J., Smatanová, I. K., Nagata, Y., Svensson, L. A., Newman, J., Takagi, M., & Damborský, J. (2000). Crystal Structure of the Haloalkane Dehalogenase from *Sphingomonas paucimobilis* UT26 *Biochemistry*, *39*(46), 14082–14086.

Melo, F., Devos, D., Depiereux, E., & Feytmans, E. (1997). ANOLEA: A www server to assess protein structures. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, *5*, 187–190.

Miyauchi, K., Adachi, Y., Nagata, Y., & Takagi, M. (1999). Cloning and Sequencing of a Novel meta-Cleavage Dioxygenase Gene Whose Product Is Involved in Degradation of γ-Hexachlorocyclohexane in *Sphingomonas paucimobilis*. *Journal of Bacteriology*, *181*(21), 6712–6719.

Miyauchi, K., Lee, H.-S., Fukuda, M., Takagi, M., & Nagata, Y. (2002). Cloning and Characterization of LinR, Involved in Regulation of the Downstream Pathway for γ-Hexachlorocyclohexane Degradation in *Sphingomonas paucimobilis* UT26. *Applied and Environmental Microbiology*, *68*(4), 1803–1807.

Mohn, W. W., Mertens, B., Neufeld, J. D., Verstraete, W., & Lorenzo, V. D. (2006). Distribution and phylogeny of hexachlorocyclohexane-degrading bacteria in soils from Spain. *Environmental Microbiology*, *8*(1), 60–68.

Muhammed, M. T., & Aki-Yalcin, E. (2019). Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chemical Biology & Drug Design*, *93*(1), 12–20.

Munsamy, G., & Soliman, M. E. S. (2017). Homology Modeling in Drug Discovery-an Update on the Last Decade. *Letters in Drug Design & Discovery*, *14*(9), 1099–1111.

Nagata, Y., Endo, R., Ito, M., Ohtsubo, Y., & Tsuda, M. (2007). Aerobic degradation of lindane (γ-hexachlorocyclohexane) in bacteria and its biochemical and molecular basis. *Applied Microbiology and Biotechnology*, *76*(4), 741.

Nagata, Y., Miyauchi, K., & Takagi, M. (1999). Complete analysis of genes and enzymes for γ-hexachlorocyclohexane degradation in *Sphingomonas paucimobilis* UT26. *Journal of Industrial Microbiology and Biotechnology*, *23*(4–5), 380–390.

Nagata, Y., Ohtomo, R., Miyauchi, K., Fukuda, M., Yano, K., & Takagi, M. (1994). Cloning and sequencing of a 2,5-dichloro-2,5-cyclohexadiene-1,4-diol dehydrogenase gene involved in the degradation of gamma-hexachlorocyclohexane in *Pseudomonas paucimobilis*. *Journal of Bacteriology*, *176*(11), 3117–3125.

Nayyar, N., & Lal, R. (2016). Hexachlorocyclohexane Contamination and Solutions: Brief History and Beyond. Emerging Model to Study Evolution of Catabolic Genes and Pathways. *Journal of Bioremediation & Biodegradation*, *07*(02).

Novak, H. R., Sayer, C., Isupov, M. N., Gotz, D., Spragg, A. M., & Littlechild, J. A. (2014). Biochemical and structural characterization of a haloalkane dehalogenase from a marine *Rhodobacteraceae*. *FEBS Letters*, *588*(9), 1616–1622.

Oakley, A. J., Klvaňa, M., Otyepka, M., Nagata, Y., Wilce, M. C. J., & Damborský, J. (2004). Crystal Structure of Haloalkane Dehalogenase LinB from *Sphingomonas paucimobilis* UT26 at 0.95 Å Resolution: Dynamics of Catalytic Residues. *Biochemistry*, *43*(4), 870–878.

Okai, M., Kubota, K., Fukuda, M., Nagata, Y., Nagata, K., & Tanokura, M. (2010). Crystal Structure of γ-Hexachlorocyclohexane Dehydrochlorinase LinA from *Sphingobium japonicum* UT26. *Journal of Molecular Biology*, *403*(2), 260–269.

Osoro, E., Wandiga, S. O., Abongo, D. A., Madadi, V. O., & Macharia, J. W. (2016). Organochlorine pesticides residues in water and sediment from Rusinga Island, Lake Victoria, Kenya. *Journal of Applied Chemistry*, *9*, 56–63.

Pal, R., Bala, S., Dadhwal, M., Kumar, M., Dhingra, G., Prakash, O., Prabagaran, S. R., Shivaji, S., Cullum, J., Holliger, C., & Lal, R. (2005). Hexachlorocyclohexane-degrading bacterial strains *Sphingomonas paucimobilis* B90A, UT26 and Sp+, having similar *Lin* genes, represent three distinct species, *Sphingobium indicum* sp. Nov., *Sphingobium japonicum* sp. Nov. And *Sphingobium francense* sp. Nov., and reclassification of [*Sphingomonas*] *chungbukensis* as *Sphingobium chungbukense* comb. Nov. *International Journal of Systematic and Evolutionary Microbiology*, *55*(5), 1965–1972.

Pant, B., Ojha, G. P., Kim, H.-Y., Park, M., & Park, S.-J. (2019). Fly-ash-incorporated electrospun zinc oxide nanofibers: Potential material for environmental remediation. *Environmental Pollution*, *245*, 163–172.

Pearce, S. L., Oakeshott, J. G., & Pandey, G. (2015). Insights into Ongoing Evolution of the Hexachlorocyclohexane Catabolic Pathway from Comparative Genomics of Ten *Sphingomonadaceae* Strains. *G3: Genes, Genomes, Genetics*, g3.114.015933.

Persson, B., Krook, M., & Jörnvall, H. (1991). Characteristics of short-chain alcohol dehydrogenases and related enzymes. *European Journal of Biochemistry*, *200*(2), 537–543.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612.

Phillips, T. M., Lee, H., Trevors, J. T., & Seech, A. G. (2006). Full-scale in situ bioremediation of hexachlorocyclohexane-contaminated soil. *Journal of Chemical Technology & Biotechnology*, *81*(3), 289–298.

Phillips, T. M., Seech, A. G., Lee, H., & Trevors, J. T. (2005). Biodegradation of hexachlorocyclohexane (HCH) by microorganisms. *Biodegradation*, *16*(4), 363–392.

Quintero, J. C., Moreira, M. T., Feijoo, G., & Lema, J. M. (2005). Anaerobic degradation of hexachlorocyclohexane isomers in liquid and soil slurry systems. *Chemosphere*, *61*, 528–536.

Raina, V., Rentsch, D., Geiger, T., Sharma, P., Buser, H. R., Holliger, C., Lal, R., & Kohler, H.-P. E. (2008). New metabolites in the degradation of α-and γ-hexachlorocyclohexane (HCH): Pentachlorocyclohexenes are hydroxylated to cyclohexenols and cyclohexenediols by the haloalkane dehalogenase LinB from *Sphingobium indicum* B90A. *Journal of Agricultural and Food Chemistry*, *56*(15), 6594–6603.

Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, *19*(12), 1572–1574.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, *61*(3), 539–542.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection*, *12*(2), 85–94.

Sadowski, M. I., & Jones, D. T. (2007). Benchmarking template selection and model quality assessment for high-resolution comparative modeling. *Proteins: Structure, Function, and Bioinformatics*, *69*(3), 476–485.

Sahu, S. K., Patnaik, K. K., Sharmila, M., & Sethunathan, N. (1990). Degradation of alpha-, beta-, and gamma-hexachlorocyclohexane by a soil bacterium under aerobic conditions. *Applied and Environmental Microbiology*, *56*(11), 3620–3622.

Sambrook, J., & Russell, D. W. (2006). Southern Blotting: Capillary Transfer of DNA to Membranes. *Cold Spring Harbor Protocols*, *2006*(1).

Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins: Structure, Function, and Bioinformatics*, *17*(4), 355–362.

Somvanshi, P., Singh, V., & Seth, P. K. (2008). Phylogenetic investigation of *Lin* genes involved in degradation of hexachlorocyclohexane (HCH). *The International Journal of Toxicology*, *4*(2).

Tabata, M., Endo, R., Ito, M., Ohtsubo, Y., Kumar, A., Tsuda, M., & Nagata, Y. (2011). The *Lin* genes for γ-hexachlorocyclohexane degradation in *Sphingomonas* sp.

MM-1 proved to be dispersed across multiple plasmids. *Bioscience, Biotechnology, and Biochemistry*, *75*(3), 466–472.

Verma, H., Bajaj, A., Kumar, R., Kaur, J., Anand, S., Nayyar, N., Puri, A., Singh, Y., Khurana, J. P., & Lal, R. (2017). Genome Organization of *Sphingobium indicum* B90A: An Archetypal Hexachlorocyclohexane (HCH) Degrading Genotype. *Genome Biology and Evolution*, *9*(9), 2191–2197.

Vihinen, M. (2021). *Guidelines for reporting protein modelling studies Authorea*. [Preprint].

Vijgen, J. (2006). The legacy of lindane HCH isomer production. *Main Report. IHPA, January*, *383*, 384.

Vijgen, J., Abhilash, P. C., Li, Y. F., Lal, R., Forter, M., Torres, J., Singh, N., Yunus, M., Tian, C., & Schäffer, A. (2011). Hexachlorocyclohexane (HCH) as new Stockholm Convention POPs−A global perspective on the management of Lindane and its waste isomers. *Environmental Science and Pollution Research*, *18*(2), 152–162.

Vijgen, J., Yi, L. F., Forter, M., Lal, R., & Weber, R. (2006). The legacy of lindane and technical HCH production. *Organohalogen Compounds*, *68*, 899–904.

Vincze, T., Posfai, J., & Roberts, R. J. (2003). NEBcutter: A program to cleave DNA with restriction enzymes. *Nucleic Acids Research*, *31*(13), 3688–3691.

Vyas, V. K., Ukawala, R. D., Ghate, M., & Chintha, C. (2012). Homology Modeling a Fast Tool for Drug Discovery: Current Perspectives. *Indian Journal of Pharmaceutical Sciences*, *74*(1), 1–17.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research*, *46*(W1), W296–W303.

Zdravkovski, Z. (2004). Theoretical study of the stability of hexachloro-and hexafluorocyclohexane isomers. *Bulletin of the Chemists and. Technologists of. Macedonia*, *23*(2), 131–137.

APPENDICES

APPENDIX 1A. Percent sequence identities and similarities of HCH dehydrochlorinase (LinA) from *Sphingobium* sp. S6 (QGJ16213.1) and *Sphingobium* sp. S8 (QGJ16206.1) to similar sequences from other HCH-degrading bacteria, based on BLOSUM62 substitution matrix.

% Seq Identity (Seq Simil.)	**QGJ16213.1**	**QGJ16206.1**	**BAI96690.1**	**APL95055.1**	sp|P59766.2	**AAU11089.2**	**AAT00794.1**	ACV91871.1	ABP93360.1
QGJ16213.1	ID	**92.3 (92.9)**	**98.7 (98.7)**	**98.7 (98.7)**	87.1 (89.7)	**98.7 (98.7)**	**98.7 (98.7)**	97.4 (98.0)	91.6 (94.2)
QGJ16206.1	**92.3 (92.9)**	ID	**93.5 (94.2)**	**93.5 (94.2)**	86.3 (89.6)	**93.5 (94.2)**	**93.5 (94.2)**	92.3 (93.5)	85.8 (89.1)

% Seq Identity (Seq Similar.)	ABP93362.1	ABG77566.1	**CAI43920.1**	EQB30645.1	CAI43917.1	CAI43918.1	**CAI43919.1**	EQB02014.1
QGJ16213.1	96.7 (96.7)	96.1 (96.7)	94.2 (94.2)	96.1 (96.1)	97.4 (97.4)	98.0 (98.0)	**98.7 (98.7)**	91.6 (94.8)
QGJ16206.1	92.3 (92.9)	92.3 (92.9)	**93.5 (94.2)**	91.6 (92.3)	92.3 (92.9)	92.9 (93.5)	**93.5 (94.2)**	86.5 (90.3)

NB: ID – Identical. The protein IDs for the respective dehydrochlorinase (LinA) sequences are as follows:

QGJ16213.1: HCH dehydrochlorinase LinA (*Sphingobium* sp. S6); **QGJ16206.1**: HCH dehydrochlorinase LinA (*Sphingobium* sp. S8); **BAI 9660.1**: Gamma–HCH dehydrochlorinase LinA (*Sphingobium japonicum* UT26S); **APL95055.1**: HCH dehydrochlorinase (*Sphingobium indicum* B90A); **sp|P59766.2**: HCH dehydrochlorinase 1 (*Sphingobium indicum* B90A); **AAU11089.2**: HCH dehydrochlorinase (*Sphingobium francense* Sp+); **AAT00794.1**: lindane (*Rhodanobacter lindaniclasticus*); **ACV91871.1**: dehydrochlorinase (*Sphingomonas* sp. HZ-1); **ABP93360.1**: HCH-dehydrochlorinase LinAa (*Pseudomonas aeruginosa*); **ABP93362.1**: HCH–dehydrochlorinase LinAb (*Pseudomonas aeruginosa*); **ABG77566.1**: dehydrochlorinase (*Sphingomonas* sp. NM05); **CAI43920.1**: dehydrochlorinase (*Sphingomonas* sp. α1-2); **EQB30645.1**: Gamma–HCH dehydrochlorinase (*Sphingobium ummariense* RL-3); **CAI43917.1**: dehydrochlorinase (*Sphingomonas* sp. γ1-7); **CAI43918.1**: dehydrochlorinase (*Sphingomonas* sp. γ12-7); **CAI43919.1**: dehydrochlorinase (*Sphingomonas* sp. γ16-1); **EQB02014.1**: Gamma–HCH dehydrochlorinase (*Sphingobium* sp. HDIP04).

APPENDIX 1B. Percent sequence identities and similarities of haloalkane dehalogenase (LinB) sequences from *Sphingobium* sp. S6 (QGJ16214.1) and *Sphingobium* sp. S8 (QGJ16207.1) to similar sequences from other HCH-degrading bacteria, based on BLOSUM62 substitution matrix.

% Seq Identity (Seq Similar.)	**QGJ16214.1**	**QGJ16207.1**	APL96138.1	BAI96793.1	**AMK21182.1**	BAF80333.1	AAX07227.1	BAF80336.1
QGJ16214.1	ID	**96.6 (96.9)**	97.6 (98.3)	97.9 (98.9)	**98.6 (99.3)**	98.3 (99.3)	97.9 (98.6)	97.9 (98.9)
QGJ16207.1	**96.6 (96.9)**	ID	94.9 (95.9)	95.2 (96.6)	**95.9 (96.9)**	95.6 (96.9)	95.2 (96.2)	95.2 (96.6)
% Seq Identity (Seq Similar.)	ACV91872.1	BAF80345.1	**ABP93361.1**	ABI93216.1	ABG77567.1	EQB33511.1	KMS56815.1	KER36945.1
QGJ16214.1	97.6 (98.6)	97.6 (98.3)	**98.6 (99.3)**	94.5 (95.6)	95.9 (96.9)	97.9 (98.6)	97.2 (98.3)	97.2 (97.9)
QGJ16207.1	94.9 (96.2)	94.9 (95.9)	**95.9 (96.9)**	94.5 (95.8)	93.2 (94.5)	95.2 (96.2)	94.5 (95.9)	94.5 (95.6)

NB: ID – Identical. The protein IDs for the respective haloalkane dehalogenase sequences are as follows:

QGJ16214.1: haloalkane dehalogenase LinB (*Sphingobium* sp. S6); **QGJ16207.1**: haloalkane dehalogenase (*Sphingobium* sp. S8); **APL96138.1**: haloalkane dehalogenase (*Sphingobium indicum* B90A); **BAI96793.1**: 1,4-TCDN hydrolase LinB (*Sphingobium japonicum* UT26S); **AMK21182.1**: haloalkane dehalogenase (*Sphingobium* sp. TKS); **BAF80333.1**: haloalkane dehalogenase (*Sphingobium* sp. SS04-1); **AAX07227.1**: haloalkane dehalogenase (*Sphingobium francense* Sp+); **BAF80336.1**: haloalkane dehalogenase (*Sphingobium* sp. SS04-2); **ACV91872.1**: haloalkane dehalogenase (*Sphingomonas* sp. HZ-1); **BAF80345.1**: haloalkane dehalogenase (*Sphingobium* sp. SS04-5); **ABP93361.1**: HCH-dehalogenase LinBa (*Pseudomonas aeruginosa*); **ABI93216.1**: LinB, partial (*Xanthomonas* sp. ICH12); **ABG77567.1** haloalkane dehalogenase (*Sphingomonas* sp. NM05); **EQB33511.1**: haloalkane dehalogenase (*Sphingobium ummariense RL-3*); **KMS56815.1**: haloalkane dehalogenase (*Sphingobium czechense LL01*); **KER36945.1**: haloalkane dehalogenase (*Sphingobium lucknowense F2*).

APPENDIX 1C. Percent sequence identities and similarities of 2,5-DDOL dehydrogenase (LinC) sequences from *Sphingobium* sp. S6 (QGJ16215.1) and *Sphingobium* sp. S8 (QGJ16208.2) to similar sequences from other HCH-degrading bacteria, based on BLOSUM62 substitution matrix

% Seq Identity (Seq Similarity)	**QGJ16215.1**	**QGJ16208.2**	**BAI95393.1**	ABE98169.1	AMK21445.1	**ABG77568.1**
QGJ16215.1	ID	**99.2 (99.6)**	**100 (100)**	99.6 (99.6)	99.6 (99.6)	**100 (100)**
QGJ16208.2	**99.2 (99.6)**	ID	**99.2 (99.6)**	98.8 (99.2)	98.8 (99.2)	**99.2 (99.6)**
% Seq Identity (Seq. Similarity)	ACV91873.1	AAN64242.1	BAA03444.1	ABP93367.1	EPR12466.1	EQB07933.1
QGJ16215.1	98.4 (98.8)	99.2 (99.2)	98.8 (99.2)	99.2 (99.2)	99.2 (99.2)	98.0 (98.4)
QGJ16208.2	97.6 (98.4)	98.4 (98.8)	98.0 (98.8)	98.4 (98.8)	98.4 (98.8)	97.2 (98.0)

NB: ID – Identical. The protein IDs for the respective dehydrogenases are described as follows: **QGJ16215.1**: 2,5-DDOL dehydrogenase LinC (*Sphingobium* sp. S6); **QGJ16208.2**: 2,5-DDOL dehydrogenase LinC (*Sphingobium* sp. S8); **BAI95393.1** 2,5-DDOL dehydrogenase LinC (*Sphingobium japonicum* UT26S); **ABE98169.1**: 2,5-DDOL dehydrogenase (*Sphingomonas* sp. BHC-A); **AMK21445.1**: 2,5-DDOL dehydrogenase LinC (*Sphingobium* sp. TKS); **ABG77568.1**: short-chain alcohol dehydrogenase (*Sphingomonas sp. NM05*); **ACV91873.1**: short-chain alcohol dehydrogenase (*Sphingomonas* sp. HZ-1); **AAN64242.1**: 2,5-DDOL dehydrogenase (*Sphingomonas paucimobilis* B90); **BAA03444.1**: 2,5-DDOL dehydrogenase (*Sphingobium japonicum*); **ABP93367.1**: HCH-dehydrogenase LinC (*Pseudomonas aeruginosa*); **EPR12466.1**: short-chain dehydrogenase (*Sphingobium chinhatense* IP26); **ACV91873.1**: short-chain alcohol dehydrogenase (*Sphingomonas* sp. HZ-1); **EQB07933.1**: short-chain dehydrogenase (*Sphingobium* sp. HDIP04).

APPENDIX 1D. Percent sequence identities and similarities of 2,5-DCHQ reductive dechlorinase (LinD) sequences from *Sphingobium* sp. S6 (QGJ16216.1) and *Sphingobium* sp. S8 (QGJ16209.1) to similar sequences from other HCH-degrading bacteria, based on BLOSUM62 substitution matrix

% Seq Identity (Seq Similarity)	**QGJ16216.1**	**QGJ16209.1**	**sp|D4Z909.1**	ABP93365.1	ABE03743.1	ABE98170.1	ACV91874.1
QGJ16216.1	ID	**98.5 (99.4)**	**99.4 (100)**	99.1 (100)	98.2 (99.1)	97.6 (98.8)	97.3 (98.5)
QGJ16209.1	**98.5 (99.4)**	ID	**99.1 (99.4)**	98.8 (99.4)	97.9 (98.5)	97.3 (98.2)	97.1 (97.9)

NB: ID Identical. The protein IDs of the respective LinD protein sequences are as follows: **QGJ16216.1**: 2,5-DCHQ reductive dechlorinase LinD (*Sphingobium* sp. S6); **QGJ16209.1**: 2,5-DCHQ reductive dechlorinase LinD (*Sphingobium* sp. S8); **sp|D4Z909.1**: 2,5-DCHQ reductive dechlorinase (*Sphingobium japonicum* UT26S); **ABP93365.1**: HCH-reductive dechlorinase LinD (*Pseudomonas aeruginosa*); **ABE03743.1**: 2,5-DCHQ reductive dechlorinase (*Sphingomonas* sp. JQL4-5); **ABE98170.1**: 2,5-DCHQ reductive dechlorinase (*Sphingomonas* sp. BHC-A); **ACV91874.1**: reductive dehalogenase (*Sphingomonas* sp. HZ-1).

APPENDIX 1E. Percent sequence identities and similarities of (chloro) hydroquinone 1,2-dioxygenase (LinE) sequences from *Sphingobium* sp. S6 (QGJ16217.1) and *Sphingobium* sp. S8 (QGJ16210.1) to similar sequences from other HCH-degrading bacteria, based on BLOSUM62 substitution matrix

% Seq Identity (Seq Similarity)	**QGJ16217.1**	**QGJ16210.1**	**Q9WXE6.1**	ABD66585.1	ABG77570.1	ABP93364.1	ACV91875.1
QGJ16217.1	ID	**97.8 (98.1)**	**99.6 (99.6)**	99.3 (99.6)	99.3 (99.3)	99.3 (99.3)	97.1 (98.4)
QGJ16210.1	**97.8 (98.1)**	ID	**98.1 (98.4)**	97.8 (98.4)	97.8 (98.1)	97.8 (98.1)	95.6 (97.1)

NB: ID – Identical. The protein IDs of the respective LinE protein sequences are described as follows: **QGJ16217.1**: (chloro) hydroquinone 1,2-dioxygenase LinE (*Sphingobium* sp. S6); **QGJ16210.1**: (chloro) hydroquinone 1,2-dioxygenase LinE (*Sphingobium* sp. S8); **Q9WXE6.1**: (chloro) hydroquinone 1,2-dioxygenase LinE (*Sphingobium japonicum* UT26S); **ABD66585.1**: hydroquinone meta-cleavage dioxygenase (*Sphingomonas* sp. BHC-A); **ABG77570.1**: meta-cleavage dioxygenase (*Sphingomonas* sp. NM05); **ABP93364.1** HCH-ring cleavage dioxygenase LinE (*Pseudomonas aeruginosa*); **ACV91875.1**: meta-cleavage dioxygenase (*Sphingomonas* sp. HZ-1).

APPENDIX 1F. Percent sequence identities and similarities of LysR-type transcriptional regulator (LinR) sequences from *Sphingobium* sp. S6 (QGJ16218.1) and *Sphingobium* sp. S8 (QGJ16211.2) to similar sequences from other HCH-degrading bacteria, based on BLOSUM62 substitution matrix

% Seq Identity (Seq Similarity)	**QGJ16218.1**	**QGJ16211.2**	**Q9ZN79.3**	ACV91876.1	ABP93363.1
QGJ16218.1	ID	**94.0 (94.7)**	**98.6 (99.0)**	97.0 (97.6)	98.3 (98.6)
QGJ16211.2	**94.0 (94.7)**	ID	94.3 (94.7)	92.7 (93.3)	94.0 (94.3)
% Seq Identity (Seq Similarity)	EPR17852.1	KER36896.1	**EQA99717.1**	EQB08246.1	AMW03649.1
QGJ16218.1	97.3 (97.7)	97.3 (97.6)	96.0 (96.6)	94.7 (94.7)	98.0 (98.3)
QGJ16211.2	93.1 (93.4)	96.6 (96.9)	**97.2 (97.9)**	90.8 (90.8)	93.7 (94.0)

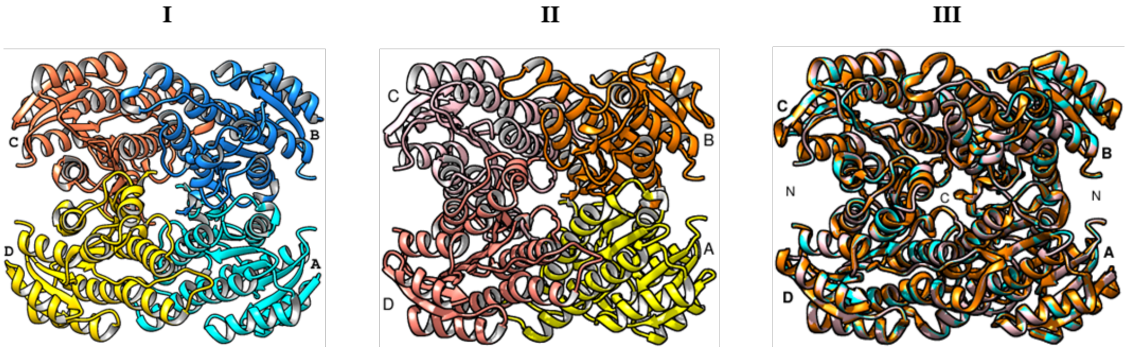
NB: ID – Identical. The protein IDs of the respective LinR sequences are as follows: **QGJ16218.1**: LysR-type transcriptional regulator LinR (*Sphingobium* sp. S6); **QGJ16211.2**: LysR-type transcriptional regulator LinR (*Sphingobium* sp. S8); **Q9ZN79.3**: HTH-type transcriptional regulator LinR (*Sphingobium japonicum* UT26S); **ACV91876.1**: transcriptional regulator (*Sphingomonas* sp. HZ-1); **ABP93363.1**: HCH-transcriptional regulator LinR (*Pseudomonas aeruginosa*); **EPR17852.1**: transcriptional regulator, partial (*Sphingobium chinhatense* IP26); **KER36896.1**: transcriptional regulator, partial (*Sphingobium lucknowense* F2); **EQA99717.1**: transcriptional regulator (*Sphingobium baderi* LL03); **EQB08246.1**: transcriptional regulator, partial (*Sphingobium* sp. HDIP04); **KEQ51345.1**: LysR family transcriptional regulator (*Sphingobium chlorophenolicum*); **AMW03649.1**: LinR (*Chlorohydroquinone sensing module vector*).

APPENDIX 1G. Percent sequence identities and similarities of 2,5-DDOL dehydrogenase (LinX) sequences from *Sphingobium* sp. S6 (QGJ16219.1) and *Sphingobium* sp. S8 (QGJ16212.1) to similar sequences from other HCH-degrading bacteria, based on BLOSUM62 substitution matrix.

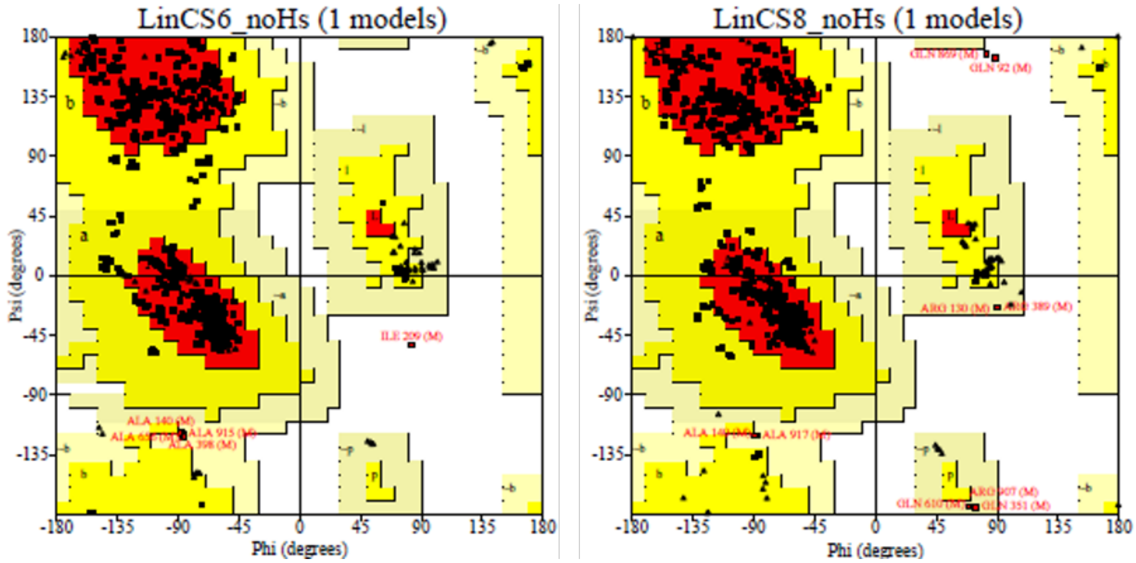
% Seq Identity (Seq Simil.)	**QGJ16219.1**	**QGJ16212.1**	**QGJ16215.1**	**QGJ16208.2**	**BAI96692.1**	EPR18614.1	EQA97108.1	ABE98171.1	BAA04939.1	KMS51485.1
QGJ16219.1	ID	**100 (100)**	**31.7 (50.1)**	**31.3 (50.1)**	**100 (100)**	99.6 (99.6)	99.6 (99.6)	99.2 (99.6)	99.2 (99.6)	99.2 (99.6)
QGJ16212.1	**100 (100)**	ID	**31.7 (50.1)**	**31.3 (50.1)**	**100 (100)**	99.6 (99.6)	99.6 (99.6)	99.2 (99.6)	99.2 (99.6)	99.2 (99.6)

NB: ID – Identical. The protein IDs of the respective LinX sequences are described as follows: **QGJ16219.1**: 2,5-DDOL dehydrogenase LinX (*Sphingobium* sp. S6); **QGJ16212.1**: 2,5-DDOL dehydrogenase LinX (*Sphingobium* sp. S8); **BAI96692.1**: 2,5-DDOL dehydrogenase LinX (*Sphingobium japonicum* UT26S); **AAR05958.1**: 2,5-DDOL dehydrogenase LinX2 (*Sphingobium indicum* B90A); **EPR18614.1**: 2,5-DDOL dehydrogenase, partial (*Sphingobium chinhatense* IP26); **EQA97108.1**: 2,5-DDOL dehydrogenase, partial (*Sphingobium quisquiliarum* P25); **ABE98171.1**: 2,5-DDOL dehydrogenase (*Sphingomonas* sp. BHC-A); **BAA04939.1**: 2,5-DDOL dehydrogenase (*Sphingobium japonicum*); **KMS51485.1**: 2,5-DDOL dehydrogenase (*Novosphingobium barchaimii* LL02).

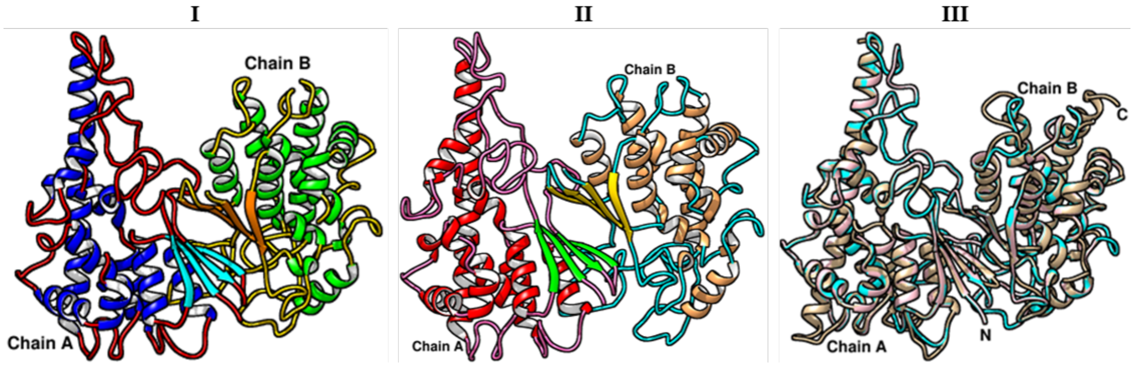
APPENDIX 2A. Modeled 3D structures of 2,5-DDOL dehydrogenase LinC from *Sphingobium* sp. S6 (LinCS6) and *Sphingobium* sp. S8. **I**) Tetrameric structure of LinCS6. Each subunit is rendered in a separate color; chain A (cyan), chain B (dodger blue), chain C (coral), and chain D (gold). **II**) Tetrameric structure of LinCS8. Each subunit is rendered in a separate color; chain A (yellow), chain B (orange), chain C (pink), and chain D (salmon). **III**) LinCS6 (cyan) and LinCS8 (pink) models superimposed with the template, 5X8H (orange). **N** and **C** are the N- and C-terminal ends, respectively of the protein. Rendering of the images was done in UCSF Chimera v1.15.



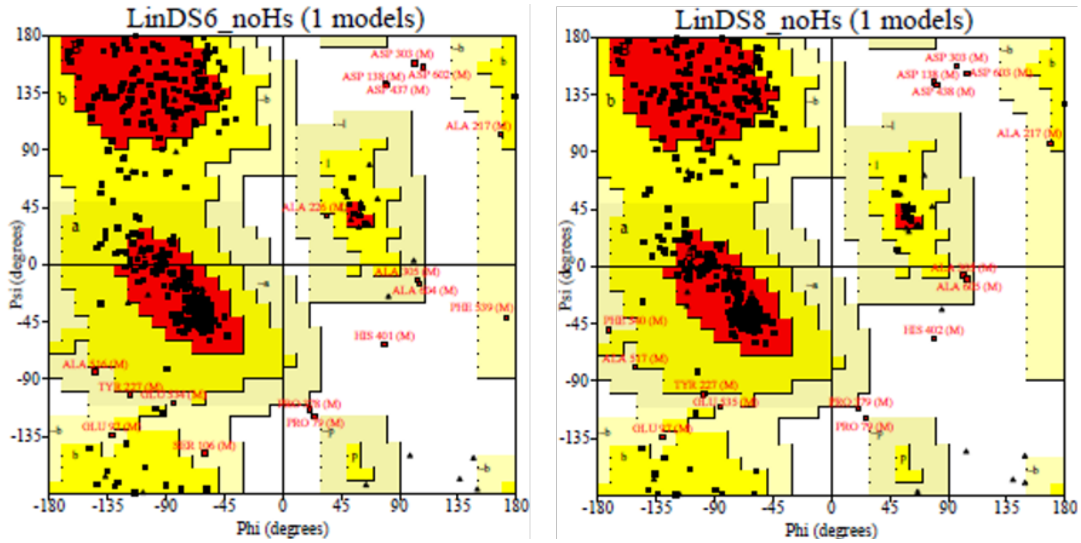
APPENDIX 2B. Ramachandran plots by PROCHECK for LinCS6 and LinCS8 models showing the φ–ψ distribution for the different regions. The core regions (marked A, B, L), additionally allowed regions (marked a, b, l, p), generously allowed regions (marked ~a, ~b, ~l, ~p), and disallowed regions are shown in red, yellow, grey and white colors, respectively. Non-proline and non-glycine residues are represented by black squares and glycine residues by black triangles. Residues in disallowed regions are shown in red.



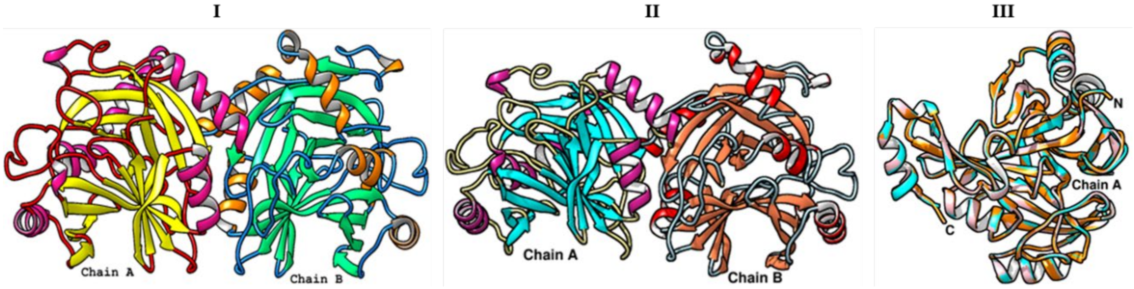
APPENDIX 3A. Modeled 3D structure of LinD from *Sphingobium* sp. S6 (LinDS6) and *Sphingobium* sp. S8 (LinDS8). **I**) Dimeric structure of LinDS6. Helices, strands, and loops (turns) of each LinD protomer are shown in different colors: α-helices (blue), β-strands (cyan), and loops (red) for chain A, and α-helices (green), β-strands (orange), and loops (yellow) for chain B. **II**) Dimeric structure of LinDS8. Helices, strands, and loops of each LinD protomer are shown in different colors: α-helices (red), β-strands (green), and loops (hot pink) for chain A, and α-helices (sandy brown), β-strands (gold), and loops (cyan) for chain B. **III**) LinDS6 (cyan) and LinDS8 (pink) models superimposed with the template 7aia (tan). **N** and **C** represent the N- and C-terminal ends, respectively of the protein. The figures were rendered using UCSF Chimera v1.15.



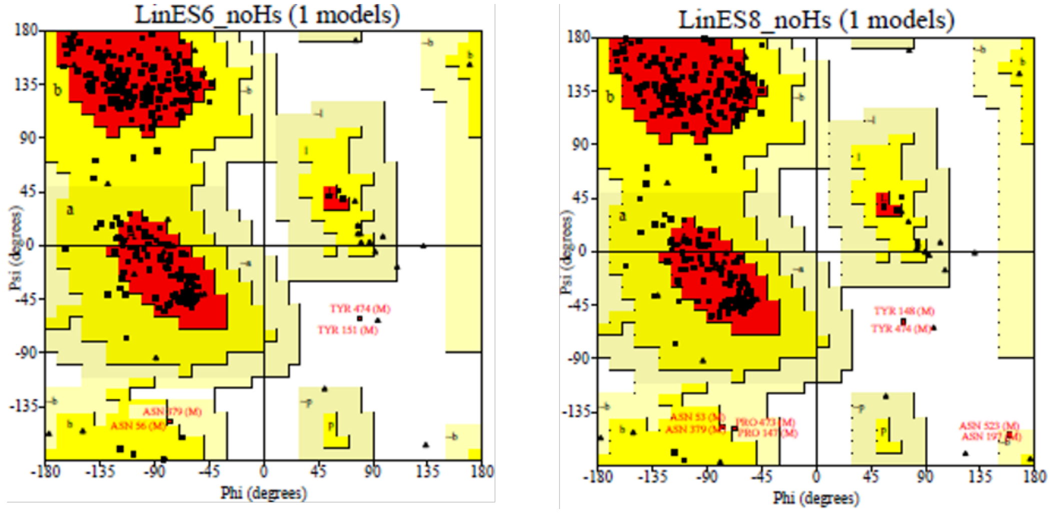
APPENDIX 3B. Ramachandran plots by PROCHECK for LinDS6 and LinDS8 models showing the φ–ψ distribution for the different regions. The core regions (marked A, B, L), additionally allowed regions (marked a, b, l, p), generously allowed regions (marked ~a, ~b, ~l, ~p), and disallowed regions are shown in red, yellow, grey and white colors, respectively. Non-proline and non-glycine residues are represented by black squares and glycine residues by black triangles. Residues in disallowed regions are shown in red.



APPENDIX 4A. Modeled 3D structures of LinE from *Sphingobium* sp. S6 (LinES6) and *Sphingobium* sp. S8 (LinES8). **I**) Dimeric structure of LinES6. Helices and strands of each LinES6 protomer are shown in different colors; Chain A: β-strands (yellow), α-helices (hot pink), and loops (red); Chain B: β-strands (light green), α-helices (orange), and loops (dodger blue). **II**) Dimeric structure of LinES8. Helices and strands of each LinES8 protomer are shown in different colors; Chain A: β-strands (cyan), α-helices (violet red), and loops (khaki); Chain B: β-strands (coral), α-helices (red), and loops (light blue). **III**) Chain A of the LinES6 (cyan) and LinES8 (pink) models superimposed with the template 4huz (orange). **N** and **C** represent the N- and C-terminal ends, respectively of the protein. The figures were rendered using UCSF Chimera v1.15.



APPENDIX 4C. Ramachandran plots by PROCHECK for LinES6 and LinES8 models showing the φ–ψ distribution for the different regions. The core regions (marked A, B, L), additionally allowed regions (marked a, b, l, p), generously allowed regions (marked ~a, ~b, ~l, ~p), and disallowed regions are shown in red, yellow, grey and white colors, respectively. Non-proline and non-glycine residues are represented by black squares and glycine residues by black triangles. Residues in disallowed regions are shown in red.



APPENDIX 5. A summary of the validation parameters by PROCHECK, MOLPROBITY, VERIFY3D, PROSAII, and ERRAT used to evaluate the structural quality of LinC, LinD, and LinE homology models generated by SWISS-MODEL

Validation parameters used to assess quality of 3D models	Theoretical three-dimensional (3D) models					
	LinCS6	**LinCS8**	**LinDS6**	**LinDS8**	**LinES6**	**LinES8**
PROCHECK:						
Residues in most favoured regions [A, B, L], %	763, 91.6%	767, 91.7%	431, 86.2%	434, 86.5%	466, 88.6%	478, 90.2%
Residues in additionally allowed regions [a, b, l, p], %	65, 7.8%	60, 7.2%	54, 10.8%	55, 11.0%	56, 10.6%	46, 8.7%
Residues in generously allowed regions [~a, ~b, ~l, ~p], %	4, 0.5%	7, 0.8%	10, 2.0%	8, 1.6%	2, 0.4%	4, 0.8%
Residues in disallowed regions, %	1, 0.1%	2, 0.2%	5, 1.0%	5, 1.0%	2, 0.4%	2, 0.4%
Procheck G-factor[a] (φ/ψ) Z-score[g]	0.20	0.55	-0.79	-0.67	-1.18	-1.42
Procheck G-factor[a] (all dihedral angles) Z-score[g]	0.24	0.83	-0.95	-0.95	-1.42	-1.66
Overall Procheck G-factor	-0.10	-0.06	-0.17	-0.18	-0.18	-0.26
MOLPROBITY:						
MolProbity score^	1.72	1.21	2.12	1.89	1.52	1.80
MolProbity clashscore	-0.48	0.65	-0.82	0.27	0.05	-0.10
VERIFY3D: 3D-1D score ≥ 0.2 (%), Z-score[g]	81.07% -4.82	75.00% -4.98	59.86%, -4.65	55.86%, -4.49	91.21%, -2.89	93.35%, -3.05
PROSAII:						
ProSAII (-ve) Z-score[g]	-0.54	-0.33	-0.83	-0.83	-1.28	-1.28
ERRAT:						
Overall quality factor (%)	93.98	95.62	79.92	82.88	90.47	84.87

[a] Residues selected using the parameter; S(phi)+S(psi)>=1.8, for dihedral angle order.

[g] Determined by comparing the standard deviation and mean of sets of 252 X-ray crystal structures of less than 500 residues; a positive value indicates a 'better' score.

^ MolProbity score is a combination of the clashscore, Ramachandran and rotamer assessments into one score, normalized by comparing with high resolution X-ray structures.
Generated using PSVS v1.5 and SAVES v6.