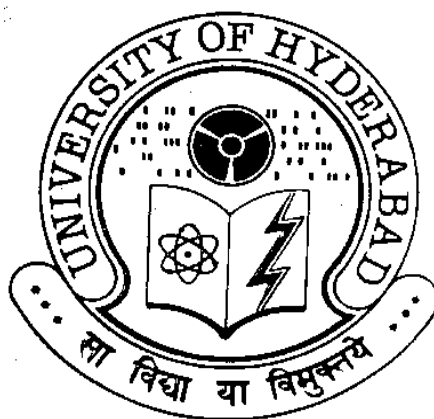


# **Studies on the PE and PPE Proteins of *Mycobacteria***

*A thesis*  
*Submitted for the degree of*  
**DOCTOR OF PHILOSOPHY**

*By*  
**Rafiya Sultana**



**School of Chemistry**  
**University of Hyderabad**  
**Hyderabad – 500 046**  
**INDIA**  
**May 2016**



**I dedicate this thesis to**

My Parents and Family



## Contents

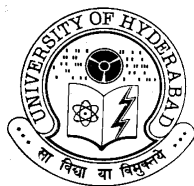
Statement	i
Certificate	ii
Declaration	iii
Acknowledgements	iv
Abbreviations	v
Chapter 1: Introduction to <i>Mycobacterium tuberculosis</i> , PE and PPE proteins, Computational Methods and General Molecular Biology Techniques	1
1.1 Tuberculosis	3
1.1.1 <i>M. tuberculosis</i>	4
1.1.2 BCG Vaccine and Drug Resistance	6
1.1.3 Cell wall virulence factors of <i>M. Tuberculosis</i>	8
1.1.4 Metabolic Pathways	9
1.1.5 The Genome of <i>M. tuberculosis</i>	10
1.1.6 The PE and PPE Multigene Families	10
1.1.7 Classification of PE Family	11
1.1.8 Classification of the PPE Family	11
1.1.9 Pathogenicity of the PE and PPE proteins	12
1.2 Computational Methods	15
1.3 Chemical Libraries	22
1.4 Virtual Screening and Molecular Docking	23
1.5 Molecular Dynamic Simulations	24
1.6 General Methods and Reagents for Biochemical Studies	24
Chapter 2: Prediction of Certain Well-Characterized Domains of Known Functions Within the PE and PPE Proteins of <i>Mycobacteria</i>	41
2.1 Introduction	43
2.2 MATERIALS AND METHODS	46
2.2.1 Sequence Searches – PSI-BLAST	46
2.2.2 Selection of Non-redundant Proteins - CD-HIT	46

2.2.3 Multiple sequence alignment - ClustalX	46
2.2.4 Phylogeny Analysis – MEGA5	46
2.2.5 Protein Fold Recognition - Phyre2	47
2.3 RESULTS AND DISCUSSION	48
2.3.1.1 Hydrolase Domain	54
2.3.1.2 Aspartic Proteinase Domain	54
2.3.1.3 Glucosyl-3-Phosphoglycerate Phosphatase Domain	56
2.3.1.4 Laminaripentaose-Producing Beta-1,3-Glucanase Domain	58
2.3.1.5 Chitinase Domain	61
2.3.1.6 Endoglucanase Domain	62
2.3.1.7 Carbohydrate Binding Domain	64
2.3.1.8 Cytochrome P450 Domain	66
2.3.1.9 Beta-Propeller	69
2.3.2 Beta-Helix	71
2.3.2.1 Acetyl Hydrolase/Cutinase Domain	72
2.3.2.2 Transmembrane Domain	74
2.4 Conclusion	76
Chapter 3: The PE-PPE Domain in <i>Mycobacterium</i> Reveals a Serine $\alpha/\beta$ Hydrolase Fold and Function: An <i>In Silico</i> Analysis	77
3.1 Introduction	79
3.2 Methods	82
3.2.1 NCBI Protein Sequence Databank	82
3.2.2 PSI-BLAST	82
3.2.3 CLUSTALW	82
3.2.4 Signal Peptide Server	82
3.2.5 FUGUE	83
3.2.6 Structure Modeling	83
3.2.7 3D Model Structure Validation	83
3.2.8 MAPSCI	83
3.3 Results And Discussion	84
3.4 Conclusion	99
Chapter 4: The PE16 (Rv1430) of <i>Mycobacterium Tuberculosis</i> is An	101

Esterase Belonging to Serine Hydrolase Superfamily of Proteins	
4.1 Introduction	103
4.2 Materials and Methods	106
4.2.1 Reagents Used	106
4.2.2 Cloning, Expression and Purification of Recombinant Rv1430 and its PE-PPE Domain	106
4.2.3 Cloning, Expression and Purification of Recombinant Ser199Ala Mutant Rv1430 PE-PPE Domain	108
4.2.4 Western Blot	108
4.2.5 Enzymatic Assay	109
4.2.6 Circular Dichroism (CD) Spectroscopy Analysis of Rv1430 and its PE-PPE Domain	110
4.2.7 Enzyme Inhibition	111
4.2.8 Effect of pH and Temperature	111
4.2.9 Effect of Salt	111
4.3 Results and Discussion	112
4.3.1 Cloning, Expression and Purification of Rv1430, PE-PPE Domain and Ser199Ala Mutant PE-PPE Domain	112
4.3.2 CD Spectroscopy Data	117
4.3.3 Rv1430 and its PE-PPE Domain have Esterase Activity	118
4.3.4 Kinetic Properties	120
4.3.5 Temperature Dependence, pH Tolerance and Effect of Salt Concentration on the Esterase Activity of Rv1430 and the PE-PPE Domain	120
4.3.6 Rv1430 belongs to Serine Hydrolase Family of Proteins	122
4.4 Conclusion	125
Chapter 5: Inhibitor Binding Studies of Rv1430 PE-PPE Domain of <i>Mycobacterium Tuberculosis</i> : Virtual Screening and Molecular Dynamic Simulations	127
5.1 Introduction	129

5. 2 Methods	132
5.2.1 Virtual Screening	132
5.2.2 MD Simulations	132
5.2.3 Binding Free Energies Calculations	133
5.3 Results and Discussion:	135
5.3.1 Homology Modeling and Virtual Screening	135
5.3.2 Virtual Screening	136
5.3.3 Molecular Docking	140
5.3.4 Molecular Dynamics Simulations	143
5.4 Conclusions	149
References	151
List of Publications	165
Plagiarism Report	166





**School of Chemistry**  
**University of Hyderabad**  
**Hyderabad – 500 046**

---

### **STATEMENT**

I hereby declare that the matter embodied in this thesis is the result of investigations carried out by me in the School of Chemistry, University of Hyderabad, Hyderabad, under the supervision of **Prof. Lalitha Guruprasad**.

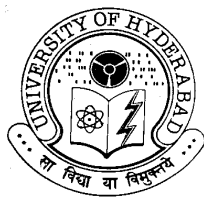
In keeping with the general practice of reporting scientific observations, due acknowledgements have been made whenever the work described is based on the finding of other investigators. Any omission which might have occurred by oversight or error is regretted.

**Hyderabad**

**May 2016**

**Rafiya Sultana**





**School of Chemistry**  
**University of Hyderabad**  
**Hyderabad – 500 046**

---

### **CERTIFICATE**

Certified that the work embodied in this thesis entitled “**Studies on the PE and PPE Proteins of *Mycobacteria***” has been carried out by **Mrs. Rafiya Sultana** under my supervision and the same has not been submitted elsewhere for any degree.

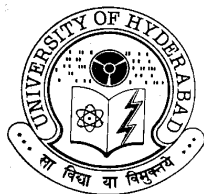
**Hyderabad**

**May 2016**

**Prof. Lalitha Guruprasad**  
**(Thesis Supervisor)**

**Dean**  
**School of Chemistry**





School of Chemistry  
University of Hyderabad  
Hyderabad – 500 046

---

### DECLARATION

I, **Rafiya Sultana** hereby declare that the thesis entitled “**Studies on the PE and PPE Proteins of *Mycobacteria***” submitted by me under the supervision of **Prof. Lalitha Guruprasad** is a bonafide research work which is free from plagiarism. I also declare that it has not been submitted previously in part or in full to this University or any other University or Institution for the award of any degree or diploma. I hereby agree that my thesis can be deposited in Shodganga/INFLIBNET.

A report on plagiarism from the University Library is enclosed.

**Name: Rafiya Sultana**

**Signature:**

**Reg. No.: 08CHPH44**

**Prof. Lalitha Guruprasad**

**(Thesis Supervisor)**



## **ACKNOWLEDGEMENTS**

First and Foremost, I would like to express my deep and sincere gratitude to my research supervisor Prof. Lalitha Guruprasad for making me a part of her research group and this esteemed university and introducing me to the world of computational biology. I sincerely thank her for showering blessings and love and her valuable suggestions, support and encouragement. I am extremely grateful to her for always been approachable, her continuous guidance and allowing me to complete my PhD besides taking up a personal career.

I am extremely grateful to Dr. Sharmistha Banerjee, School of Life sciences, University of Hyderabad for collaboration and providing me an opportunity to learn molecular biology and techniques and allowing me to perform the experiments in her lab.

I express my sincere thanks to the former dean Prof. M.V. Rajasekharan and present dean Prof. M. Durga Prasad for providing infrastructure for carrying out the research work in the school.

I would like to thank my research Doctoral Committee members Prof. M. Durga Prasad and Prof. Susanta Mahapatra for encouraging me in completing the research work. I thank all teachers of school of chemistry.

I am very grateful to CSIR, New Delhi, for providing financial support.

I deeply thank all my labmates Dr. Karunakar, Dr. Gopi, Rajender and Swati Singh, Bala Divya, Sateesh, Divya Jyothi, Mageed and project students Ashwin and Santhosh for their constant support and for providing a peaceful environment in the lab.

I thank my life sciences labmates Dr. Atoshi banerjee, Dr. Ronald Benjamin, Rakesh, Mani Harika, Harini Challa and Arshad Rizvi.

I would like to thank my friends in the School of Chemistry Saritha, Bhargavi and Sowmya and all the research scholars of School of Chemistry and especially who helped me in need.

I would like to thank my Degree College teacher Dr. Veeraiah for his continuous guidance.

My special thanks to my College Principal Dr. R.V. Devadasu garu and colleagues.

I would like to thank all non-teaching staff, School of Chemistry. I thank all the people who have helped me during my research work.

I am very grateful to Almighty for blessing me with beautiful parents Taswar Ali and Sabiya. I take this opportunity to express my deep gratitude to my parents, I reached this stage only because of their support and their dreams. I thank my in-laws, my husband Ahmad and my dearest son Afnan for their support. I thank all my family members for their unconditional love and support.

My special thanks to my chacha Riyazath Ali for his affection and support. I thank all the children in the family Dr. Sana, Dr. Sajid, Mujju, Basith, Sania, Asra, Ayesha, Neha, Ayaz, Ibrahim, Owais for always cherishing me.

Above all, I thank Almighty for blessing me with such a nice loving and supporting family and nice friends.

**Rafiya Sultana**



## ABBREVIATIONS

TB	Tuberculosis
<i>Mtb</i>	<i>Mycobacterium tuberculosis</i>
HIV	Human immunodeficiency virus
NTM	Non-tuberculosis <i>mycobacteria</i>
PDB	Protein databank
BCG	Bacille Calmette-Guérin
DOTs	Directly Observed Therapy
MDR	Multi-drug resistant strains
XDR	Extensively drug-resistant strains
RNTCP	Revised National Tuberculosis Control Programme
LAM	Lipoarabinomannan
mAGP	Mycolyl arabinogalactan–peptidoglycan complex
PIMs	Phosphatidylinositol mannosides
ORFs	Open Reading Frames
PGRSs	Polymorphic repetitive sequences
MPTRs	Major polymorphic tandem repeats
PE	Pro-Glu
PPE	Pro-Pro-Glu
3D	Three-dimensional
BLAST	Basic Local Alignment Search Tool
NCBI	National Center for Biotechnology Information
PSI-BLAST	Position Specific Iterative BLAST
PSSM	Position specific scoring matrix
SSEs	Secondary structure elements
PHYRE2	Protein Homology/AnalogY Recognition Engine
PROCHECK	A program that check the stereochemical quality and geometry of the protein structures
MAPSCI	Multiple Alignment of Protein Structures and Consensus Identification

CD-HIT	Cluster Database at High Identity with Tolerance
MEGA	Molecular Evolutionary Genetic Analysis
Bp	Base Pairs
MD	Molecular Dynamics
PCR	Polymerase Chain Reaction
IPTG	Isopropyl $\beta$ -D-thiogalactoside
MGLPs	Methylglucose Lipopolysaccharides
LPase	Laminaripentaose-producing beta-1,3-glucanase
GH-C	Glycoside hydrolase clan
CBD	Carbohydrate binding domain
NRPs	Non-ribosomally synthesised peptides
CULPs	Cutinase like proteins
pNPC2	<i>p</i> -Nitrophenyl acetate
pNPC4	<i>p</i> -Nitrophenyl butyrate
pNPC6	<i>p</i> -Nitrophenyl caproate
pNPC8	<i>p</i> -Nitrophenyl caprylate
pNPC10	<i>p</i> -Nitrophenyl caprate
pNPC12	<i>p</i> -Nitrophenyl laurate
pNPC14	<i>p</i> -Nitrophenyl myristate
pNPC16	<i>p</i> -Nitrophenyl palmitate
RMSD	Root Mean Square Deviation
SIE	Solvated Interaction Energies
$\Delta G$	Binding free energies
Vdw	Van der Waals
DS	Discovery Studio

# **Chapter 1**

## **Introduction to *Mycobacterium tuberculosis*, PE and PPE proteins, Computational Methods and General Molecular Biology Techniques**



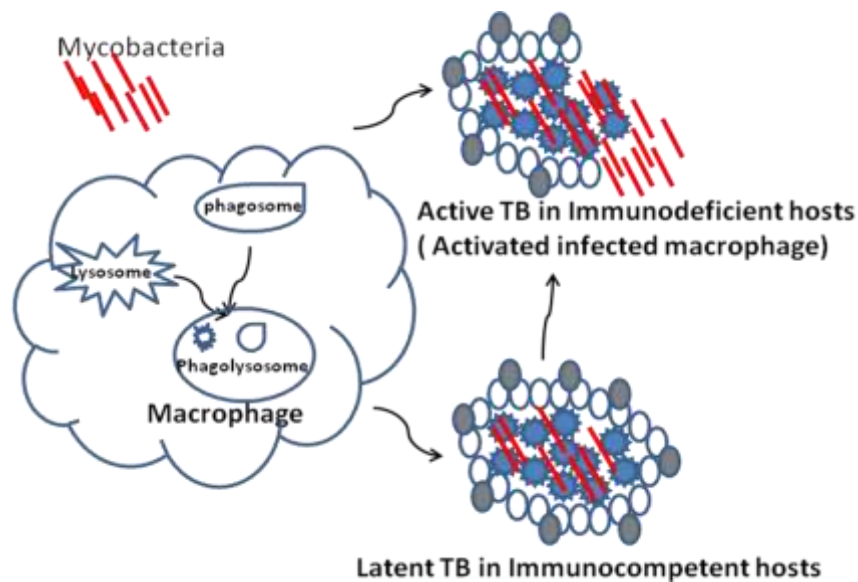
## 1.1 Tuberculosis

Tuberculosis (TB) is an airborne disease primarily caused by the bacterial pathogen *M. tuberculosis* (*Mtb*). TB remains a major global health problem and one of the main causes of mortality around the world. Two million deaths are caused worldwide by TB each year and it remains the second leading cause of death from an infectious disease, after the human immunodeficiency virus (HIV). Despite the availability of effective treatments, there is a need to develop new therapeutic agents to combat the spread of *Mtb* infection (Koul et al., 2004; WHO, 2014).

*Mtb* primarily infects humans in an aerosol droplet form by entering into the alveolar space in the lungs causing severe fever, cough and chest pain. The first contact of the bacterial pathogen is thought to be with resident macrophages/immune cells. Once within macrophages, *Mtb* has the ability to avoid the bactericidal mechanisms of phagocytic cells and therefore can replicate themselves. The aerosolized bacterium can bind and invade epithelial cells, to replicate and establish the bacterial infection thereby avoiding the potentially hostile environment of the macrophage (Bermudez and Goodman, 1996; Smith, 2003). When contacted with *Mtb*, humans build an effective response in the lungs primarily inhibiting the growth of *Mtb*, where the bacterium stays in a dormant form; this condition is often referred to as latent TB (Thillai et al., 2014). After latent infection, at the first opportunistic moment *Mtb* becomes active (Berry et al., 2013). Individuals with HIV/immunocompromised conditions are at a greater risk of developing active TB (Fogel, 2015) as shown in Figure 1.1.

The tests commonly used to detect and diagnose latent infection are 1. Tuberculin skin test (TST) and 2. Interferon-gamma release assays (IGRAs) (Thillai et al., 2014). Both these methods work by measuring the response of T cells to TB antigens (Cruz-Knight and Blake-Gumbs, 2013).

*Mycobacteria* can be classified into three groups. 1. Species that cause TB in humans or in animals, which includes *M. tuberculosis* complex: *M. tuberculosis*, *M. bovis*, *M. africanum*, *M. microti* and *M. canetti*. 2. Species that cause Hansen's disease or leprosy which includes *M. leprae* and *M. lepromatosis*, and 3. *Mycobacteria* that does not contain *Mtb* complex and that does not develop Hansen's disease or leprosy are known as non-tuberculosis *mycobacteria* (NTM) or MOTT (*Mycobacteria* Other Than TB).



**Figure 1.1.** Schematic representation of invading *Mtb* by initiating phagocytosis

NTM are environmental organisms, they opportunistically cause diseases in animals or humans (Wallace, 2010). NTM often present in soil and various water sources, which include other *mycobacteria* like *M. avium* complex is one of the most commonly observed pathogens in immunosuppressed patients, while *M. kansasii*, *M. malmoense* and *M. xenopi* are predominant in immunocompetent people. NTM can cause pulmonary disease resembling TB, lymphadenitis, skin disease or disseminated disease. Both *Mtb* and NTM can cause chronic lung infections but only TB spreads from person to person by simply inhalation of organisms, discharged/expectorated into the air. NTM infections are acquired directly from the environment (Kendall et al., 2011).

### 1.1.1 *M. tuberculosis*

In 1882, Robert Koch identified *Mtb* whose Latin-originated name describes the rod like shape bacillus (Keshavjee and Farmer, 2012) then known as the causative agent of TB and hence tubercle bacillus. He was awarded the Nobel Prize in Medicine and Physiology in 1905, for his work on TB, the bacterium is also known as Koch's bacillus. ([http://www.nobelprize.org/nobel\\_prizes/medicine/laureates/1905/](http://www.nobelprize.org/nobel_prizes/medicine/laureates/1905/))

The origin of *Mtb* is that bacteria in the genus *Mycobacterium*, like other actinomycetes, initially found in soil and some species evolved to live in mammals. Previously it was thought that the domestication of cattle that occurred 10,000- 25,000 years ago, allowed the passage of a mycobacterial pathogen from domestic animals to humans, during this adaptation the bacterium evolved to the closely related new host *Mtb*.

Explicitly, *M. bovis*, which causes a TB-like disease in cattle, was the evolutionary precursor of *Mtb* (Stead, 1997). A study of the distribution of deletions and insertions in the genomes of the *Mtb* complex proved the independent evolution of both *Mtb* and *M. bovis* from another precursor species, possibly related to *M. canettii*.

Tubercle bacilli, which are thought to most closely resemble the progenitor of *Mtb* are human and not animal pathogens (Brosch et al., 2002; Smith et al., 2009). However *Mtb* is an obligate pathogen and has no natural reservoir outside humans, where its main target cells are macrophages. Mostly, parasitization of macrophages by pathogenic *mycobacteria* involves the inhibition of numerous host-cell processes, which allows them to survive within the host cells. Pathogenic bacteria inhibit host processes by the fusion of phagosomes and lysosomes, providing antigen, apoptosis and the stimulation of bactericidal responses due to the activation of pathways involving protein kinases activated by mitogen, interferon- $\gamma$  and  $\text{Ca}^{2+}$  signalling (Koul et al., 2004). The success of *Mtb* during the parasitization of macrophages involves a modulation of the normal progression of the phagosome into an acidic and hydrolytically active phagolysosome, that also avoids the development of localized, productive immune responses against *Mtb* in the host (Meena and Rajni, 2010).

*Mtb* belongs to the genus *Mycobacterium*, that comprises filamentous Gram-positive bacteria that are distinguished by complex surface lipids. *Mtb* is typically visualized by Ziehl–Neelsen (acid-fast) staining and appears as a rod-shaped red bacillus and is surrounded by an impermeable and thick cell wall/capsule that is made of peptidoglycans, polysaccharides, glycolipids and lipids that primarily contains long-chain fatty acids, such as mycolic acid. Unlike other bacteria, *Mtb* does not form spores but has the capability to become dormant, a nonreplicating state characterized by low metabolic activity and phenotypic drug resistance (Gengenbacher and Kaufmann, 2012). Dormancy is a state where the bacillus remains inactive within infected tissue and reflects metabolic shutdown resulting from the action of a cell-mediated immune response that can contain but not eradicate the infection. As immunity wanes, the reactivation of dormant bacteria starts, resulting an outburst of disease even often many decades after the initial infection. The molecular basis of dormancy of the pathogenic bacteria and its reactivation remains difficult to be understood but is expected to be genetically programmed and to involve intracellular signaling pathways (Cole et al., 1998).

Indian scientists have contributed in many ways to *Mtb* like X-ray crystallographic studies, to name few such as NADP dependent DNA ligase, NrdH family

of redox proteins, transcription regulatory protein, *Mtb* ribosome recycling factor and etc (Phulera and Mande, 2013; Saikrishnan et al., 2005; Shrivastava and Ramachandran, 2007; Srivastava et al., 2005). Nearly 30 unique structures were determined by the Indian scientists that are now available at protein databank (PDB) (Arora et al., 2011). The characterization of PE/PPE protein families in *Mtb*, their role in the virulence of the bacterium was also explained by indian scientists (Akhter et al., 2012). For example PE 11 that belongs to pathogenic *mycobacteria* was overexpressed during bacterial infection, PPE Rv2430c encourages strong B-cell response (Choudhary et al., 2003; Singh et al., 2016).

### **1.1.2 BCG Vaccine and Drug Resistance**

TB remains a burden to the society since ancient times because of its infectious nature and several millions of annual deaths all over the world. Live attenuated Bacille Calmette-Guérin (BCG) isolated in 1908 by and named after Calmette and Guerin in Lille, France, is still the only vaccine available today which is being administered as immunization for over 90 years with astonishing safety records for the prevention of TB in humans. Tubercle bacilli were initially cultivated on glycerin and potato medium and later on led to the development of this vaccine from attenuated tubercle bacillus (Calmette, 1922).

However, there is a significant variation in the efficacy of immunization with BCG which remains controversial (Luca and Mihaescu, 2013). Upon infection, the TB therapy involves the combination of drugs to avoid resistance. Initially it started with the effective medications like streptomycin and *p*-aminosalicylic acid in 1944, then came the revelation of "triple therapy" streptomycin, *p*-aminosalicylic acid and isoniazid in 1952, which assured cure; then in 1970s it was found that isoniazid and rifampicin could reduce the duration of treatment from 18 to 9 months; in 1980's it was observed that adding pyrazinamide to these drugs cured the disease in only 6 months (Iseman, 2002).

Despite the availability of directly observed therapy (DOTS) and the BCG vaccine, the tubercle bacillus mostly remains naturally resistant to many antibiotics which makes the treatment difficult (Cole and Telenti, 1995; Snider and La Montagne, 1994). The emergence of multi-drug resistant strains (MDR), extensively drug-resistant strains (XDR) and an alarming rise in the number of TB patients co-infected with HIV, have stressed the urgency of developing new novel intervention strategies against TB (Andersen and Doherty, 2005). MDR-TB strains are resistant to anti-TB drugs such as



isoniazid and rifampicin, while XDR-TB strains are resistant to additional drugs including fluoroquinolone and one of the anti-TB injectable drugs such as kanamycin, capreomycin, or amikacin (Mani et al., 2014).

As per the reports presented in TB India 2014, there is a decrease in TB mortality rate by 42% in 2012 compared to 1990 reports. India initiated a Revised National Tuberculosis Control Programme (RNTCP) in 1997, to eradicate mortality and prevalence of TB due to drug resistance and HIV co-infected *Mycobacterium tuberculosis*. Due to the high population in India, there is a tuberculosis endemicity and further, the strain type varies from place to place. India is the second leading country with MDR-TB cases according to WHO 2010 reports where Andhra Pradesh is leading with highest cases reported in 2009. To combat MDR-TB and evaluate the extent of mismanagement of drugs, RNTCP developed new strategies. One of the reasons for drug resistance is either mismanagement of drug regime or the transmittance of primary drug resistant TB from person to person. According to RNTCP reports, in 2011 nearly 19,178 MDR-TB suspects were examined in India from 2007 to 2010 and 5,365 cases of MDR-TB were diagnosed, and initiated treatment for 3,610 MDR-TB patients. Due to many TB control programmes, nearly 2.3 million cases are reported in 2014 out of 8.6 million annual cases that occur globally, thus making India world's highest TB populated country (RNTCP Status Report 2011) <http://www.tbcindia.nic.in/>.

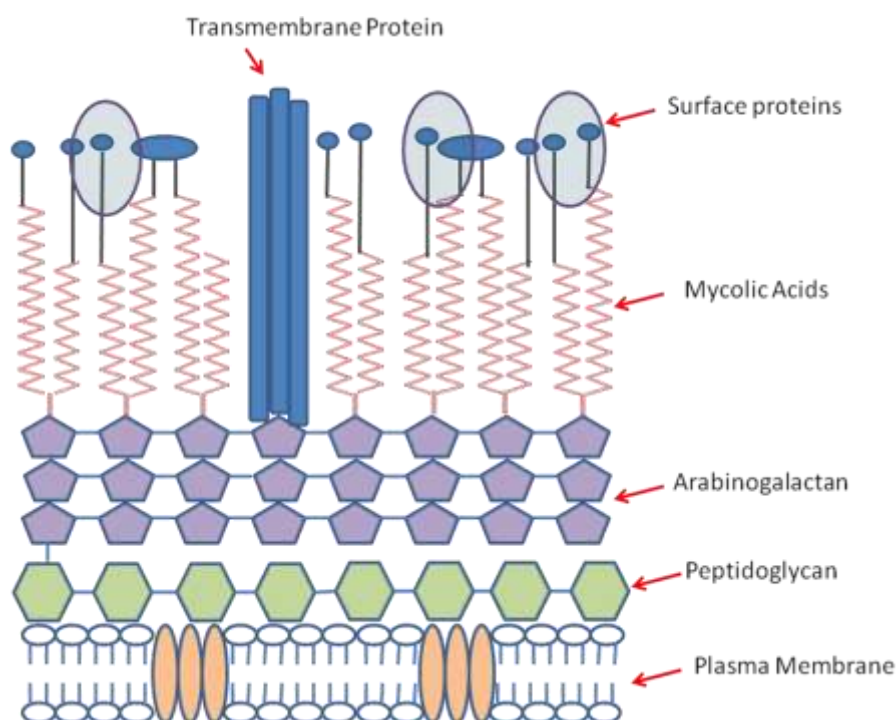
Resistance of the pathogenic *Mtb* is mainly due to the highly hydrophobic cell envelope acting as a permeability barrier (Brennan, 1994) that includes many potential resistance determinants like hydrolytic or drug-modifying enzymes such as  $\beta$ -lactamases and aminoglycoside acetyl transferases, and many potential drug-efflux systems that include 14 members of the major facilitator family and numerous ABC transporters. Understanding of these putative resistance mechanisms will help and promote better use of existing drugs and facilitate the conception of new therapies (Cole et al., 1998).

Recently the drug called Bedaquiline was approved by FDA that can attack *Mtb* in different ways from the available current treatment. Bedaquiline was discovered by the scientists of Janssen pharmaceuticals targetting mainly MDR-TB. Bedaquiline works by inhibiting the ATP synthase, an important enzyme in the ATP synthesis of *Mtb*, which leads to the death of bacteria (Mahajan, 2013).

### 1.1.3 Cell Wall Virulence Factors of *M. tuberculosis*

When *mycobacteria* are stained by gram staining, there is no decolorization by acid alcohol and therefore these bacteria are classified as acid-fast bacilli. This is due to high content of mycolic acids, long chain cross-linked fatty acids and other lipids in the bacterial cell wall (Daffe and Draper, 1998). Mycolic acid and other lipids are linked to underlying arabinogalactan and peptidoglycan (Ratledge, 1982). A variety of unique lipids, such as lipoarabinomannan (LAM), trehalose dimycolate and phthiocerol dimycerate that are linked noncovalently with the cell membrane appear to play an important role in the virulence of *Mtb* (Glickman and Jacobs, 2001). Most of the exported proteins and protective antigens of *Mtb* are a triad of related gene products called the antigen 85 complex, these have fibronectin binding capacity and thus play an important role in disease pathogenesis (Belisle et al., 1997). LAM is also a major constituent of mycobacterial cell wall and has been shown to induce tumor necrosis factor release from the macrophages (Chatterjee et al., 1992), which plays a prominent role in bacterial cell death. It was shown that LAM acts at several levels which can scavenge potentially cytotoxic oxygen free radicals, inhibiting protein kinase C activity and block the transcriptional activation of gamma interferon inducible genes in human macrophages such as cell lines, and hence it is a chemically defined virulence factor causing the persistence of *mycobacteria* inside mononuclear phagocytes (Chan et al., 1991).

The cell wall of *mycobacteria* is composed of two segments, upper and lower, as shown in Figure 1.2. Beyond the plasma membrane is peptidoglycan (PG) covalently bound to arabinogalactan (AG), which in turn is attached to the mycolic acids with their long meromycolate. This is called as the cell wall core, the mycolyl arabinogalactan-peptidoglycan (mAGP) complex. The upper segment of the cell wall comprises of free lipids, some with longer fatty acids complementing the shorter fatty acid chains, while some with shorter fatty acids complementing the longer chains. Interspersed somehow are the cell wall proteins, the phosphatidylinositol mannosides (PIMs), the phthiocerol containing lipids, lipomannan (LM) and LAM. When cell walls are disrupted/extracted with various solvents, the free lipids, proteins LAM and PIMs gets solubilized, while the complex mAGP remains as the insoluble residue. In simplistic terms, it can be considered that these lipids, proteins and lipoglycans are the signaling effector molecules in the disease process, while the insoluble part is essential for the viability of the cell. These components should be considered in the context of new drug development (Brennan, 2003).



**Figure 1.2.** The cell wall of *Mtb*. Sixty percent of the cell wall is composed of lipid; peptidoglycan, arabinogalactan and in particular mycolic acids which is proposed to cause virulence

Esterases or lipases are hydrolases that play significant role in lipid metabolism from prokaryotes to eukaryotes. Sequence analysis by Cole et al., 1998 revealed that there are at least 250 enzymes related to lipid metabolism which includes extracellular secreted enzymes, integrated cell wall enzymes and intracellular esterases/lipases (Camus et al., 2002; Cole et al., 1998). Most of the mycobacterial genes involved in lipid metabolism, cell division, chromosomal partitioning and secretion are required during infection in mouse model (Lamichhane et al., 2005; Sasseti and Rubin, 2003).

#### 1.1.4 Metabolic Pathways

The complete genome sequence indicates that the pathogenic *mycobacteria* can synthesize all the necessary amino acids, vitamins and enzyme co-factors, even if some of the pathways may differ from other bacteria. *Mtb* has the ability to metabolize a variety of hydrocarbons like carbohydrates, alcohols, ketones and carboxylic acids (Cole et al., 1998; Ratledge, 1982). For the past 20 years no new drugs were developed though there is an emergency for the development of new *Mtb* drugs with increasing resistance of the present leading anti-TB drugs. Therefore, Amir et al., 2014 identified reliable drug targets using computational approaches for *Mtb*. They have also shown that there are 5

unique pathways present in the *Mycobacterium* but not in the *Homo sapiens*, they are metabolism of C5-branched dibasic acid, methane metabolism, carbon fixation pathways, lipopolysaccharide biosynthesis, and peptidoglycan biosynthesis involving 60 new non-homologous targets identified using KEGG database of metabolic pathways.

### **1.1.5 The Genome of *M. tuberculosis***

The complete genome sequence of *Mtb* (H37Rv strain) consists of 4,411,529 base pairs (bp), with a G + C content of 65.6%. Approximately 4000 open reading frames (ORFs) were identified in the genome, accounting for 91% of the potential coding capacity. A few of these genes appear to have in-frame stop codons or frameshift mutations and may either use frameshifting during translation or correspond to pseudogenes. Several regions showing higher than average G + C content were detected; these correspond to sequences belonging to a large gene family that includes the polymorphic G + C rich sequences (Cole et al., 1998).

Interesting discovery of two extensive families of novel glycine-rich proteins, which may be of immunological significance as they were predicted to be abundant and potentially polymorphic antigens were named as the PE and PPE genes (Cole et al., 1998).

### **1.1.6 The PE and PPE Multigene Families**

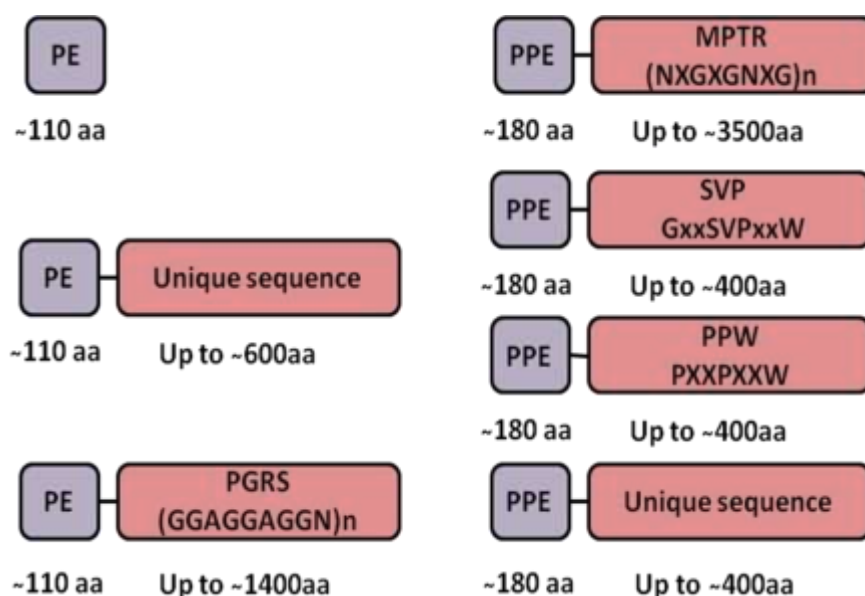
Two large unrelated families of acidic, glycine-rich proteins, the PE and PPE families represent about 10% of the coding capacity of *Mtb* genome, whose genes are clustered and are often based on multiple copies of the polymorphic repetitive sequences (PGRS), and major polymorphic tandem repeats (MPTR), respectively (Hermans et al., 1992; Poulet and Cole, 1995). The PE family with ~100 genes and the PPE family with ~70 genes are exemplified by the presence of Pro-Glu (PE) motif at positions 8 and 9 in most of PE proteins and Pro-Pro-Glu (PPE) motif at positions 8, 9 and 10 in most of the PPE proteins near the N-terminus. The PE and PPE protein families are characterized by highly conserved N-terminal domains with approximately 110 and 180 amino acid residues, respectively towards N-terminal domain, that is followed by a C-terminal segment that varies in size, sequence and repeat copy number (Figure 1.3) (Cole et al., 1998; Gordon et al., 2001).

### **1.1.7 Classification of PE Family**

PE family sequences are generally categorized into three subgroups based on C-terminal extensions, few members contain only the N-terminal domain, whereas most have C-terminal extensions ranging in size from 100 to 1,400 residues (Figure 1. 3) The first group of 29 members contains only the PE domain, the second group of eight members contains the PE domain followed by a unique sequence (i.e. these sequences do not show significant homology with other members of the PE family), and the largest group comprises 67 members which exhibit a PE domain followed by multiple repetitive tandem repeats of Gly-Gly-Ala or Gly-Gly-Asn, termed as PGRS domain and have a high glycine content (up to 50%) (Balaji et al., 2007; Cole et al., 1998). Considerable variation among PGRS sequences of *Mtb* clinical strains speculated that PGRS sequences could be a source of antigenic variation (Mukhopadhyay and Balaji, 2011; Talarico et al., 2007; Talarico et al., 2008).

### **1.1.8 Classification of the PPE Family**

The PPE protein family was also categorized into at least four subgroups based on C-terminal extensions (Figure 1.3), one of which constitutes the MPTR class characterized by the presence of multiple, tandem copies of the motif Asn-X-Gly-X-Gly-Asn-X-Gly. The second subgroup contains a characteristic, well-conserved GxxSVPxxW motif around position 350, whereas the third subgroup contains ~20 PPE proteins corresponding to the conserved C-terminal region comprising 44 amino acid residues with GFxGT and PxxPxxW sequence motifs and some of the last subgroup of PPE genes share remote similarity at the C-terminus (Adindla and Guruprasad, 2003; Cole et al., 1998) as shown in Figure 1.3.



**Figure 1.3.** Classification of the PE and PPE protein families in *Mtb*

### 1.1.9 Pathogenicity of the PE and PPE proteins

The PE and PPE family proteins are highly polymorphic, localize to the cell surface either secreted or expressed during the host infection (Kruh et al., 2010). These observations and results support the possibility of antigenic variation, mycobacterial virulence and pathogenesis (Banu et al., 2002). Although many PE and PPE proteins are recognized by the immune system during infection (Sayes et al., 2012), it remains unclear whether they are involved in antigenic variation (McEvoy et al., 2012). The findings that PE/PPE genes are duplicated and expanded in the genomes of pathogenic *mycobacteria* during the course of evolution lends further support to their role in virulence (Gey van Pittius et al., 2006).

The PE and PPE genes are not randomly distributed in the genome but often form operons, suggesting that PE and PPE proteins interact with each other. The crystal structure of the complex of PE and PPE domains showed directly that they form a heterodimer (Strong et al., 2006). The PE25-PPE41 protein complex has been shown to induce increased humoral and cell mediated immune response (Tundup et al., 2008). The crystal structure of a putative enzymatic domain of PE\_PGRS16 was reported in 2013 (Barathy and Suguna, 2013), and PE25/PPE41 dimer structure (in complex with EspG) only followed in the year 2014, underlining the inherently challenging nature of these proteins. Nevertheless, the PE and PPE domains have no detectable sequence resemblance or structural homology to other protein families, and their function remains unknown (Ekiert and Cox, 2014; Korotkova et al., 2014). With many potential roles of

the PPE and PE proteins as discussed above, it is important that further studies have to be performed to understand their activity. If extensive antigenic variability or reduced antigen presentation were indeed found, this would be significant for vaccine design and for understanding protective immunity in TB (Cole et al., 1998). The subcellular location of the PE and PPE proteins is largely unknown. PPE family protein Rv1808 is associated with the cell wall and exposed on the cell surface. Physical binding of Rv1808 to TLR2 manipulated the host cytokines via MAPK and NF- $\kappa$ B signaling pathways (Deng et al., 2014).

The PE and PPE family proteins like Rv1168c and Rv2430c exhibit strong immune response against the sera from clinically active TB patients and could distinguish TB patients from *M. bovis* BCG-vaccinated healthy controls, indicating that probably these proteins are highly expressed during active pathogenesis of *Mtb* (Choudhary et al., 2003; Khan et al., 2008; Nair, 2014).

Several PE and PPE proteins are upregulated during stress condition, only during macrophage infection and in host granulomas, supporting their role in pathogenesis of *mycobacteria* (Voskuil et al., 2004). Despite these efforts, several ORFs in PE and PPE gene clusters are largely unannotated with regard to their biochemical activity with very few exceptions such as, Rv3097c (LipY<sub>tub</sub>). The C-terminal domain of Rv3097c, a homologous protein belonging to hormone-sensitive lipase family is characterized by the conserved GX<sub>2</sub>SG active-site motif that hydrolyzes extracellular lipids and also it is the most highly expressed gene during as early as 24 hours post infection (Cascioferro et al., 2007; Daleke et al., 2011; Deb et al., 2006; Mishra et al., 2008; Srivastava et al., 2007). The PE protein Rv3097 has been shown to possess lipase enzymatic function. The PE13 and PPE18 genes, a part of the Rv0485 regulon encodes a putative transcription regulator. Mutation studies of Rv0485 indicates that it is *Mtb* virulent protein (Goldstone et al., 2009).

The PE\_PGRS region of protein Rv1818c (PE\_PGRS33) plays an important role in cellular immune response by influencing the antigen processing and presentation. Further analysis indicated that expression of complete PE\_PGRS33 protein in the non-pathogenic fast-growing *M. smegmatis*, offers better survival advantage in the infected macrophages (Delogu and Brennan, 2001; Dheenadhayalan et al., 2006). The PE domain of PE\_PGRS33 was found to be crucial for cell wall localization, while the PGRS domain affects the bacterial shape and colony morphology (Delogu et al., 2004; Mishra et al., 2008). Some PE and PPE members (Rv1818c, Rv1917c and Rv3873) were shown to be

associated with the cell wall (Delogu et al., 2004). A PPE family member, Rv3018c was shown to attenuate the growth in macrophages (Camacho et al., 1999). Some PPE proteins such as Rv1807, Rv3873 and Rv3872 play an important role in mycobacterial growth (Sasseti and Rubin, 2003).

Even after close to two decades of the availability of wealth of genome sequence of *Mtb*, TB persists as a major disease in the world and a major challenge to scientific world. One of the major outcomes of the genome sequencing of this pathogenic bacteria are PE and PPE multigene families which are the causative agents for the virulence of the bacteria. It is therefore, important to work on the identification and characterization of the PE and PPE proteins that are unique to *mycobacteria* virulence and these are extremely important for understanding the mechanism of pathogenesis in TB. These PE and PPE proteins are unique to *mycobacteria*, have evolutionarily expanded protein families preferentially in the virulent *mycobacteria* (Gey van Pittius et al., 2006) and are absent in the human host, making them ideal for diagnosis and drug targeting for the design of new anti-TB drugs.

Since the number of sequenced genomes are increasing rapidly, bioinformatics and computational methods play a major role in annotating and assigning the function of uncharacterized proteins, thus making biochemical characterization of unknown proteins easier with reasonable efficiency. Mostly sequence and three dimensional (3D) structure based methods are used in annotating protein functions. Computational methods have been developed to predict the catalytic residues important for the protein activity, thereby their local spatial arrangements can be used to categorize protein fold and function. In addition, these initiatives will add considerable value to the current volume of structural genomics data by reducing the number of absent or inaccurate functional annotations of the proteins. Thus, the challenge of assignment of function to proteins can be achieved by implementing new and reliable computational methods. Also, once the fold and function for a particular protein is available it may serve as a biological receptor to design new anti-TB potential drug candidates/inhibitors with high affinity and selectivity using various computer aided-drug design tools such as molecular docking methodologies (Speck-Planche et al., 2010). Molecular docking techniques can be used to screen the large databases available to select new and potent drug candidates/inhibitors, which further can be synthesized and biologically tested against pathogenic *mycobacteria*.



## **1.2 Computational Methods**

### **1.2.1 BLAST**

The Basic Local Alignment Search Tool (BLAST) is a program developed and maintained by the National Center for Biotechnology Information (NCBI). BLAST is widely used for comparing amino acid sequences or nucleotide sequences of different proteins or the nucleotides of DNA sequences for sequence similarity search. Using BLAST search researcher can compare a query sequence of interest with a database of sequences, and can identify homologous sequences that are similar to the sequence provided by the user above a certain threshold. BLAST uses a basic algorithm which is heuristic and rapidly finds local alignments with scores that are statistically considerable. The BLAST algorithm first looks for close matches to words within database sequences, then searches these matches within longer and high-scoring local alignments finally providing the results of the alignments ranked according to E-value. A low E-value is indicative of matching between the aligned sequences (Altschul et al., 1990).

### **1.2.2 PSI-BLAST**

Position Specific Iterative BLAST (PSI-BLAST) can find matches with sequences that were scored too low to be considered in a normal BLAST search. PSI-BLAST constructs a multiple alignment from BLAST output data, it processes this alignment into a position specific scoring matrix (PSSM), then it uses this matrix to search more matching words in the database that can be used to detect even more distant homologs. The aim of PSI-BLAST is to concentrate on the alignment of positions that are important, while allowing for more variability in areas that are not so important. PSI-BLAST uses iterative searches till no new sequences are identified or a user set limit threshold is reached (Altschul et al., 1997). The PSI-BLAST program can also be used to search for homologous 3D structures in PDB available at [www.rcsb.org/](http://www.rcsb.org/), in order to select appropriate templates for constructing 3D structure models of the template sequence using the comparative modeling methods. The BLAST methodology is available at <http://blast.ncbi.nlm.nih.gov/Blast.cgi>.

### **1.2.3 Sequence Alignment**

#### **1.2.3.1 CLUSTALW**

ClustalW (Thompson et al., 1994) is one of the most widely used program for multiple sequence alignment. The 'W' signifies a specific version that is improvised from original Clustal program.

Following steps are involved in CLUSTAL algorithm.

1. The method calculates all possible pairwise alignments and reports the score for each pair of alignment
2. Calculates a guide tree based on the pairwise distances using neighbor joining algorithm
3. Finds the two most closely related sequences
4. Aligns the sequences using progressive method in the following way
  - a. Calculates the consensus of the alignment
  - b. Replaces the two sequences with the consensus
  - c. Finds the two next-most closely related sequences
  - d. Iterate until all sequences have been aligned
5. Expands the consensus sequence alignment with the original sequence alignment
6. Finally reports the multiple sequence alignment

#### **1.2.3.2 CLUSTALX 2.1**

CLUSTALX is a graphical interface and a flexible advancement to the CLUSTALW multiple sequence alignment program which can be used for both nucleotide and amino acid sequences. CLUSTALX displays the multiple sequence alignment on the screen with an opportunity for versatile coloring theme highlighting the conserved residues. It also provides a pull-down menu bar which provides both customary multiple sequence and profile alignment, similar to the CLUSTALW. New features of this program includes: the facility to detect the suspected regions and realign selected residue sequence, user can cut-and-paste sequences to alter the order of the alignment, then select the sequence of the alignment realign and put it back into the original sequence alignment (Thompson et al., 1997). CLUSTALW is available at <http://www.ebi.ac.uk/Tools/clustalw2/>.

### **1.2.4 Signal Peptide Server**

The signal peptide is a short peptide sequence mostly present at the N-terminus initiating the secretory pathways and determines the efficiency of secretion in newly synthesized proteins. The SignalP is a program to predict the presence, absence and location of signal peptide/signal sequence cleavage sites in amino acid sequences of Gram-positive/Gram-negative bacteria. In this method cleavage sites/signal peptides are predicted based on Hidden Markov models and combination of some artificial neural networks (Emanuelsson et al., 2007). Signal peptide server is available at <http://www.cbs.dtu.dk/services/SignalP/>.

### **1.2.5 Protein Structure Databank**

PDB is a 3D structural database of large biological macromolecules, for example proteins and nucleic acids, and it was initially established at Brookhaven National Laboratories, USA in 1971. The structures are solved by biologists and biochemists from all over the world using X-ray crystallography, cryoelectron spectroscopy or NMR spectroscopy. This database may help researchers in understanding the drug target and disease thereby helping in designing new drug candidates using computational tools and thus creating a niche for molecular biologists. The available PDB file format can be viewed by using many freely available graphic software. PDB is accessible at [www.rcsb.org/](http://www.rcsb.org/) (Berman et al., 2000).

### **1.2.6 Modeling Methods**

#### **1.2.6.1 FUGUE**

FUGUE is a computational technique to find distant homology by sequence and 3D structure comparison in order to develop structural and functional relationship. FUGUE uses global-local algorithm method to align and compare sequence structure pair, when lengths vary significantly, or else uses the global algorithm. It utilizes structure-dependent gap penalties and environment-specific substitution tables to evaluate the matching, insertions and deletions using local environment. The gap penalty at every position of the structure is assessed using its solvent accessibility, its position with respect to the secondary structure elements (SSEs) and the preservation of the SSEs. Consequently, FUGUE generates alignments which represent a better relatedness between the amino acid sequences of the query protein and the template structure. FUGUE scans 3D structural database, evaluates the sequence-structure compatibility scores finally

providing potential homologues and alignments (Shi et al., 2001). It involves three steps, the first step constructs the environment-specific amino acid substitution tables using homologous structure alignments which are based on high-resolution 3D coordinates of the known proteins, the second step generates a database of structural profiles from structural alignments using the structure-dependent gap penalties and environment-specific amino acid substitution tables and the last step aligns the sequence alignment against each profile in the profile library. The statistical significance of each comparison, is evaluated to predict the possible evolutionary relationships and the sequence-structure compatibility. FUGUE is available at (<http://tardis.nibio.go.jp/fugue/prfsearch.html>).

### **1.2.6.2 Homology Models**

When experimentally determined protein structures are not available, comparative or homology modeling provides a useful method to build a 3D model for a protein target that is related to at least one of the known protein structures (Blundell et al., 1987; Browne et al., 1969; Marti-Renom et al., 2000; Sali and Blundell, 1993). From the 3D structural comparison of homologous proteins it is obvious that 3D structures are more conserved in evolution than their protein sequences because the mutual orientation of the secondary structures are similar, further amino acid replacement mostly occurs at the surface. The 3D framework of the homologous proteins can therefore be used for modeling a protein with unknown tertiary structure. First, one has to align the fragment of the known high homologous protein sequence with the unknown which provides the basis for model building. After the mainchain construction, the next step is to model the sidechain conformations. The simple method of modeling sidechain residues is to use the most probable conformation where no useful guidance is available from equivalent sidechain residues of the homologous proteins (Sutcliffe et al., 1987a; Sutcliffe et al., 1987b).

MODELLER is a program which uses homology or comparative modeling method to predict a reliable 3D model structure from its amino acid sequence by satisfaction of spatial restraints with respect to probability density functions. It is based on the alignment between the sequences to be modeled with the sequence of known experimentally determined template structure. The program automatically builds a model based on target-structural alignment for all non-hydrogen atoms in the protein structure by the satisfaction of the spatial restraints that includes loops which are not homologous to target sequence and minimizes final 3D model. The proteins built using this method on

low sequence homology can yield good quality and reliable models with great accuracy comparable to low resolution experimentally determined structures (Sali and Blundell, 1993). Therefore comparative homology modeling using rigid body assembly, matching segments and satisfying spatial restraints are the most reliable methods and widely used.

#### **1.2.6.3 PHYRE2**

Protein Homology/AnalogY Recognition Engine (PHYRE2) program can also be used for predicting the structure of an unknown protein. This method also follows comparative/homology modeling method for predicting protein structure by sequence-structure comparison because proteins are most probably related with their 3D structures than their protein sequences. PHYRE2 provides results in a way that user can see the template sequence matches in a color coded pattern based on confidence and coverage of the sequence length. The presence of secondary structural elements such as  $\alpha$ -helices,  $\beta$ -strands and disordered regions are also presented in a colored confidence bar. This method involves profile-profile or hidden Markov model - hidden Markov model alignment procedure for building 3D structure of a target protein sequence (Kelley et al., 2015). Using PHYRE2 user can also find ligand binding sites/pockets which can be further utilized by researchers in understanding the disease targets. PHYRE2 is available at <http://www.sbg.bio.ic.ac.uk/phyre2>.

### **1.2.7 Homology Model Validation**

#### **1.2.7.1 PROCHECK**

A program that checks the stereochemical excellence and geometry of the protein structures (PROCHECK). PROCHECK helps researches to evaluate the modeled 3D protein structures. PROCHECK calculates two torsion/dihedral angles for amino acids in the modeled protein (Ramachandran map) (Ramachandran et al., 1963). The first angle is called phi ( $\phi$ ) angle present around the bond between N-C $\alpha$  and the other angle is called Psi ( $\psi$ ) angle present around the bond C $\alpha$ -C along the peptide backbone. These torsion angles are restricted to certain values called allowed angles providing the needed flexibility to proteins such that the proteins can be moulded in a certain fold otherwise few torsion values may cause instability in protein fold due to rise in steric interactions between mainchain and the sidechain atoms of the amino acid residues, these angles are called disallowed angles. There is another possible flat torsion angle fixed to 180° called omega ( $\omega$ ). This planarity is developed due to the resonance by the partially double

bonded peptide bond that restricts the rotation around the C-N bond. Thus the arrangement of the amino acids in space of the protein 3D structures can be verified by scrutinizing the Ramachandran angles. In this way PROCHECK is used to evaluate the detailed stereochemistry of the protein structure, by checking residue-by-residue geometry (Laskowski RA, 1993). PROCHECK is available at <http://nihserver.mbi.ucla.edu/SAVES/>.

#### **1.2.7.2 Verify\_3D**

Verify\_3D is developed for evaluating the protein models constructed by various modeling tools. The basic concept in this program involves determining the compatibility of a built 3D protein model with its own amino acid primary chain by means of assigning a stereochemistry based on its local environment of each residue, described based on the following steps; 1. overall residue positions in model 2. the fraction of sidechain area of the residues buried in the fold that is covered by polar atoms like O, N and etc, and 3. the localization of SSEs such as  $\alpha$ -helix,  $\beta$ -sheet and loop; and finally comparing the results to good structures. For known 3D protein structures, the 3D profile score for the amino acid sequence of the model is high. The profile score of a model depends on its size and its validity. The 3D profile score of correct models enhance with molecular weight of the protein (Bowie et al., 1991). 3D profile score is a 3D-1D score that measures the compatibility of the amino acid sequence with its own 3D structure. This is most widely used in the homology modeling projects to validate the structures. 3D-1D score is plotted against each residue position to expose local regions or environment of the relatively high/low 3D-1D compatibility (Luthy et al., 1992).

#### **1.2.8 MAPSCI**

Multiple Alignment of Protein Structures and Consensus Identification (MAPSCI) follows heuristic algorithm. Users either upload the coordinates of the proteins or directly obtain them from PDB. MAPSCI provides the results in PIR formats which can be viewed using visualization software. MAPSCI is iterative method of computing the multiple structural alignments for the given set of proteins, it also generates consensus structure by representing every single protein as set of triplets which approximates the multiple structural alignment and minimizes the sum of pairwise distances between consensus and to the other proteins. MAPSCI is available at <http://www.geom-comp.umn.edu/mapsci/>. (Ilinkin et al., 2010).

### **1.2.9 Sequence Comparison**

#### **1.2.9.1 CD-HIT**

Cluster Database at High Identity with Tolerance (CD-HIT) is used to cluster large numbers of proteins or nucleotide datasets. CD-HIT program is faster than other tools in comparing sequences and database search tools by reducing manpower. This program is based on the incremental clustering algorithm. In this program sequences are first arranged in decreasing order of their sequence lengths. The longest sequence is taken as a representative for the first cluster. Now, each sequence is compared with the representatives of existing earlier chosen clusters. If the similarity in the sequence with any chosen representative occurs above a certain threshold, it is grouped and placed into that cluster. Otherwise, that sequence is taken as a representative for a new cluster. Short word filtering (counting words) is the algorithm applied by CD-HIT for the analysis of each sequence comparison to prove if the resemblance between two sequences is below the clustering limit. If this is not confirmed then an actual sequence alignment method is used for sequence comparison (Li and Godzik, 2006).

#### **1.2.9.2 MEGA**

Molecular Evolutionary Genetic Analysis (MEGA) is a computer program used to analyse DNA and protein sequence alignments and estimate evolutionary distances and inferring the phylogenetic trees. It involves three steps. First step is to provide the DNA or protein sequences of interest to the MEGA program. Second step is the alignment of sequences using ClustalW and MUSCLE (Edgar, 2004). Third step is the construction of a phylogenetic tree from the second step of aligned sequences (Kumar et al., 1994; Tamura et al., 2011).

#### **1.2.9.3 Phylogenetic Tree**

Phylogenetic tree also known as evolutionary tree is a method of representing the evolutionary relationships among a set of biological species or organisms called taxa. The tips of the tree joined together indicates descendent and the nodes of the tree indicates the common ancestors of those descendents. Phylogenetic tree is of three types

1. Rooted tree: It has a unique node representing the most common ancestor
2. Unrooted tree: No proper ancestor root

3. Bifurcating tree: It can be either rooted or unrooted bifurcating tree. Rooted bifurcating tree is a binary tree consisting of two descendents. Unrooted bifurcating tree is a unrooted binary tree

#### **1.2.9.4 Bootstrapping**

Bootstrapping is a computational tool used for assessing the phylogenetic trees introduced by Felsenstein (Felsenstein, 1985). Bootstrapping is based on Efron's bootstrap re-sampling technique, for example for the given sequences it generates new set of sequences by randomly choosing nucleotides. A tree is rebuilt with these set of sequences and verifies whether the same nodes are recovered i.e compared with the original tree. This is repeated several times until a 95% bootstrap value is obtained i.e. the same node is recovered after 95 to 100 iterations (Efron et al., 1996).

### **1.3 Chemical Libraries**

A chemical library is a database of chemical compounds which are generally used for high throughput screening in drug discovery. Database is a storehouse with compound information like purity, structure, quantity and etc. These databases of chemical compounds are useful for the researchers in drug discovery. The selected drug target is screened with a set of chemical compounds depending on the active site of the target. Chemical libraries are classified based on their structure such as lead like, natural product like, peptide like and etc. These databases are usually synthesized and maintained by chemists. The chemical space around these compounds is usually large but increases with the molecular weight of the compound. Initially, the screening of these chemical libraries will be carried out against a particular drug target and then the chemicals with the desired activity are chosen for the preliminary hits. After primary screening the hits are again tested for their activity. Once hit chemical is confirmed, commonalities among the different functional groups are studied for a particular chemical subspace and if needed chemical library is extended in that particular subspace by synthesizing more derivatives to the hit compound. Now, these library of compounds will be validated from hit to lead in the active molecule drug discovery development (Huggins et al., 2011).

Several chemical libraries are available. For example,

National Cancer Institute (NCI): <http://cactus.nci.nih.gov/>

MDL Inc: [http://www.mdl.com/products/experiment/available\\_chem\\_dir/index.jsp](http://www.mdl.com/products/experiment/available_chem_dir/index.jsp)

ZINC database: <http://blaster.docking.org/zinc/>



## 1.4 Virtual Screening and Molecular Docking

Virtual screening of databases is a common technique used in the early stage of drug discoveries by evaluating very large databases which reduces experimental efforts and time in making a drug by medicinal and computational chemists. Virtual screening is classified into two types, structure based screening i.e. target based docking, another is ligand based virtual screening i.e., using active library of compounds as templates. Structure based screening is called molecular docking. Docking is most widely used computational technique in the drug development process for structure based drug design to calculate protein - ligand interactions and also binding affinities (Schneider and Bohm, 2002).

Here ligand is a small molecule docked into a protein that is a target structure with all possible conformations and orientations where each docking position is considered as pose. Further to screen the energetically allowed/favorable pose, each pose is validated with a value known as dock score. These scores are based on the pose of the ligand in the target protein by taking into account various terms such as electrostatic interactions and van der waals interactions between the ligand and the target protein. This process is iterated for all the molecules available in the database used and ranked in the order of scores. A high score indicates the good binding of the ligand in the target protein or a best fit and is considered as active ligand/inhibitor. Another more specific way of evaluating the scores is by calculating the affinity score for the best fit ligands that can be obtained from the following equation, where the change in free energy term upon binding of a ligand molecule to the target protein can be described as a sum of individual contributions (Kroemer, 2007).

$$\Delta G_{\text{bind}} = \Delta G_{\text{int}} + \Delta G_{\text{solv}} + \Delta G_{\text{conf}} + \Delta G_{\text{motion}}$$

$G_{\text{int}}$  is a specific ligand–receptor interactions

$G_{\text{solv}}$  is the interactions of ligand and target protein with solvent

$G_{\text{conf}}$  is the conformational changes in the ligand and the target protein

$G_{\text{motion}}$  indicates motions in the target protein and the ligand during the complex formation

## **1.5 Molecular Dynamic Simulations**

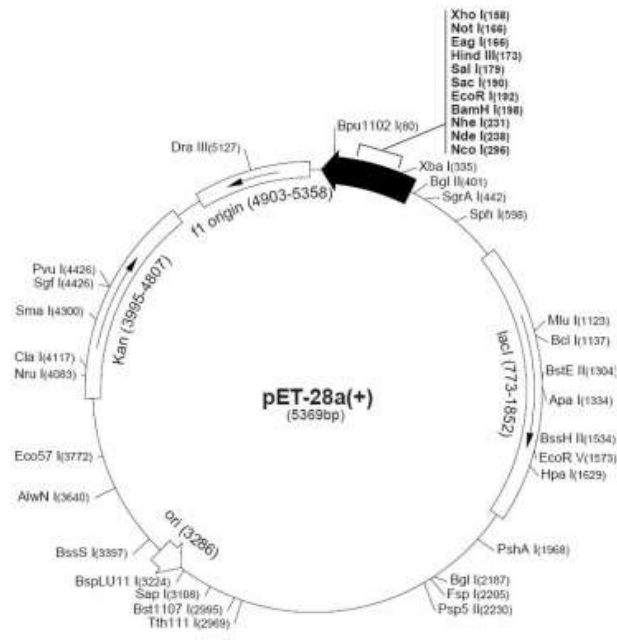
Molecular dynamic simulations (MD) is a computer based method applied to investigate the dynamic behaviour of the atoms and the molecules on the whole. The atoms and molecules are allowed from fast to slow internal motions. This method determines the interaction between the particles by solving Newton's equations of motions. This method is now widely applied to study the dynamic behaviour, thermodynamic stability, appropriate folding of biological molecules like proteins. Molecular mechanical force fields are applied to investigate the forces and potential energies of the interacting particles (Karplus and McCammon, 2002; Snow et al., 2005).

MD simulations can be applied to investigate the proper orientations of protein and ligand. Docking is a technique where the screening of large libraries of inhibitors/ligand/drug-like molecules is carried out that allows them to use the vast conformational space in a short time. Sometimes, docking does not allow the flexibility of the protein and is restricted upon ligand binding. However, MD simulations allows induced fit of the protein and ligand i.e. protein is flexible and the effect of solvent molecules can also be studied. However, MD simulations are computationally expensive and time-consuming therefore, the combination of the two techniques is used as a standard method to reduce time and fast screening of large libraries of inhibitors. Therefore molecular docking is done first, then MD simulations are carried out to search appropriate conformations of the protein receptor, refinement of the protein - ligand complex and calculation of accurate binding energies (Alonso et al., 2006).

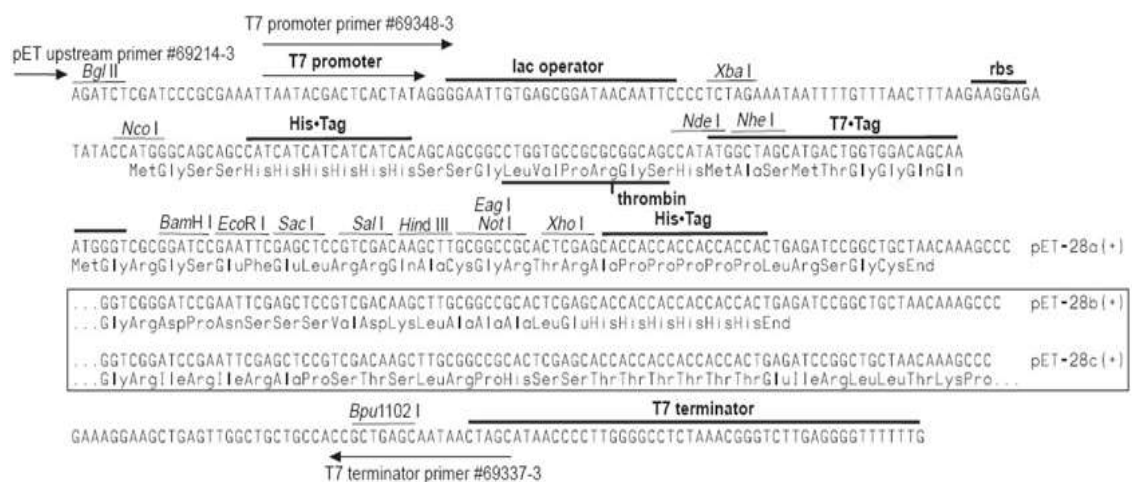
## **1.6 General Reagents and Methods for Biochemical Studies.**

### **1.6.1.1 pET-28a Vector for High-Level Expression of Recombinant Proteins in *E. coli***

The pET-28a vector is designed with N-terminal His tag/thrombin cleavage site/T7 tag configuration and an optional C-terminal His tag sequence, kanamycin resistance and restriction enzyme cloning. These sites are shown in the circular map below (Figure 1.4 and 1.5). T7 RNA polymerase allows transcription of the cloning/expression region of the coding strand. The fl origin in the pET-28a vector is framed such that the infection with helper phage produces single-stranded DNA that correspond to the coding strand. Hence, sequencing of the single stranded DNA (virions) must be performed with the T7 terminator primer.



**Figure 1.4.** pET-28a-c vector circular map (Source: Novagen, [http://www.helmholtz-muenchen.de/fileadmin/PEPF/pET\\_vectors/pET-28a-c\\_map.pdf](http://www.helmholtz-muenchen.de/fileadmin/PEPF/pET_vectors/pET-28a-c_map.pdf))



**Figure 1.5.** pET-28a-c vector cloning or expression region (Source: Novagen, [http://www.helmholtz-muenchen.de/fileadmin/PEPF/pET\\_vectors/pET-28a-c\\_map.pdf](http://www.helmholtz-muenchen.de/fileadmin/PEPF/pET_vectors/pET-28a-c_map.pdf))

### 1.6.1.2 Preparation of Media, Transformation buffer, Glycerol Stock , *E. coli* DH5a Competent Cells, *E. coli* BL21 Competent Cells, Gel loading dye and IPTG Luria Bertani (LB) Broth

LB is a widely used bacterial culture medium in the laboratories. The LB medium was prepared by dissolving 10 g of tryptone, 5 g of yeast extract and 10 g of NaCl in ~800 mL of distilled water. The contents were stirred and finally made up to 1 L with distilled

water. The pH of the medium was adjusted to 7.2 with 1 N NaOH and then sterilized by autoclaving at 15 lb/sq from 121-124 °C for 15 min. When required antibiotic was added after cooling the medium to 45 °C. Whenever necessary LB agar plates were prepared by adding 1.5% agar to LB broth, before pouring into the plates, sterilized LB broth containing agar.

### **Super Optimal Broth (SOB) Media**

1. Measured 900 mL of distilled H<sub>2</sub>O
2. Added 20 g Tryptone
3. Added 5 g Yeast Extract
4. Added 0.5 g NaCl
5. Added 10 mL of 250 mM KCl
6. Added 5 mL of 2 M MgCl<sub>2</sub>
7. Added 5 mL of 2 M MgSO<sub>4</sub>
8. Adjusted to 1 L with distilled H<sub>2</sub>O
9. Sterilized by autoclaving

### **Transformation buffer**

Prepared 0.5 M PIPES, adjusting the pH to 6.7 with KOH.

Mixed the following to get the required concentrations

Reagent	Amount/liter	Final concentration
MnCl <sub>2</sub> .4H <sub>2</sub> O	10.88 g	55 mM
CaCl <sub>2</sub> .2H <sub>2</sub> O	2.20 g	15 mM
CaCl <sub>2</sub> .2H <sub>2</sub> O	18.65 g	250 mM
PIPES (0.5 M, pH 6.7,)	20 mL	10 mM
H <sub>2</sub> O	1 L	

Then it is filtered using a Millipore filter and stored at -20 °C.

### **Glycerol Stock Preparation**

Added 0.5 mL bacterial culture in a sterile eppendorf tube and added 0.5 mL of sterile 80% (v/v) glycerol solution and stored at -20 °C.

### **DH5α Competent Cells**

DH5α supports blue/white screening, recA1 and endA1 mutations in DH5α™ increases insert stability and improves the quality of plasmid DNA prepared. DH5α cells have the following benefits-

1. High transformation efficiency and wide range of efficiencies from  $>1 \times 10^6$  to  $>1 \times 10^9$  transformants/μg
2. Enhances the plasmid yield and quality due to endA1 mutation
3. Allows blue/white screening of recombinant clones due to lacZΔM15. Also increases insert stability due to recA1 mutation

### **Competent Cell Preparation (*E. coli* DH5α)**

1. Sterilized glassware, media and buffers
2. Day 1: Streaked out frozen glycerol stock of bacterial cells onto LB plate (no antibiotics were added since these cells do not have a plasmid in them). Allowed to grow overnight at 37 °C
3. Day 2: Prepared primary inoculum by selecting a single colony of *E. coli* (DH5α cells) from fresh LB agar plates and inoculating a 10 mL culture of LB (no antibiotics). Allowed to grow culture at 37 °C in shaker overnight
4. DH5α cells were grown in LB broth of 100 ml at 37 °C with primary inoculum of 1% till the cell density reached to an absorbance 0.6 OD at 600 nm
5. Cultures were chilled on ice for 30 min. Harvested the cells by centrifugation at 4000 rpm for 15 min at 4 °C
6. Decanted the supernatant and the pellet was then resuspended in 40 mL of ice cold 0.1 M MgCl<sub>2</sub> and incubated on ice for 30 min
7. After incubation harvested the cells by centrifugation at 4000 rpm for 15 min at 4 °C, Decanted the supernatant
8. Resuspended the pellet in 10 mL of 0.1 M CaCl<sub>2</sub> and incubated on ice for 60 min. After incubation harvested the cells by centrifugation at 4000 rpm for 15 min at 4 °C. Decanted the supernatant
9. Cells were resuspended gently in 15% sterile glycerol and frozen 200 μl aliquots in liquid nitrogen and later stored at -80 °C

### ***E. coli* BL21 (DE3) CodonPlus-RIL Competent Cells**

BL21-CodonPlus competent cells are bacterial cells obtained from Stratagene for high levels of expression. These cells are designed such that they increase the expression levels of heterologous proteins in *E. coli* which otherwise are inadequate due to the shortage of certain tRNAs that are plentiful in the organisms containing heterologous proteins. Forced high-level expressions causes scarcity of tRNA molecules thereby disrupting the translation resulting in truncated protein expression or sometimes no protein expression. BL21-CodonPlus strains are designed such that they contain extra copies of tRNAs that mostly reduce the translation of heterologous proteins in *E. coli* and resolve codon bias problem. Therefore BL21-CodonPlus (DE3)-RIL cells are widely used strains for high-level protein expression. These cells provide easy induction in T7 promoter derived vectors for example pCAL and pET vectors. Further, BL21-CodonPlus (DE3)-RIL cells are designed such that they hold extra copies of the *argU*, *ileY*, *leuW* and *proL* tRNA genes. The role of these genes is to encode tRNAs that recognize the codons of some amino acids like arginine codons AGA and AGG, the leucine codon CUA, and the isoleucine codon AUA, respectively. The CodonPlus-RIL strains contain tRNAs which reduce the translation of heterologous proteins from containing AT-rich genomes.

### ***E. coli* BL21 Competent Cells Preparation**

1. Sterilized glassware, media and buffers.
2. Day 1: Streaked out frozen glycerol stock of bacterial cells onto LB plate (with the appropriate antibiotics added). Allowed to grow overnight at 37 °C
3. Day 2: Prepared primary inoculum by selecting a single colony of *E. coli* (BL21 cells) from fresh LB agar plates and inoculating a 10 mL culture of LB (with appropriate antibiotics). Allowed to grow culture at 37 °C in shaker overnight
4. DH5 $\alpha$  cells were grown in LB broth of 100 ml at 37 °C with primary inoculum of 1% till the cell density reached to an absorbance 0.6 OD at 600 nm
5. Harvested the cells by centrifugation at 4000 rpm for 15 min at 4 °C
6. Decanted the supernatant and the pellet was then resuspended in 40 mL of ice cold 0.1 M MgCl<sub>2</sub>
7. Harvested the cells by centrifugation at 4000 rpm for 15 min at 4 °C, decanted the supernatant

8. Resuspended the pellet in 10 mL of 0.1 M  $\text{CaCl}_2$  and incubated on ice for 45 min. After incubation harvested the cells by centrifugation at 4000 rpm for 15 min at 4 °C. Decanted the supernatant
9. Cells were resuspended gently in 15% sterile glycerol and frozen 200  $\mu\text{l}$  aliquots in liquid nitrogen and later stored at -80 °C

### **Gel loading dye (6x)**

1. 0.25% Bromophenol blue
2. 0.25% Xylene Cyanol blue
3. 30% Glycerol in water
4. Appropriate amount of DNA samples were mixed with 4  $\mu\text{l}$  of 6X loading buffer and loaded into the wells of submerged gel. Electrophoresis was carried out at 100 volts till the bromophenol blue reaches the end of the gel

### **Isopropyl $\beta$ -D-thiogalactoside (IPTG)**

1 M IPTG stock solution was prepared by dissolving 0.238 g of IPTG in 10 mL of sterile double distilled water. The stock solution was stored in 1 mL aliquots in eppendorf tubes at -20 °C. When required the stock solution was thawed and adequate amount was added to the medium to get 1 mM working concentrations of IPTG.

### **1.6.1.3 Polymerase Chain Reaction (PCR)**

Following steps are involved in PCR

1. Denaturation: DNA is heated at 95 °C for 20-40s. The double-stranded DNA melts and yields a single-stranded DNA
2. Annealing: This is done at medium temperatures ~ 54 °C (depending on GC content) for 20-40 s, the primers anneal with the single-stranded template DNA the polymerase enzyme attaches and starts copying the complementary part of the template. The annealing temperature is maintained 3-5 °C lower than the melting temperature ( $T_m$ ) of the primers
3. Extension/Elongation: This step is done at ~ 72 °C. In this step polymerase synthesizes new DNA strand that is complementary to the template DNA by adding dNTPs

4. Final elongation: This step is performed at a temperature of 70–74 °C for 5–15 min after the last PCR cycle to make sure that the remaining single-stranded DNA is fully extended

For every cycle, a single molecule of double-stranded template DNA is amplified into two separate molecules of double-stranded DNA. These two molecules are further used for amplification in the next cycle. These cycles are repeated such that more copies of DNA are generated, increasing the number of template copies exponentially.

#### **1.6.1.4 PCR Purification Protocol**

In this protocol (Qiagen) single or double stranded DNA fragments are separated from PCR. Fragments ranging from 100 bp to 10 kb are purified, using QIAquick spin columns in a micro centrifuge.

1. Added 5 volumes of Buffer PB to 1 volume of the PCR sample and mixed thoroughly
2. Placed the Qiaquick spin column in a 2 mL collection tube
3. Applied the sample to QIAquick spin column and centrifuged for 60 s
4. Discarded the flow-through and kept the QIAquick spin column back in the tube
5. Washed the column with 0.75 mL PE Buffer and centrifuged for 60 s
6. Discarded the flow-through and centrifuged for an additional 1 min at maximum speed
7. Kept the QIAquick column in a clean 1.5 mL micro centrifuge tube
8. Eluted the DNA by adding 10 µl of EB Buffer or water to the center of the membrane and waited for 1 min, then centrifuged for 60 s

#### **1.6.1.5 Agarose Gel Electrophoresis**

Tris Acetic acid EDTA (TAE) Buffer (stock 50 X)

<b>Reagent</b>	<b>Amount/liter</b>
Tris base	242 g
Glacial acetic acid	57.1 mL
0.5 M EDTA	100 mL
Distilled water	900 mL
Final volume	1 L



1. Measured 1 g of agarose. Agarose gels are commonly prepared in concentrations of 0.7% to 2% depending on the size of DNA needed to be separated. This is achieved by simply adjusting the amount of starting agarose (i.e. 2g/100mL will give you 2%)
2. Poured agarose powder into microwavable flask along with 100 mL of 1xTAE
3. Microwaved for 1-3 min (until the agarose is completely dissolved)
4. Now agarose solution is allowed to cool for 5 min, then added ethidium bromide (EtBr) approximately 0.2-0.5 µg/mL. EtBr binds to the DNA and allows visualization of the DNA under ultraviolet (UV) light
5. Poured the agarose into a gel tray with the well comb in place
6. Allowed newly poured gel to sit at room temperature for 20-30 minutes, until it has completely solidified

#### **1.6.1.6 Gel Extraction of DNA**

This protocol (Qiagen) is designed to extract and purify DNA fragment of 70 bp to 10 kb from standard agarose gels in TAE or TBE Buffer. Up to 400 mg agarose can be processed per a spin column.

1. Excised carefully the DNA fragment from the agarose gel with a clean, sharp scalpel. Minimized the size of the gel by removing extra agarose
2. Weighed the scalped gel slice in a colorless tube. Added 3 volumes of QG Buffer to 1 volume of gel (100 µl ~ 100 mg)
3. Incubated at 50 °C for 10 min until gel slice is completely dissolved. Mixed thoroughly by vortexing the tube to enable and dissolve the gel during incubation. Solubilised agarose completely when required increased the incubation time
4. After the gel slice is dissolved the color of the mixture turned to yellow. When colour of the mixture is orange or violet, added 10 µl of 3 M Sodium acetate pH 5.0 and mixed. The color turns yellow
5. Added 1 gel volume of isopropanol and mixed thoroughly. For 100 mg agarose slice added 100 µl of isopropyl alcohol
6. Now placed the QIAquick column in the 2 mL collection tube
7. Applied the sample to the QIAquick column which allows binding of DNA, centrifuged for 1 min
8. Discarded the flow-through and kept the column back in collection tube

9. Added 0.5 mL of QG Buffer to QIAquick column and centrifuged for 1 min
10. Washed the QIAquick column with 0.75 mL of PE Buffer and centrifuged for 1 min Discarded the flow-through and centrifuged for an additional 1 min at 13000 rpm
11. Placed the QIAquick column into a clean 1.5 mL micro centrifuge tube
12. Eluted the DNA by adding 40 µl of EB Buffer or water to the center of the QIAquick membrane. Allowed it to stand for 1 min and centrifuged at maximum speed for 1 min

#### **1.6.1.7 Transformation into *E. coli* DH5α cells**

1. The frozen competent DH5α cells were allowed to thaw by placing them on ice bath  
1-2 µl of plasmid having desired gene was added and incubated on ice for 30 min
2. After 30 min, the cells were subjected to heat shock at 42 °C for exactly 90 s and immediately chilled on ice for 5 min
3. 1 mL of LB broth was added and incubated at 37 °C for 60 min
4. The cells were collected by centrifugation at 4000 rpm for 10 min and again suspended in 200 µl of LB broth
5. Cells were plated on LB agar plates containing required concentration of antibiotic. The plates were then incubated at 37 °C to allow the 12 h for colonies to appear

#### **1.6.1.8 Plasmid Purification Protocol**

Followed method of Qiagen

1. Picked a single colony from a freshly streaked selective bacterial plate and inoculated a primary culture of 2–3 mL LB medium containing the appropriate antibiotic. Incubated for approximately 8 h at 37 °C with vigorous shaking (approximately 300 rpm)
2. Diluted the primary culture 1/500 to 1/1000 into 3 mL selective LB medium. Grown at 37 °C for 12–16 h with vigorous shaking (approximately 300 rpm)
3. Harvested the bacterial cells by centrifugation at 6000 x g for 15 min at 4 °C
4. Resuspended the bacterial pellet in 0.3 mL of Buffer P1. RNase A has been added to Buffer P1 before use

5. Added 0.3 mL of Buffer P2, mixed thoroughly by briskly inverting the sealed tube 4–6 times, and incubated at room temperature (15–25 °C) for 5 min
6. Added 0.3 mL of chilled Buffer P3, mixed immediately and thoroughly by vigorously inverting 4–6 times, and incubated on ice for 5 min. Precipitation is enhanced by using chilled Buffer P3 and incubation on ice. After adding P3 Buffer, a fluffy white material forms (like precipitate) and the lysate becomes less viscous. This fluffy white material contains genomic DNA, proteins, cell debris and etc. The lysate was mixed thoroughly. Centrifuged for 10 min at maximum speed of 10,000–13,000 rpm in a microcentrifuge. Removed supernatant containing plasmid DNA promptly
7. The supernatant from above step is added to the QIAGEN-tip 20 and allowed it to enter the resin by gravity flow
8. Washed the QIAGEN-tip 20 with 2 x 2 ml Buffer QC. Allowed Buffer QC to move through the QIAGEN-tip by gravity flow
9. DNA is eluted with 0.8 mL Buffer QF. Collected the eluate in a 1.5 mL or 2 mL microcentrifuge tubes
10. Precipitated DNA by adding 0.7 volumes of room temperature isopropyl alcohol to the eluted DNA. Mixed and centrifuged immediately at  $\geq 15,000 \times g$  rpm for 30 min in a microcentrifuge. Carefully decanted the supernatant
11. DNA pellet was washed with 1 mL of 70% ethyl alcohol and centrifuged at  $15,000 \times g$  for 10 min. Carefully decanted the supernatant without disturbing the pellet
12. Air-dried the pellet for 5–10 min and redissolved the DNA in a TE buffer, pH 8.0, or 10 mM Tris·HCl, pH 8.5)

#### **1.6.1.9 Expression of Protein by IPTG Induction**

1. On Day 1, a single bacterial colony of BL21 cells carrying plasmid was inoculated into 5 mL of LB medium containing kanamycin (35µg/mL) and was grown by incubating at 37 °C with vigorous shaking (200 rpm)
2. On day 2, 10 mL of LB media was inoculated (it is not necessary to include antibiotics for expression) with 2 % of the overnight culture
3. Culture was grown at 37 °C with vigorous shaking until the absorbance at 600 nm reaches to 0.4-0.6

4. 1 mL aliquot of cells prior to IPTG induction was removed, centrifuged in a micro centrifuge, the supernatant was separated carefully. The cell pellet was frozen at -20 °C. This was the time zero sample
5. IPTG was added to a final concentration of 1 mM and was left to grow for 4- 6 h
6. After incubation, expression was checked with coomassie stained protein gel, 1 mL sample was removed and centrifuged as described in Step 4. Cell pellet was frozen at -20 °C and induced samples were checked for expression. Checked expression in both the total cell extract soluble and insoluble fractions. If the target protein is insoluble, repeat expression at a lower temperature (15 to 30 °C)
7. Cell pellets were suspended in 50 µl of 4 X SDS-PAGE sample buffer
8. 30 µl of each of the pellet samples were loaded on 18% SDS-PAGE after boiling for 5 min and electrophoresis was carried out

### 1.6.2 SDS- PAGE Electrophoresis

12% Resolving gel,( pH 8.8) volume		4% stacking gel,( pH 6.8) volume	
1.5 M Tris-HCl buffer	1.25 mL	0.5 M Tris-HCl buffer	0.416 mL
30% acrylamide	3 mL	30% acrylamide	0.5 mL
10% SDS solution	33.3 µl	10% SDS solution	33.3 µl
10% APS solution	33.3 µl	10% APS solution	67 µl
TEMED	7 µl	TEMED	8 µl
H <sub>2</sub> O	0.684 mL	H <sub>2</sub> O	2.5 mL

#### 4X SDS sample buffer (8 mL)

2.5 mL 0.5 M Tris-HCl pH 6.8

0.8 g SDS

4 mL glycerol

20% β-mercaptoethanol or dithiothreitol (DTT)

0.08 mL bromophenol blue slurry

Above all contents were mixed and made up to the volume to 8 mL with deionized water

Stored in 1 mL aliquots at -70 °C.

### Coomassie Blue Staining Protocol

## Procedure

First step is to fix the gel in fixing solution for 1 h or overnight containing 50% methanol and 10% glacial acetic acid with gentle agitation.

Second step is staining the gel in staining solution containing 0.1% coomassie brilliant blue R-250, 50% methanol and 10% glacial acetic acid for 20 min with gentle agitation.

Last step is destaining the gel in destaining solution containing 40% methanol and 10% glacial acetic acid by changing the solution several times till the background of the gel is completely destained.

After this the gels can be stored in the storage solution containing 5% glacial acetic acid.

## Silver Stain Protocol

1. Fixed the gel in fixation solution containing 50% methanol, 12% acetic acid and 37% formaldehyde (100 mL) for 30 minutes or overnight
2. Washed gel twice with at least 30 minutes per wash (50% ethanol)
3. Incubated gels in sodium thiosulfate (10 g in 50 mL of double distilled water) for 60 s exactly
4. Washed gels with double distilled water twice
5. Added silver nitrate solution (0.1 g silver nitrate and 30  $\mu$ l formaldehyde in 50 mL double distilled water) and allowed the gels to stain for 30 min in dark
6. Washed gels with double distilled water twice
7. Added developer solution (3 g sodium carbonate 30  $\mu$ l formaldehyde in 50 mL double distilled water). Kept it for 10 min
8. The reaction was stopped by adding destain (12% acetic acid)
9. Washed gels with double distilled water twice

## 1.6.3 Western Blotting

### 1x Towbin buffer

Reagent	Amount/Liter
Tris HCl	3 g
Glycine	14.4 g

Dissolved the above in double distilled water, added 400 ml of methanol, adjusted volume to 1 L with double distilled water.

10x TBS buffer - 100 mL

Reagent	Amount/Liter
Tris HCl	1.21 g
NaCl	8.77 g

Adjusted pH with HCl (~ 3 mL) to 7.4; made up the volume to 100 mL.

**1xTBST:** 1x TBS prepared from 10x Stock, add 0.05% of tween 20.

Blocking solution-1x TBS with 5% casein.

Ponceau Stain-100 g of Ponceau salt in 5% acetic acid.

Primary antibody (Mouse anti-His monoclonal antibody)-1:2000 dilution

Secondary antibody (peroxidase-conjugated Goat anti Mouse IgG)-1:2000 dilutions

Developer- 250 mL

Packet A 4 g

Packet B 22 g

Dissolve first A then add B

Fixer- 250 mL

60.75 g dissolved in water.

### Protocol for Western Blotting

1. Soaked the SDS-PAGE gel and methanol pretreated PVDF membrane in 1x Towbin buffer for 30 minutes
2. Placed the PVDF membrane in the sandwich chamber with 2 fiber pads and 2 filter papers all soaked in transfer buffer with facing the black side of the sandwich down
3. The sandwich is black side, fiber pad, filter paper, SDS gel, membrane depending on the size of the protein, filter paper, fiber pad and red side. After each layer roll out to remove the air bubbles
4. Run the gel at 4 °C (in the cold room) at 40 volts, overnight
5. After the trans-blot, air dried the blot and stained with the Ponceau reagent and after visualization of the protein destained the blot by 1xTBST buffer

6. Blocked the nonspecific sites with blocking solution
7. Kept in primary antibody on a rocker at 4 °C (in the cold room), overnight
8. Removed the primary antibody and washed the blot 3 times with 1xTBST buffer, each for 5 minutes
9. Kept the blot in the secondary antibody at room temperature for 1 h on a rocker
10. Removed the secondary antibody and washed 3 times with 1xTBST buffer, each for 5 min
11. Chemiluminescence reagent was prepared by mixing enhanced luminol reagent from brown bottle and oxidizing reagent from white bottle immediately before use
12. Incubated the chemiluminescence reagent with the membrane for 30 seconds
13. The blot was developed with Supersignal west pico chemiluminiscent substrate (GeneX) and visualized on a versa doc V5 (Bio-Rad)

#### **1.6.4 Purification of Recombinant Protein by Affinity Chromatography**

##### **Denaturing Buffers for Talon™ column**

Equilibration buffer/ Wash buffer (pH 7.0)

50 mM sodium phosphate

300 mM sodium chloride

8 M urea

Imidazole elution buffer (pH 7.0)

50 mM sodium phosphate

300 mM sodium chloride

150 mM Imidazole

##### **Column regeneration buffer (pH 5.0)**

20 mM MES

0.1 M NaCl

##### **Column Preparation**

1. Thoroughly resuspended the TALON Resin
2. Transferred the required amount of resin suspension to a column, for 100 µl bed volume of resin 200 µl resin was added
3. Centrifuged for 2 min at 2000 rpm. Supernatant was separated carefully
4. 10 bed volumes of equilibration buffer was added and mixed briefly to pre-equilibrate the resin

5. Centrifuged for 2 min at 2000 rpm. Supernatant was taken out carefully

### **Sample Preparation**

1. Cell culture was harvested (cells expressing the desired protein checked by SDS-PAGE) by centrifugation at 1,000–3,000 x g for 15 min. Supernatant was removed
2. Cell pellet was resuspended by vortexing in 2 mL of chilled equilibration buffer per 25 ml of culture  $\leq 100$  mL For cultures  $> 1$ L, resuspended the pellet in 1–2% of the original culture volume
3. This cell suspension was incubated for 20–30 min. Freeze thawed in order to lyse the cells
4. Centrifuged for 15 min at 12000 rpm. Supernatant was taken out carefully and loaded on the pre-equilibrated column
5. Column was gently agitated for 30 min on a platform shaker to allow the polyhistidine-tagged protein to bind the resin
6. Centrifuged for 2 min at 2000 rpm. Supernatant was taken out carefully that is the flow through (FT)
7. Column was washed thrice by adding 10 bed volumes of equilibration/wash buffer first with 10 mM imidazole, later 20 mM imidazole and finally with 30 mM imidazole respectively to remove non- specific binding of protein
8. Suspension was gently agitated for 10 min on a platform shaker to promote thorough washing
9. All the washes were collected and kept at 4 °C for SDS-PAGE
10. Polyhistidine-tagged protein bound to the resin was eluted thrice by 200  $\mu$ l of elution buffer containing 150 mM imidazole (E)

All the samples were run on SDS-PAGE to determine the extent of purification

### **1.6.5 Gel filtration**

Gel filtration is a technique used for purification of proteins/biomolecules. It is also called as size exclusion chromatography. This technique is based on the differences in the size of the biomolecules to be separated as they pass through the column packed with the gel filtration material. Gel filtration is also used for the refolding of the denatured/unfolded proteins. In gel filtration column, resin is made up of porous material/matrix with physical stability and inertness. Here resin is stationary phase and the buffer is a mobile phase. Sephadex G-10, G-25 and G-50 are resins used for this purpose. Sephadex is



cross-linked dextran gel available in various fractionation range that vary in the degree of the polymerization. During gel filtration, larger proteins travel faster than the smaller proteins and are eluted first.



## **Chapter 2**

### **Prediction of Certain Well-Characterized Domains of Known Functions Within the PE and PPE Proteins of *Mycobacteria***



## 2.1 Introduction

As discussed in Chapter 1, TB caused by *Mtb* remains a major global health problem and one of the main causes of death around the world (Zaman, 2010). Although DOTS and BCG vaccine are available, the tubercle bacillus remains naturally resistant to many antibiotics, making the treatment complicated and difficult (Cole and Telenti, 1995; Snider and La Montagne, 1994). One of the major issues associated with TB patients is its co-infection with HIV, significant variations in the efficacy of immunization with BCG and the drug resistance resulting in the resurgence of MDR and XDR-TB (WHO, 2006). Thus stressing the need for the development of novel strategies against TB (Andersen and Doherty, 2005). The drug resistance is partially due to the highly potential hydrophobic cell envelope (permeability barrier) (Brennan, 1994), drug-modifying enzymes and potential drug-efflux systems (Cole et al., 1998). Hence efforts should be made in developing new anti-TB therapeutic agents that could be administered for shorter duration and most effective in completely killing the pathogen present in the human host (Koul et al., 2004; WHO, 2014) and to control the infection of highly drug resistant strains (Udwadia et al., 2012).

The complete nucleotide sequence of *Mtb* H37Rv strain comprising ~4,000 genes, contains two new gene families; PE and PPE, accounting for ~10% of the total genome (Cole et al., 1998). These proteins are characterized by highly conserved N-terminal domains with approximately 110 and 180 amino acid residues, respectively. The names PE and PPE for these proteins are due to the presence of amino acid sequence motifs, Pro-Glu and Pro-Pro-Glu, respectively towards the N-terminus (Akhter et al., 2012; Cole et al., 1998; Sampson, 2011). Mostly these genes are arranged in such a way that the PE genes are always followed by the PPE genes and spread all over the genome (Tundup et al., 2006).

The PE and PPE genes are known to be present in pathogenic and non-pathogenic *mycobacteria* but not yet in non-mycobacterial species (Abdallah et al., 2009; Ishikawa et al., 2004; Nair, 2014). PE and PPE genes have a strong evolutionary selection in the pathogenic *mycobacteria* since their expansion is linked to the ESAT-6 gene clusters that has role in immunopathogenesis (Gey van Pittius et al., 2006). The functional role of few PE and PPE proteins are identified so far. The PE and PPE proteins are highly polymorphic and localized to the cell wall/surface and possess immunological role.

These proteins are anticipated to be a source of antigenic variation and liable for virulence of pathogen. Subsequently, the complete genome sequencing of several strains

of *Mycobacterium* (Bentley et al., 2012; Garnier et al., 2003; Kim et al., 2012; Li et al., 2005a; Stinear et al., 2008; Vissa and Brennan, 2001) identified the existence of variable numbers of PE and PPE genes. Single nucleotide polymorphisms were observed to be greater in these genes compared to the non-PE and non-PPE genes. These genes were proposed as possible vaccine candidates (Akhter et al., 2012; Cole et al., 1998; Sampson, 2011).

Comparative analysis of the PE and PPE families in *Mtb* H37Rv (virulent) and H37Ra (avirulent strain) revealed genetic differences in several single nucleotide variations, insertions and deletions (Kohli et al., 2012). Comparative genomics of the *M. avium* complex members revealed several polymorphisms in the PE and PPE family members and the presence of some unique members in PPE family that have been implicated for applications in diagnostics (Mackenzie et al., 2009).

Among the revealed functional characterization of the PE and PPE proteins, the N-terminal PE domain of PE\_PGRS33 (Rv1809) is crucial for protein localization to the cell wall in *M. marinum* and *M. tuberculosis* (Cascioferro et al., 2011; Zumbo et al., 2013). Antigenic properties of some of the proteins like operonic PE25 (Rv2431), PPE41 (Rv2430) and the complex PE25/PPE41 showed that PPE41 and PE25/PPE41 complex stimulate significant B cell response compared to that of PE25 protein (Tundup et al., 2006; Tundup et al., 2008). The upregulation of PPE32 i.e. Rv1808 in several conditions describe its role in the host innate immune reaction (Cascioferro et al., 2007; Deng et al., 2014). In murine macrophages cell lines the gene PE\_PGRS63 (Rv3097c) is highly expressed as early as overnight post-infection (Srivastava et al., 2007) and higher expression of some of the PE/PPE genes Rv0977, Rv1361c and Rv1840c were observed in human macrophages post-infection (Dubnau et al., 2002).

Phylogeny to functional prediction of PE and PPE family members in *Mtb* and their pathogenicity have been recently reported (Fishbein et al., 2015). Further functional studies of the PE and PPE family members in *Mtb* have reported their localization in cell wall, cytosol and membrane, and their functions have been implicated in cell wall, virulence, detoxification, adaptation, insertion sequences, lipid metabolism, intermediary metabolism, respiration and cell processes (Fishbein et al., 2015). In view of the fact that the PE and PPE family members are distinctive to *Mycobacterium*, and absent in the human host and other related bacteria, these proteins are appropriate targets for drug design to combat and detain the pathogenicity of *Mtb*.

Despite the availability of the complete *Mtb* genome sequence very few PE, PPE proteins were structurally and functionally characterized. Since wet-lab experimental validation is often time consuming and tedious, computational methods can be used for predicting the structure and activity of proteins that can be further experimentally validated. Most often, comparative sequence analysis is implemented in functional and structural annotation of a proteome, although there exists few limitations because these methods mainly rely on the comparative sequence homology (Anand et al., 2011). The protein fold and conservation of active site residues are determinant factors of molecular function. Therefore, whenever experimental structures are unavailable, computer-based protein modeling methods may be employed to predict the structure and possible function (Pearson, 2013). Often, large proteins evolve as domains to facilitate independent folding and function, without considering their location in the protein sequence (Swathi Adindla, 2013). Recently re-annotation of several unannotated proteins of *Mtb* using computational methods by Ramakrishnan et al added substantial information on probable structures and functions of unknown proteins. In this case, proteins are annotated by means of assigned structural and functional domain families. They have also drawn structural and functional inferences for DUFs (Domains of Unknown Function) that comprising 89% of *Mtb* proteins that belong to functionally uncharacterized protein domain families (Ramakrishnan et al., 2015).

In this chapter, we have analyzed the sequences of PE and PPE proteins from several completely sequenced mycobacterial species. Our results provide clues into the evolutionary relationship between the PE and PPE proteins and identified well-characterized domains present in these mycobacterial proteins.

## **2.2 MATERIALS AND METHODS**

### **2.2.1 Sequence Searches – PSI-BLAST**

The amino acid protein sequences corresponding to the PE and PPE family from *Mtb* H37Rv strain were retrieved from the NCBI databank (<http://www.ncbi.nlm.nih.gov/>). The program PSI-BLAST is used for iterative and reciprocal search to identify regions corresponding to the PE and PPE sequence regions ([www.ncbi.nlm.nih.gov/BLAST/](http://www.ncbi.nlm.nih.gov/BLAST/)) against the protein sequences belonging to 60 mycobacterial species. The PSI-BLAST program looks for related proteins by constructing a profile or a PSSM that is generated from multiple sequence alignment of proteins only if detected above a given threshold score (Altschul et al., 1997). The results obtained were manually inspected to confirm the protein family and to exclude protein sequences of unrelated families.

### **2.2.2 Selection of Non-Redundant Proteins - CD-HIT**

The large dataset of mycobacterial protein sequences obtained from the PSI-BLAST/homology searches had a high percentage of redundancy. We have used the CD-HIT program (Li and Godzik, 2006) that has an ability to efficiently handle huge datasets of protein sequences i.e. millions of protein sequences, in order to exclude redundant proteins and short-list a dataset of protein sequences based on 40% sequence similarity cut-off value ([http://weizhong-lab.ucsd.edu/cdhit\\_suite/cgi-bin/index.cgi](http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi)).

### **2.2.3 Multiple Sequence Alignment - ClustalX**

The PE and PPE protein sequences of mycobacterial species retrieved earlier were aligned separately using ClustalX 2.1. It generated a dendrogram by performing a heuristic pairwise progressive sequence alignment. This dendrogram was used to build the multiple sequence alignment (Jeanmougin et al., 1998). The parameters used for multiple sequence alignment of the PE and PPE proteins were; “10” for gap opening penalty, “0.2” for gap extension and "Gonnet Series" was selected for protein weight matrix.

### **2.2.4 Phylogeny Analysis – MEGA5**

Phylogenetic trees were constructed separately for PE and PPE proteins with draw tree clustering algorithm in ClustalX based on the neighbor joining option in clustering algorithm. The phylogenetic tree thus obtained was explored using MEGA 5.0 (Tamura et



al., 2011). The representative protein sequences based on the phylogenetic trees generated were selected for further study.

### **2.2.5 Protein Fold Recognition - Phyre2**

For functional domain analysis, the non-PE and non-PPE sequence regions in the PE and PPE proteins were selected. PHYRE2 (<http://www.sbg.bio.ic.ac.uk/phyre2>) is a computational aid for predicting protein structure, function and mutations (Kelley et al., 2015). PHYRE2 is basically useful to identify the protein folds of distantly related sequences and adopts advanced methods for building the 3D model. The validation of the model constructed from the protein sequence of interest can be judged based on confidence score obtained, percentage of sequence identity and the sequence alignment between template and the target structure, secondary and tertiary structures of the model constructed, domain composition, conservation of the residues which are functionally important i.e. active site and finally the model quality, such as stereochemistry.

## 2.3 RESULTS AND DISCUSSION

Mycobacterial species are well known to abundantly comprise variable numbers of the PE and PPE family proteins. Several variations in these protein families are due to synonymous and non-synonymous single nucleotide polymorphism (SNPs), in-frame insertions and deletions that occur in protein coding regions resulting in their altered encoded amino acid sequence that in turn affects physicochemical properties (Kohli et al., 2012; Mackenzie et al., 2009; McEvoy et al., 2012). These features indicate the observed dissimilarity between the mycobacterial species.

The N-terminal domain of PE and PPE family proteins, and the Gly-rich sequence regions in these family of proteins give limited clues on their probable structure and function. The structure-based characterization of the protein folds with certain functional annotation are more useful, which was therefore applied in the current work using PHYRE2 tools for the non PE and non PPE regions of the representative members of these protein families. Based on the existence of functional protein domains these investigations shed light on the evolution of pathogenic mycobacterial species. The list of mycobacterial species for the assessment of structural folds and function of the PE and PPE protein families analyzed in this work are shown in the Table 2.1.

**Table 2.1.** List of mycobacterial species studied for the PE and PPE protein analyses.

<b>Mycobacterium Species</b>	<b>Nature of infection</b>
<i>M. abscessus</i> subsp. <i>bolletii</i> str. GO 06	Opportunistic infections
<i>M. africanum</i> GM041182	Human tuberculosis in West Africa
<i>M. avium</i> subsp. <i>paratuberculosis</i> K-10	The causative agent of Johne's disease in cattle and other ruminants
<i>M. bovis</i> BCG str. Pasteur 1173P2	Vaccine
<i>M. canettii</i> CIPT 140010059	Pulmonary tuberculosis
<i>M. chubuense</i> NBB4	Non-pathogenic strain , versatile hydrocarbon degrader
<i>M. gilvum</i> PYR-GCK	Non-pathogenic strain, Also useful in the application of bioremediation processes since they can degrade high-molecular-weight polycyclic aromatic hydrocarbons (PAHs)
<i>M. gilvum</i> Spyr1	None
<i>M. indicuspranii</i> MTCC9506 M	Immunotherapeutic against leprosy and also approved as a vaccine against it
<i>M. intracellulare</i> MOTT-64	Cause tuberculosis in birds, and pulmonary and disseminated infections in immunocompromized humans
<i>M. leprae</i> Br4923	An unculturable obligate pathogen that causes leprosy in humans

<i>M. liflandii</i> 128FXT	Frog pathogen
<i>M. marinum</i> M	Pathogen of fish and amphibia, is a near relative of <i>MTB</i> , the etiologic agent of tuberculosis in humans
<i>M. neoaurum</i> VKM Ac-1815D	synthesizes the valuable steroid precursor 4-androstene-3,17-dione as a major product from sitosterol
<i>M. rhodesiae</i> NBB3	Rare cases of peritonitis, also involved in PAH degradation pathway
<i>M. smegmatis</i> str. MC2 155	Nonpathogenic species
<i>M. ulcerans</i> Agy99	Causes Buruli ulcer in humans
<i>M. vanbaalenii</i> PYR-1	Non-pathogenic strain, useful for bioremediation
<i>M. yongonense</i> 05-1390	Pulmonary disease
<i>M. arupense</i>	Tenosynovitis and osteoarticular infections
<i>M. thermoresistibile</i> ATCC 19527	Human Pathogen
<i>M. aromaticivorans</i>	Powerful degrading capacity to polycyclic aromatic compounds (PAHs)
<i>M. asiaticum</i>	<i>M. asiaticum</i> pneumonia, human infection
<i>M. bohemicum</i>	Human, Veterinary Down's syndrome and tuberculosis
<i>M. caprae</i>	Tuberculosis among animals and, to a limited extent, in humans
<i>M. colombiense</i>	Respiratory disease and disseminated infection in immunocompromised HIV patients, as well as lymphadenopathy in immunocompetent children
<i>M. conceptionense</i>	opportunistic pathogen. Infections include skin and soft tissue infection characterized by slowly progressive granulomatous inflammation, lymphadenitis, skeletal and pulmonary infections, and catheter-related, disseminated infection in immunocompromised patients
<i>M. cosmeticum</i>	Cosmetic infection and from a nail salon
<i>M. europaeum</i>	Higher risk of Lung infection in immunocompromised hosts/ patients
<i>M. fortuitum</i>	Occasionally cause miscellaneous human infections, including skin and soft tissue infections, post-surgical wound infections, lymphadenitis, and catheter-related infections, but it can also cause lung disease.
<i>M. genavense</i>	Disseminated infection, Opportunistic pathogen, gastrointestinal disorders.
<i>M. hassiacum</i>	Non pathogenic
<i>M. heraklionense</i>	A chronic tenosynovitis associated with trauma and foreignbody introduction in an otherwise healthy individual
<i>M. immunogenum</i>	Implicated in cutaneous infection in both healthy and immunosuppressed patients.
<i>M. iranikum</i>	Infrequent human pathogen
<i>M. kyorinense</i>	Pathogenic for humans and have substantial clinical effects.
<i>M. lentiflavum</i>	Infections involve in skin or lymph node. Also, it has been isolated from pleural effusions, ascites, and from lung tissues, chronic pulmonary infection in immunodominant

	patients
<i>M. lepromatosis</i>	Lepromatous leprosy (LL) and diffuse lepromatous leprosy (DLL)
<i>M. llatzerense</i>	Abdominal Abscess
<i>M. mageritense</i>	<i>M. mageritense</i> pneumonia in an immunocompromised patient.
<i>M. kansasii</i> ATCC 12478	Pulmonary disease in immunocompromised individual. Found in aquatic environment
<i>M. nebraskense</i>	Nodular pulmonary disease
<i>M. neoaurum</i>	Bloodstream infection is rare and occurs most often in immunocompromised hosts present with undifferentiated fever and have an indwelling venous catheter
<i>M. obuense</i>	Capable of Forming a Black Product from p-Aminosalicylate and Salicylate, Heat-killed <i>M. obuense</i> is immunomodulatory and has been used to direct the immune response in the treatment of cancers - notably pancreatic cancer and malignant melanoma
<i>M. orygis</i>	Causative agent of tuberculosis in animals and humans
<i>M. parascrofulaceum</i>	Opportunistic pathogen, pulmonary infection
<i>M. rhodesiae</i>	Pulmonary disease
<i>M. senegalense</i>	Cause disease among cattle in east Africa
<i>M. setense</i>	Traumatic chronic skin abscess associated with osteitis
<i>M. septicum</i>	In the immunosuppressed patient, these organisms cause serious infections such as catheter-related bacteraemia or disseminated disease
<i>M. simiae</i>	lymphadenitis in an immunocompetent children, pediatric case, mostly in pulmonary and reticuloendothelial system
<i>M. arupense</i>	Pulmonary infection and tenosynovitis
<i>M. triplex</i>	Causes episodic in Immunocompetent Host
<i>M. tusciae</i>	Cervical lymphadenitis in immunocomponent host
<i>M. vaccae</i>	Low pathogenicity for humans, heat-killed <i>M. vaccae</i> immunotherapeutic agent
<i>M. vulneris</i>	Non-tuberculosis opportunistic pathogen
<i>M. xenopi</i>	Pulmonary infection
<i>M. gastri</i> 'Wayne'	Reports of pediatric infection Casual resident of human stomach

The PHYRE2 results that have been identified with ‘high’ confidence and are described as "certain" were initially considered. Then the proteins corresponding to significant coverage over full-length of the sequence were chosen as the feasible folds. The 3D models created by PHYRE2 based on the templates identified and the sequence alignments to each of the templates were further manually inspected to examine whether the functionally significant active site residues and the important residues for the cofactor binding were conserved.

Under the circumstances where such conservation was found, the corresponding domains were assigned as the probable protein fold. In this chapter we have discussed the different domains identified in the PE and PPE proteins. The 3D model thus generated using PHYRE2 was visualized on graphics to inspect the protein fold and active site region by comparing with the template structural fold used for model building. A representative list of the PE and PPE proteins from various mycobacterial species and their probable structural template and structural fold/function is shown in Table 2.2.

**Table 2.2.** A representative list of proteins from mycobacterial species and their probable structural template and structural fold/function.

Gene Name	PE/PPE Family	Species	Structure template PDB ID	Structural fold/function
WP_003401047.1	PE	<i>M.tuberculosis</i> complex	3AJA:A	Hydrolase Domain
WP_003910048.1	PE	<i>M. africanum</i>	3AJA:A	Hydrolase Domain
WP_003407378.1	PE	<i>M. tuberculosis</i> complex	3AJA:A	Hydrolase Domain
WP_023369542.1	PE	<i>M. kansasii</i>	3AJA:A	Hydrolase Domain
WP_023369544.1	PE	<i>M. kansasii</i>	3AJA:A	Hydrolase Domain
WP_015354155.1	PE	<i>M. liflandii</i>	3AJA:A	Hydrolase Domain
WP_015356865.1	PE	<i>M. liflandii</i>	3AJA:A	Hydrolase Domain
WP_012396468.1	PE	<i>M. marinum</i>	3AJA:A	Hydrolase Domain
WP_012395068.1	PE	<i>M. marinum</i>	3AJA:A	Hydrolase Domain
WP_012395100.1	PE	<i>M. marinum</i>	3AJA:A	Hydrolase Domain
WP_012392353.1	PE	<i>M. marinum</i>	3AJA:A	Hydrolase Domain
WP_023368197.1	PE	<i>M. kansasii</i>	3AJA:A	Hydrolase Domain
WP_015356629.1	PE	<i>M. liflandii</i>	3AJA:A	Hydrolase Domain
WP_003413473.1	PPE	<i>M. tuberculosis</i> complex	3AJA:A	Hydrolase Domain
WP_023372897.1	PPE	<i>M. kansasii</i>	3AJA:A	Hydrolase Domain
WP_012393340.1	PPE	<i>M. marinum</i>	3AJA:A	Hydrolase Domain
Rv1800	PPE	<i>M. tuberculosis</i> H37Rv	3AJA:A	Hydrolase Domain
WP_003910441.1	PPE	<i>M. tuberculosis</i>	3AJA:A	Hydrolase Domain

		complex		
WP_003909943.1	PPE	<i>M. tuberculosis</i> complex	3AJA:A	Hydrolase Domain
Rv3539	PPE	<i>M. tuberculosis</i> H37Rv	3AJA:A	Hydrolase Domain
WP_023371977.1	PE	<i>M. kansasii</i>	3D7R:B	Hydrolase Domain
ETW23397.1	PE	<i>M. gastri</i> 'Wayne'	3D7R:B	Hydrolase Domain
WP_036353694.1	PE	<i>M. asiaticum</i>	3D7R:B	Hydrolase Domain
WP_036353699.1	PE	<i>M. asiaticum</i>	3D7R:B	Hydrolase Domain
CPR13259.1	PE	<i>M. bohemicum</i> DSM 44277	3D7R:B	Hydrolase Domain
WP_003416104.1	PE	<i>M. orygis</i>	3D7R:B	Hydrolase Domain
WP_023371984.1	PPE	<i>M. kansasii</i>	3D7R:B	Hydrolase Domain
WP_012393384.1	PPE	<i>M. marinum</i>	3D7R:B	Hydrolase Domain
ETW23392.1	PPE	<i>M. gastri</i> 'Wayne'	3D7R:B	Hydrolase Domain
CPR07079.1	PPE	<i>M. bohemicum</i> DSM 44277	3D7R:B	Hydrolase Domain
WP_010886098.1	PE	<i>M. tuberculosis</i> complex	4EHC:A	Aspartic proteinase domain
WP_003899118.1	PE	<i>M. tuberculosis</i> complex	4EHC:A	Aspartic proteinase domain
Rv0977	PPE	<i>M. tuberculosis</i> H37Rv	4EHC:A	Aspartic proteinase domain
CPR05380.1	PPE	<i>M. bohemicum</i> DSM 44277	4EHC:A	Aspartic proteinase domain
WP_023365779.1	PE	<i>M. kansasii</i>	4EHC:A	Aspartic proteinase domain
WP_015355774.1	PE	<i>M. liflandii</i>	4EHC:A	Aspartic proteinase domain
WP_011740369.1	PE	<i>M. ulcerans</i>	4EHC:A	Aspartic proteinase domain
WP_014000539.1	PE	<i>M. canettii</i>	4EHC:A	Aspartic proteinase domain
WP_036358526.1	PE	<i>M. asiaticum</i>	4EHC:A	Aspartic proteinase domain
WP_036409477.1	PE	<i>M. gastri</i>	4EHC:A	Aspartic proteinase domain
WP_036412329.1	PE	<i>M. gastri</i>	4EHC:A	Aspartic proteinase domain
WP_003403850.1	PE	<i>M. tuberculosis</i> complex	4PZ9:B	Glucosyl-3-phosphoglycerate phosphatase
WP_015357328.1	PE	<i>M. liflandii</i>	4PZ9:B	Glucosyl-3-phosphoglycerate phosphatase
CPR12073.1	PE	<i>M. bohemicum</i> DSM 44277	4PZ9:B	Glucosyl-3-phosphoglycerate phosphatase
WP_012394280.1	PE	<i>M. marinum</i>	3GD9:A	Laminaripentaose-producing beta-

				1,3-glucanase domain
WP_015355330.1	PE	<i>M. liflandii</i>	2DSK:A	Chitinase Domain
WP_011739252.1	PE	<i>M. ulcerans</i>	2DSK:A	Chitinase Domain
WP_012395848.1	PE	<i>M. marinum</i>	2DSK:A	Chitinase Domain
WP_036414736.1	PE	<i>M. gastri</i>	2DSK:A	Chitinase Domain
WP_023367572.1	PE	<i>M. kansasii</i>	2DSK:A	Chitinase Domain
MGAST_01715	PE	<i>M. gastri</i> 'Wayne'	1OA4:A and 2NLR	Endoglucanases
CPR09297.1	PE	<i>M. bohemicum</i> DSM 44277	1OA4:A and 2NLR	Endoglucanases
CPR09297.1	PE	<i>M. bohemicum</i> DSM 44277	3NDY	CMB domain (CBD domain)
ETW25608.1	PE	<i>M. gastri</i> 'Wayne'	3NDY	CMB domain (CBD domain)
WP_036414736.1	PE	<i>M. gastri</i>	3NDY	CMB domain (CBD domain)
AGC62230.1	PE	<i>M. liflandii</i> 128FXT	4L0E	CytochromeP450 domain
WP_023369269.1	PE	<i>M. kansasii</i>	3JRO:A	$\beta$ -propeller fold
Rv0980c	PE	<i>M. tuberculosis</i> H37Rv	1QNI:E	$\beta$ -propeller fold
WP_013988789.1	PE	<i>M. africanum</i>	1GQ1: B	$\beta$ -propeller fold
WP_013988787.1	PE	<i>M. tuberculosis</i> complex	1FWX:B	$\beta$ -propeller fold
CMB12570.1	PE	<i>M. tuberculosis</i>	3HRP	$\beta$ -propeller fold
WP_012396836.1	PPE	<i>M. marinum</i>	3NB2:B	$\beta$ - helix like
WP_023366810.1	PPE	<i>M. kansasii</i>	3NB2:B	$\beta$ - helix like
WP_023368051.1	PPE	<i>M. kansasii</i>	3NB2:B	$\beta$ - helix like
WP_023371336.1	PE	<i>M. kansasii</i>	3D59:B	Acetyl hydrolase/cutinase like fold
WP_015356298.1	PE	<i>M. liflandii</i>	3D59:B	Acetyl hydrolase/cutinase like fold
WP_036353055.1	PE	<i>M. asiaticum</i>	3D59:B	Acetyl hydrolase/cutinase like fold
CPR09862.1	PE	<i>M. bohemicum</i> DSM 44277	3D59:B	Acetyl hydrolase/cutinase like fold

### 2.3.1.1 Hydrolase Domain

The PE and PPE proteins of some mycobacterial species (*M. tuberculosis*, *M. bovis*, *M. africanum*, *M. canettii*, *M. marinum*, *M. kansasii*, *M. liflandii*, *M. heckeshornense*, *M. bovis BCG strain*, *M. asiaticum*, *M. caprae*, *M. nebraskense*, *M. gordonae*, *M. haemophilum*, *M. lentiflavum*, *M. simiae*, *M. triplex*, *M. sinense*, *M. arupense*, *M. heraklionense*, *M. neworleansense*) that were predicted to comprise a conserved C-terminal domain with prominent serine  $\alpha/\beta$  hydrolase fold are shown in the supplementary data (Appendix S1). This domain was built on the crystal structures of templates msmeg\_6394 from *M. smegmatis* str. MC2155 (PDB\_ID: 3AJA:A) and from *Staphylococcus aureus* (PDB\_ID: 3D7R:B) as described in the Chapter 3. These template structures correspond to hydrolase family that exhibit an overall  $\alpha/\beta$  serine hydrolase fold with central five to eight  $\beta$ -strands flanked by  $\alpha$ -helices on either side and has the functional characterization of lipases. Most hydrolase family proteins are described by a  $\beta$ -sheet core connected by  $\alpha$ -helices that forms an  $\alpha/\beta/\alpha$  sandwich with a characteristic pentapeptide sequence motif GxSxG, and conserved Ser, Asp and His as the catalytically important residues (Hotelier et al., 2004).

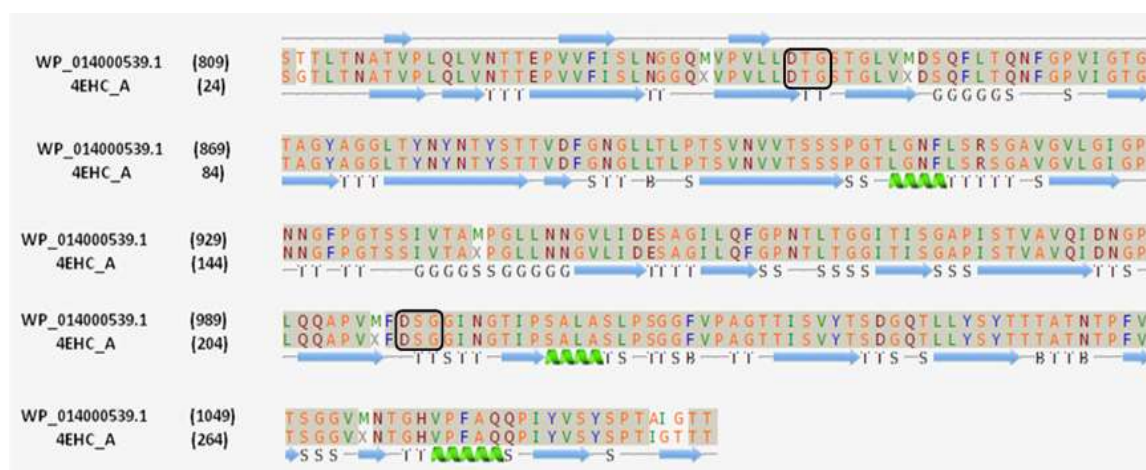
### 2.3.1.2 Aspartic Proteinase Domain

The PE proteins from mycobacterial species (*M. tuberculosis*, *M. bovis*, *M. caprae*, *M. africanum*, *M. caprae*, *M. orygis*, *M. canettii*, *M. gordonae*) were predicted to comprise the aspartic proteinase domain shown in the supplementary data (Appendix S2). This domain was identified in the PE\_PGRS16 (Rv0977) protein and its 3D crystal structure was solved recently by Bharathy and Suguna that was deposited with PDB\_ID: 4EHC (Barathy and Suguna, 2013). This aspartic proteinase domain has a low sequence similarity to Rv0977, HIV proteinase with a characteristic pepsin-fold, signature motifs of pepsin DTG and DSG and catalytic active site architecture.

The overall fold consists of a six  $\beta$ -strands located at the centre of the fold formed by the contribution of 3 strands in between the N- and C-terminal domains. The C-terminal domain comprises the conserved DTG motifs. These motifs are crucial for hydrolysis of the substrate containing peptide bond. In the present work, this type of fold for some of the PE proteins, for example, WP\_011740369.1, WP\_036358526.1, WP\_014000539.1 and WP\_036412329.1 (refer Appendix S2 for detailed list) consisting of a C-terminal 280 amino acid length domain was predicted to be the aspartic proteinase domain. This domain in the mycobacterial PE proteins was built on the crystal structure



of PDB\_ID: 4EHC\_A using PHYRE2. The models constructed comprise the two conserved DTG and DSG motifs. The alignment of the sequence to the template (~97% identity) generated by PHYRE2 is shown in Figure 2.1.A. The corresponding structure alignment of the model and template is shown in Figure 2.1.B and suggests the overall similarity in the protein fold which is characteristic of aspartic proteinases. In the hydrolysis reaction of a peptide, aspartic acid residues from the conserved motif play an important role, one of the aspartic acid acts as a general base while the other is a general acid which is followed by the nucleophilic attack facilitated by the catalytic water molecule (Dunn, 1989). In general, aspartic proteinase hydrolyses a peptide bond between hydrophobic residues. For instance, renin an aspartic proteinase exclusively cleaves angiotensinogen converting it to decapeptide angiotensin by preferentially cleaving Leu-Leu bond (Poulsen et al., 1976). Most of these proteins are drug targets such as HIV proteinase in AIDS and renin in hypertension (Cooper, 2002).



**Figure 2.1.A.** Sequence alignment of model WP\_014000539.1 and template PDB\_ID: 4EHC:A



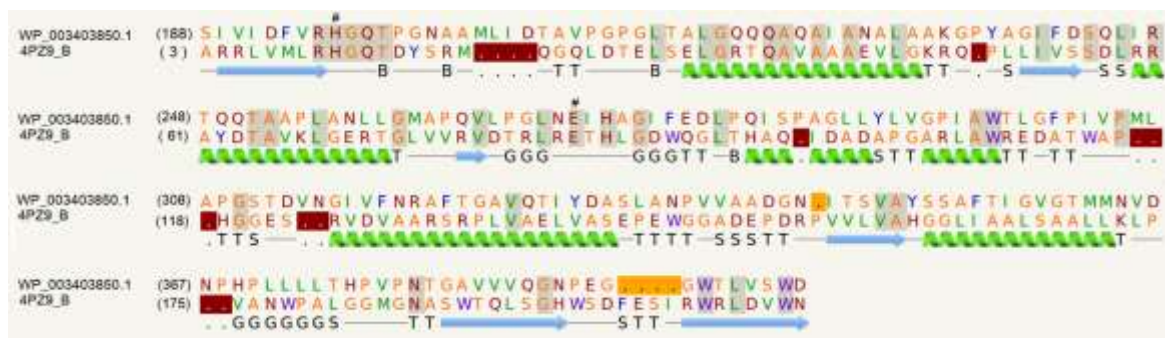
**Figure 2.1.B.** Structure alignment of model WP\_014000539.1 (green) and template PDB\_ID: 4EHC:A (pink)

### 2.3.1.3 Glucosyl-3-Phosphoglycerate Phosphatase Domain

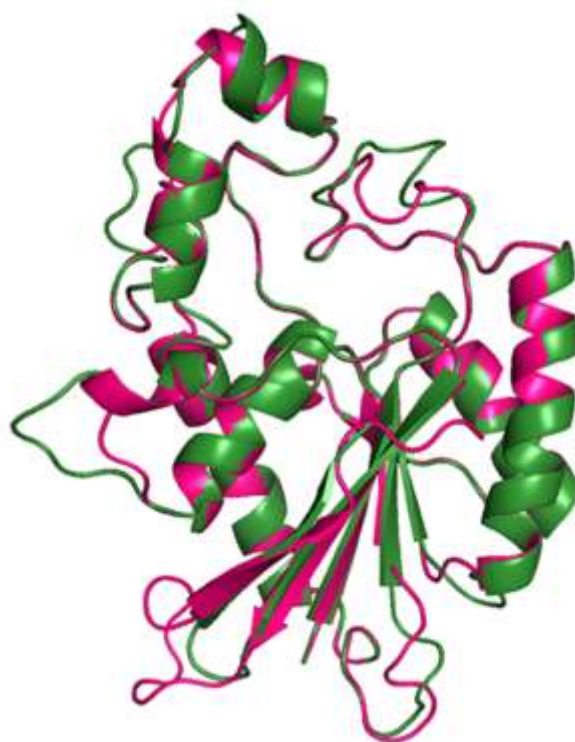
The list of PE proteins from mycobacterial species (*M. tuberculosis*, *M. bohemicum*, *M. haemophilum*, *M. canettii*, *M. kansasii*, *M. africanum*, *M. bovis*, *M. marinum*, *M. liflandii*, *M. xenopi*, *M. heckeshornense*, *M. gordonae*) comprising the glucosyl-3-phosphoglycerate phosphatase domain is shown in the supplementary data (Appendix S3). Some of these PE proteins such as WP\_003403850.1, WP\_015357328.1 and CPR12073.1 recognized the template PDB\_ID: 4PZ9:B corresponding to the mycobacterial glucosyl-3-phosphoglycerate phosphatase Rv2419c (Zheng et al., 2014).

This enzyme comprises of a single domain which has a central  $\beta$ -sheet with flanking  $\alpha$ -helices on either side and is known to catalyze the second step in the biosynthesis of methylglucose lipopolysaccharides (MGLPs) pathway. The synthesis of mycolic acids is regulated by MGLPs that is a vital lipid component in the mycobacterium cell wall. Therefore it can serve as an important target for the development of anti-TB drugs. The alignment of the sequence to the template (~21% identity) generated by PHYRE2 is shown in Figure 2.2.A and highlights regions of secondary structure that was used for constructing a 3D structure. The corresponding

structure alignment of the model and template is shown in Figure 2.2.B and suggests the overall similarity in the protein fold.



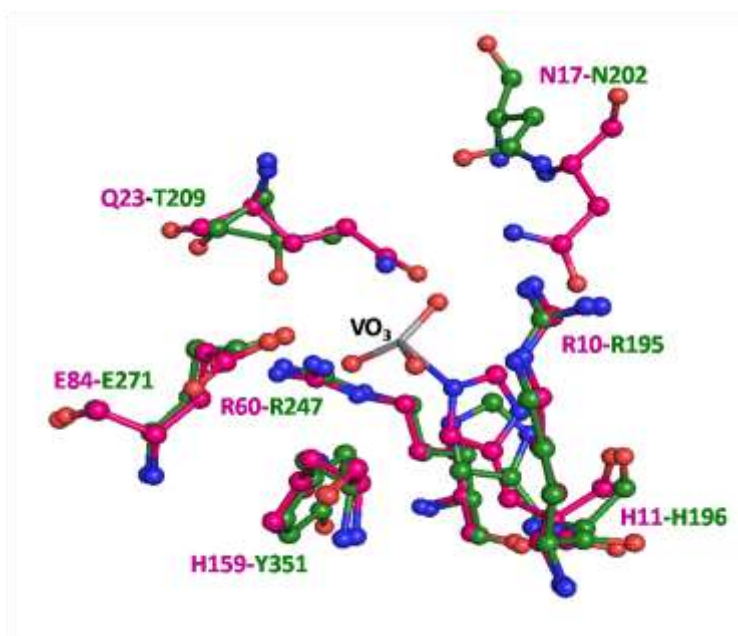
**Figure 2.2.A.** Sequence alignment of model WP\_003403850.1 and template PDB\_ID: 4PZ9:B



**Figure 2.2.B.** Structure alignment of model WP\_003403850.1 (green) and template PDB\_ID: 4PZ9:B (pink)

In the crystal structure (PDB\_ID: 4QIH), the active site is positioned in a positively charged cleft placed above a central  $\beta$ -sheet. Explicit electron density of a vanadate ion covalently linked to His11 (numbering according to PDB\_ID: 4PZ9) similar to the phosphohistidine intermediate and acetate ion was observed. In WP\_003403850.1

too, the catalytic residues His11 and Glu84 are conserved (Figure 2.2.C) and almost all the ligand binding residues are near to the acetate and vanadate, for instance, Arg10, His11, Asn17, Gln23, Arg60, Glu84, His159 and Leu209 which are important for enzymatic activity are conserved, except Gln23 which was replaced by Thr209 and His159 by Tyr351. According to the structure analyses, most residues being conserved, especially residues close to vanadate, suggests that the protein function is also likely to be conserved. From the structure analyses we believe that these subtle mutations would still preserve the protein function.



**Figure 2.2.C.** Superimposition of active site residues (ball and stick) in model WP\_003403850.1 (green) and template PDB\_ID: 4PZ9:B (pink)

#### 2.3.1.4 Laminaripentaose-Producing Beta-1,3-Glucanase Domain

Another interesting domain identified during this work is laminaripentaose-producing beta-1,3-glucanase (LPHase) domain present in some of the PE proteins. The list of PE proteins from mycobacterial species (*M. marinum*, *Mycobacterium sp. 012931*, *M. ulcerans str. Harvey*) comprising the laminaripentaose-producing beta-1,3-glucanase domain is shown in the supplementary data (Appendix S4). The PE protein, for example, WP\_012394280.1 from *M. marinum* C-terminal region of this protein recognized the template PDB\_ID: 3GD9:A, the crystal structure of LPHase in complex with laminaritetraose (Wu et al., 2009). LPHase belongs to glycoside hydrolase family 64 protein that hydrolyses a lengthy polysaccharide  $\beta$ -1,3-glucan to specific pentasaccharide oligomers. Glycoside hydrolases are grouped into families based on the similarity in the

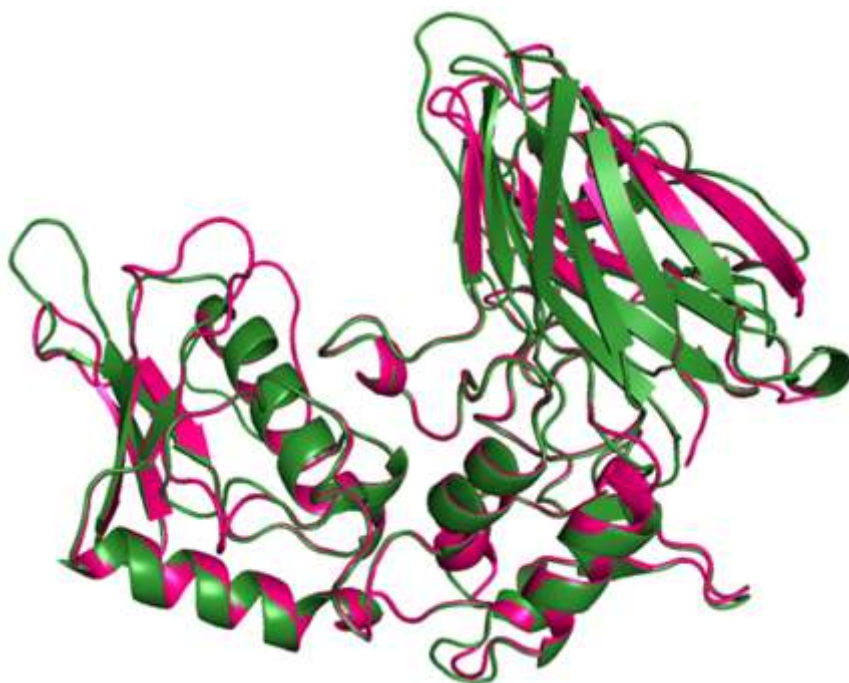


sequences and have been further classified into clans based on the resemblance in the overall fold, active site pocket and the catalytic mechanism (Henrissat and Bairoch, 1996; Henrissat et al., 1995; Henrissat and Davies, 1997).

The protein structure comprises of a crescent-like fold; a barrel shape domain and a mixed ( $\alpha/\beta$ ) domain that forms a wide-open groove in between the two domains. The sequence alignment obtained with highlighting the secondary structures predicted, is shown in Figure 2.3.A and has ~24% identity. The structural overlay of the model and template is shown in Figure 2.3.B that suggests the protein has overall similar fold and certain variable loop regions. The glycoside hydrolases are known to catalyze the hydrolysis of the glycosidic bond between carbohydrates or between a carbohydrate and non-carbohydrate moieties (Davies and Henrissat, 1995). Based on the nature of the organism, these enzymes play a variety of roles, such as degradation of biomass by cellulases (disintegrate celluloses into small carbohydrate moieties), exhibit pathogenesis in the activity of influenza virus neuraminidase (Colman, 1994). They are also involved in the normal cellular metabolic procedure along with glycosyl transferase in the formation and breaking of glycosidic bonds (Sinnott, 1990).

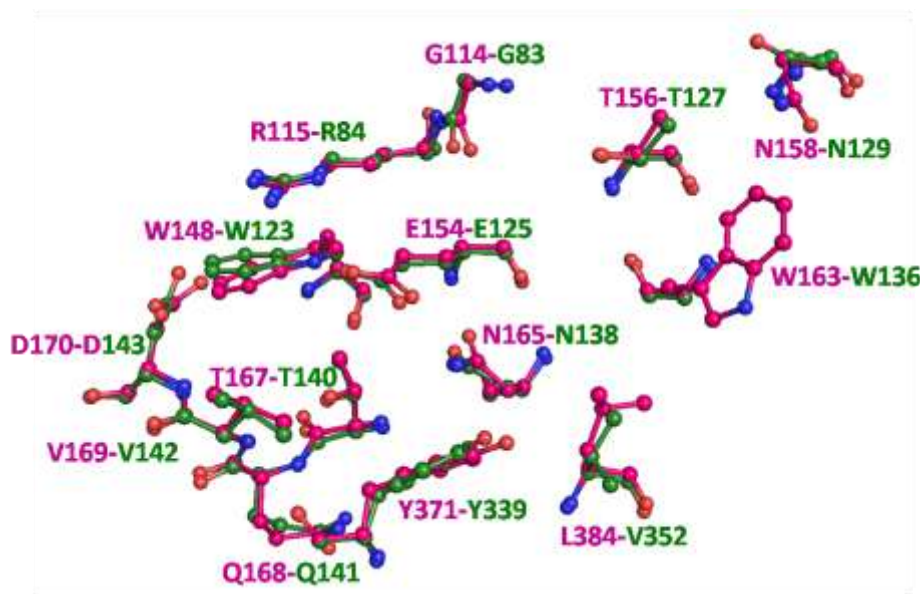


**Figure 2.3.A.** Sequence alignment of model WP\_012394280.1 and template PDB\_ID: 3GD9:A



**Figure 2.3.B.** Structure alignment of model WP\_012394280.1 (green) and template PDB\_ID: 3GD9:A (pink)

Our model suggests the conservation of catalytic residues; Glu154 and Asp170 (numbering according to the PDB\_ID: 3GD9) as shown in Figure 2.3.C. Between the N and C-terminal domains, the model contains an electronegatively charged wide groove comprising several conserved residues that include the above catalytic residues and four amino acid residues; Thr156, Asn158, Trp163 and Thr167 involved in sugar binding that accommodate the laminaritetraose molecule. According to the crystal structure of LPHase (PDB\_ID:3GD9), the enzyme uses a direct displacement mechanism that involves Glu154 and Asp170 residues through acid-base catalysis in the cleavage of  $\beta$ -1,3-glucan to specific  $\alpha$ -pentasaccharide oligomer moiety.

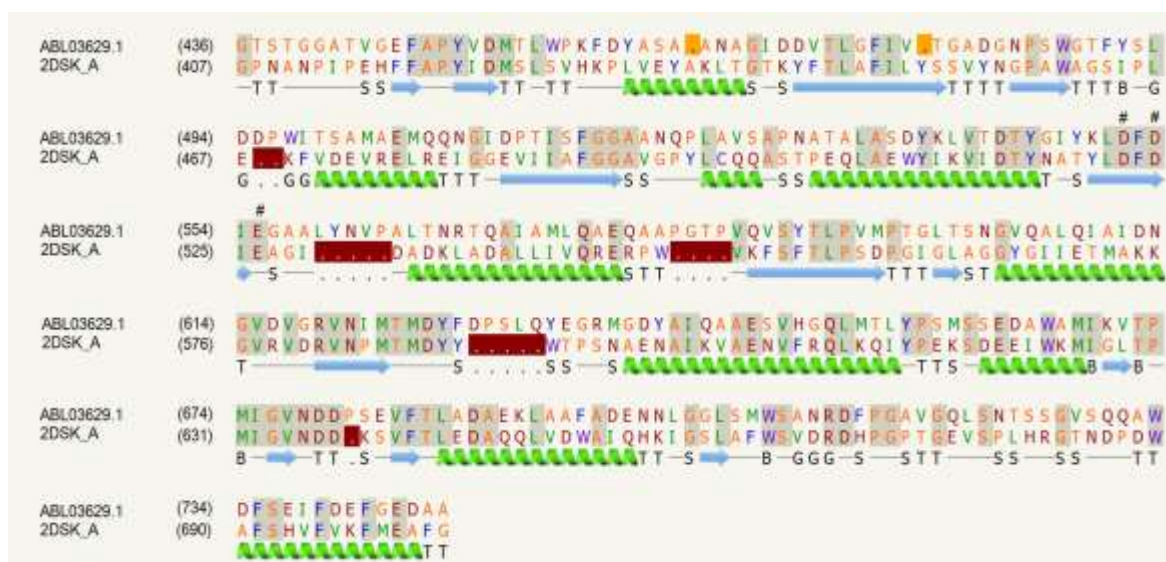


**Figure 2.3.C.** Superimposition of active site residues (ball and stick) in model WP\_012394280.1 (green) and template PDB\_ID: 3GD9:A (pink)

### 2.3.1.5 Chitinase Domain

Some of the PE proteins were found to comprise the chitinase domain. The list of PE proteins from mycobacterial species (*M. liflandii*, *M. marinum*, *Mycobacterium sp.* 012931, *M. ulcerans str. Harvey*, *M. gastri*, *M. goodnae*) comprising the chitinase domain is shown in the supplementary data (Appendix S5). Some of these PE proteins, for instance, ABL03629.1, WP\_015355330.1, WP\_023367572.1, WP\_036414736.1 and WP\_012395848.1 identified the template PDB\_ID: 2DSK:A which corresponds to the structure of chitinase domain. From the sequence alignment generated by PHYRE2 the catalytic site residues; Asp522, Asp524 and Glu526 that form the characteristic DXDXE motif observed in the crystal structure were conserved in these PE proteins as shown for one of the illustrative examples in Figure 2.4.A that shares ~37% sequence identity. The model was constructed on the crystal structure of the hyperthermophilic chitinase domain from *Pyrococcus furiosus* (PDB\_ID: 2DSK:A). The overall structure of the chitinase domain comprises a TIM-barrel fold with a tunnel-like active site, commonly observed in family 18 chitinases. The high degree of the overall structural similarity is shown in Figure 2.4.B. The chitinases are classified into two families (families 18 and 19 in the CAZy database; <http://www.cazy.org/>) according to amino acid sequence similarity (Henrissat, 1991) and hydrolyzes chitin polymer of  $\beta$ -1,4-linked N-acetylglucosamine.





**Figure 2.4.A.** Sequence alignment of model ABL03629.1 and template PDB \_ID: 2DSK:A



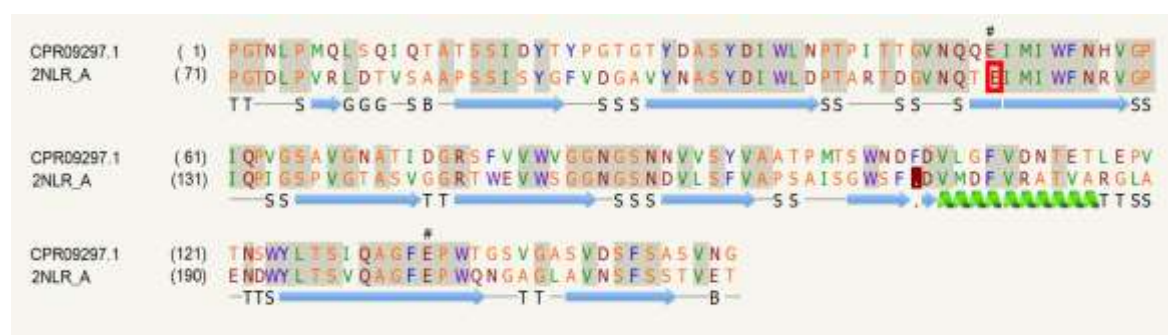
**Figure 2.4.B.** Structure alignment of model ABL03629.1 (green) and template PDB \_ID: 2DSK:A (pink)

### 2.3.1.6 Endoglucanase Domain

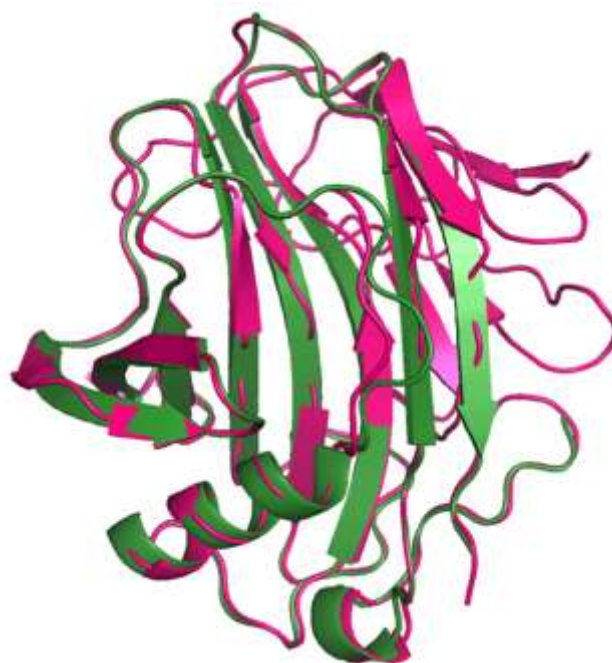
Some of the PE proteins, for example, MGAST\_01715 and CPR09297.1 have a conserved C-terminal region. These regions of PE proteins were constructed on PDB\_IDs: 1OA4:A and 2NLR:A. The list of PE proteins from mycobacterial species (*M. kansasii*, *M. gastri*, *M. goodnae*, *M. bohemicum* DSM 44277, *M. asiaticum*) comprising the endoglucanase domain is shown in the supplementary data (Appendix S6). The crystal



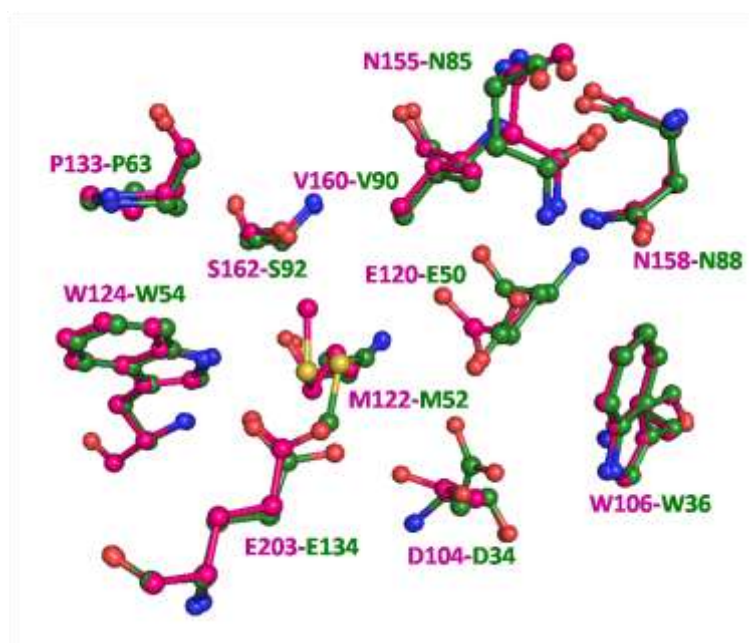
structure with PDB\_IDs: 1OA4:A and 2NLR:A corresponds to the endoglucanases family. These endoglucanase family belongs to glycoside hydrolase clan (GH-C groups) families of 11 xylanases and 12 cellulases, that contribute a jelly-roll topology. The structure primarily comprises of two anti-parallel  $\beta$ -sheets that forms a long substrate-binding pocket. The catalytic mechanism of these enzymes involves a double-displacement mechanism via covalent glycosyl enzyme intermediate, which is subsequently hydrolyzed with acid-base assistance, via oxocarbenium ion transition state and performs the catalysis with net retention of anomeric configuration (Davies, 1997; Koshland, 1953). Figure 2.5.A shows the sequence alignment (~58% identity) along with the catalytic residues and the secondary structural information. The structural comparison is shown in Figure 2.5.B. At one face of the enzyme there is a concave like surface made by a larger  $\beta$ -sheet that resulted in a wide substrate-binding cleft (Torronen et al., 1994). Further, this cleft has two invariant catalytic residues Glu120 and Glu203 (amino acid numbering according to PDB\_IDs: 2NLR) pointing towards the active site cleft from opposite sides. These residues are also conserved in PE proteins as shown in Figure 2.5.C. A long loop across the substrate-binding cleft terminates at the reducing end called as “cord”, commonly observed in all structures of family 11 and family 12. Certain residues located in the loop, importantly, Pro133 is conserved throughout clan GH-C members. The purpose of this loop remains unidentified, but there is a speculation that there is a probable loop movement over substrate binding (Torronen et al., 1994).



**Figure 2.5.A.** Sequence alignment of model CPR09297.1 and template PDB\_IDs: 2NLR



**Figure 2.5.B.** Structure alignment of model CPR09297.1 (green) and template PDB\_IDs: 2NLR (pink)



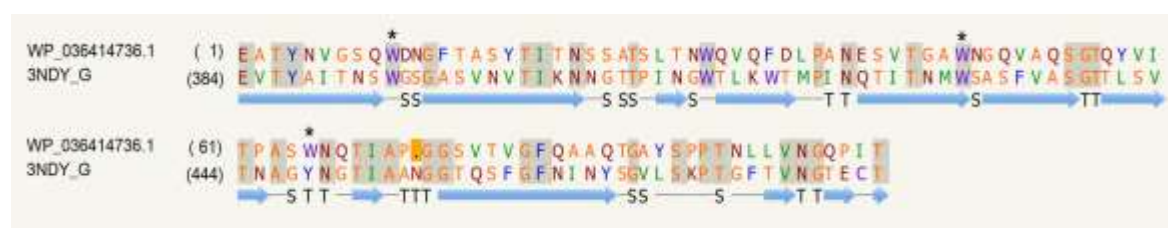
**Figure 2.5.C.** Superimposition of active site residues (ball and stick) in model CPR09297.1 (green) and template PDB\_IDs: 2NLR (pink)

### 2.3.1.7 Carbohydrate Binding Domain

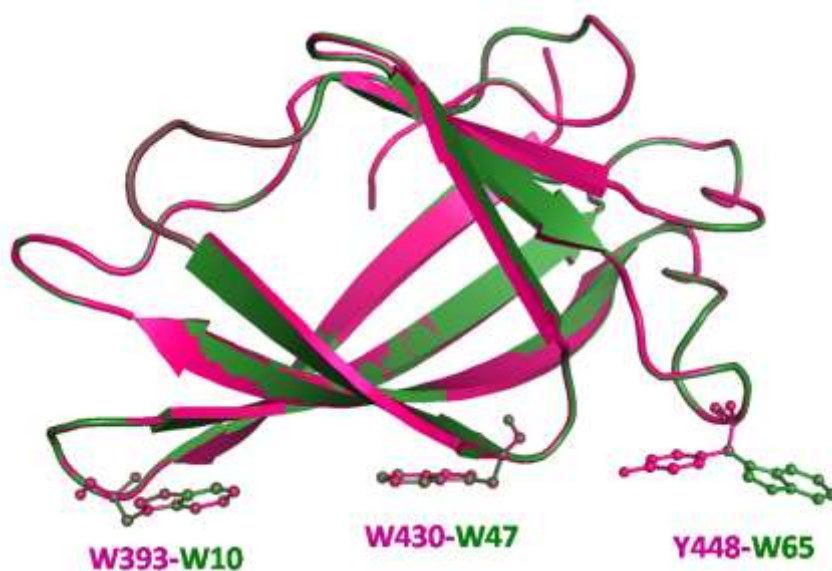
Another important domain found was carbohydrate binding domain in PE proteins from mycobacterial species (*M. kansasii*, *M. gastri*, *Mycobacterium* sp. 012931, *M. bohemicum* DSM 44277) shown in the supplementary data (Appendix S7). In the cellulose degradation method, the binding of the cellulosomes or cellulolase enzyme to the cell

surface is mediated via carbohydrate binding domain (CBD). In general CBD comprises of ~100 amino acid residues (Tormo et al., 1996). CBDs are separated from the cellulose domain by a short amino acid linker region and the enzymatic degradation is carried out by the cellulolytic domain (Din et al., 1994; Din N, 1991).

From our analysis, we observed that the PE proteins CPR09297.1 comprising an endoglucanase domain described above has one CBD, where as ETW25608.1 has two successive CBDs that are separated by 87 amino acids. The two proteins with CBD domains are linked with C-terminal glycosyl hydrolase 12 (GH12) family domains discussed above. Likewise, another PE protein; WP\_036414736.1, comprises a CBD that is linked to a glycosyl hydrolase 18 (GH18) family domain. The 3D structures of these CBDs were built on the CBD domain present in the endoglucanase D from the species *Clostridium cellulovorans* (PDB\_ID: 3NDY). The sequence alignment obtained from PHYRE2 (~35% identity) is shown in Figure 2.6.A. The alignment demonstrates that the secondary structure is mainly comprised of  $\beta$ -strands. The structural overlay of model with template containing nearly 100 amino acids is shown in Figure 2.6.B which revealed a  $\beta$ -sheet containing eight major  $\beta$ -strands. Three conserved hydrophobic amino acids (Trp, Trp, Trp/Tyr) 'strip' are situated on the loops connecting the  $\beta$ -strands in the model and the template are shown in Figure 2.6.B. The pi-electron density in the aromatic rings of the 'strip' contacts the cellulose hydrophobic region and impel the enzymes to carry out their catalytic functions (Doxey et al., 2010; Johnson et al., 1996).



**Figure 2.6.A.** Sequence alignment of model WP\_036414736.1 and template PDB\_ID: 3NDY

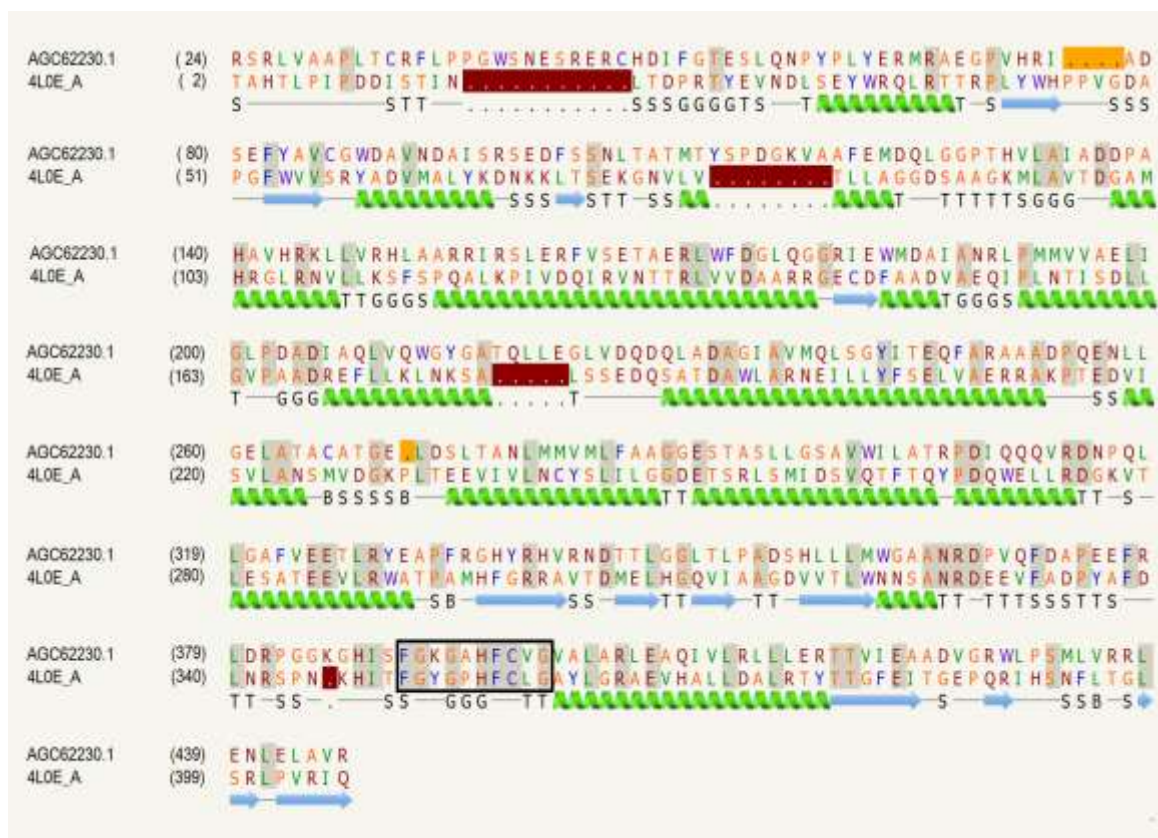


**Figure 2.6.B.** Structure alignment of model WP\_036414736.1 (green) and template PDB\_ID: 3NDY (pink)

### 2.3.1.8 Cytochrome P450 Domain

The PE protein AGC62230.1 with 1761 amino acid residues from *M. liflandii* 128FXT consists of a 500 amino acid C-terminal domain. For the C-terminal region of this PE protein, PHYRE2 identified cytochrome P450 fold from an organism *Streptomyces sp. Acta* 2897 (PDB\_ID:4L0E) (Uhlmann et al., 2013). Cytochrome P450s belong to the class of heme cofactor binding superfamily of proteins that is present in all domains of life. The remarkable high diversity in terms of sequences and functions resulted in an expansion of cytochrome P450 family. They catalyse many reactions, for instance, carbon heteroatom oxygenation, dealkylation, epoxidation, aromatic hydroxylation, reduction and dehalogenation (Danielson, 2002). A unique consensus sequence motif; 'FXXGXXXCXG' is present in all cytochrome P450s. This motif is located in between helices K and L forming a heme binding decapeptide loop (Song et al., 1993). This motif was also observed in the PE protein AGC62230.1 from *M. liflandii* 128FXT (Figure 2.7.A).



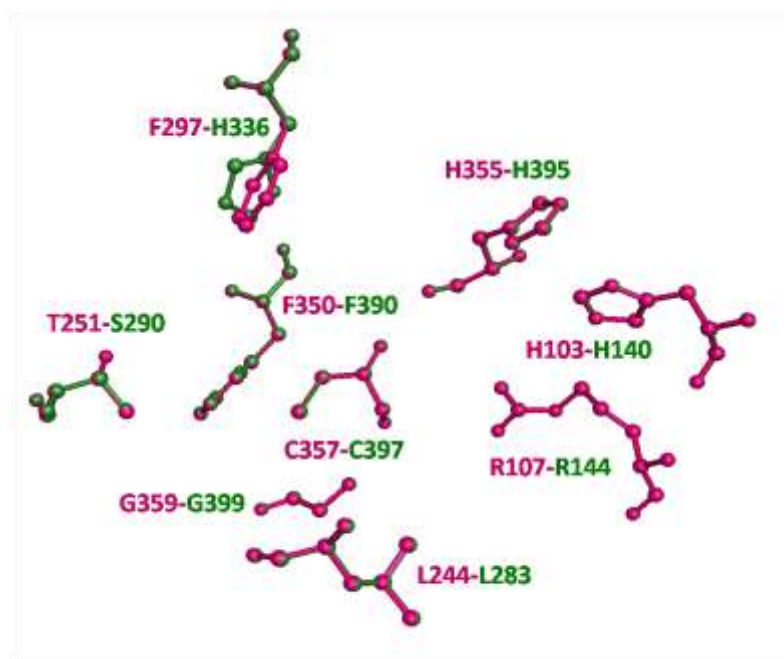


**Figure 2.7.A.** Sequence alignment of model AGC62230.1 and template PDB\_ID: 4L0E

The non-ribosomal peptide synthetases (NRPSs) are involved in the synthesis of diverse peptides known as non-ribosomally synthesised peptides (NRPs). NRPs have many applications in medical world (Hur et al., 2012). One of the important modifications found in NRPs is the preferential  $\beta$ -hydroxylation in various amino acids as well as the hydroxylation of weaker C-H bonds (Cryle, 2011) that are catalyzed by cytochrome P450 enzymes. In the PDB\_ID:4L0E, the cytochrome P450 encoded (sky32) is associated with the skylamycin biosynthesis gene cluster. The cyclodepsipeptide skylamycin A is an inhibitor moiety for the platelet derived growth factor signaling pathway isolated from streptomyces (Pohle et al., 2011). The crystal structure of sky32 (PDB\_ID: 4L0E) is responsible for the  $\beta$ -hydroxylation of three separate amino acids at positions 5 ( $\beta$ -hydroxyphenylalanine), 7 ( $\beta$ -hydroxy-OMe-tyrosine), and 11 ( $\beta$ -hydroxyleucine). The sequence alignment (~23% identity) corresponding to the cytochrome P450 domain in the above PE protein and sky32 mainly comprising  $\alpha$ -helices is shown in Figure 2.7.A. The comparison of the overall fold shown in Figure 2.7.B reveals the high degree of structural similarity. From the heme cofactor binding positions in both structures it can be seen that they are highly similar as indicated in Figure 2.7.C.



**Figure 2.7.B.** Structure alignment of model AGC62230.1 (green) and template PDB \_ID: 4L0E (pink)



**Figure 2.7.C.** Superimposition of active site residues (ball and stick) in model AGC62230.1 (green) and template PDB \_ID: 4L0E (pink)

### 2.3.1.9 Beta-Propeller

Adindla et al., in 2003 reported certain PE family proteins to comprise YVTN repeats (Adindla et al., 2004). These repeats contain 40-45 amino acid residues present in tandem copies along the protein sequence and located towards the C-terminus. We have identified several PE proteins from mycobacterial species (*M. tuberculosis*, *M. bovis*, *M. africanum* MAL020173, *M. caprae*, *M. orygis* 112400015, *M. canettii* CIPT 140070010, *M. haemophilum*, *M. marinum*) that contain the above repeats as shown in the supplementary data (Appendix S8). Some of these PE proteins containing YVTN repeats are WP\_023369269.1, CCP43730.1, WP\_013988789.1 and WP\_013988787.1. These proteins were modeled on the crystal structures of nitrous oxide reductase from *Pseudomonas nautical* (PDB\_ID: 1QNI:E) (Brown et al., 2000b), nitrous oxide reductase from *P. denitrificans* (PDB\_ID: 1FWX:B) (Brown et al., 2000a), cytochrome cd1 nitrite reductase (PDB\_ID: 1GQ1:B) (Gordon et al., 2003) and nup84-nup145c-sec13 (PDB\_ID: 3JRO:A) (Brohawn and Schwartz, 2009).

These diverse proteins commonly comprise a 6-8 bladed  $\beta$ -propeller fold. Typically,  $\beta$ -propellers contain 4-8 blades that are arranged circularly around a central axis where each blade consists of 4  $\beta$ -strands (Adindla et al., 2007). Several other diverse ~40-45 amino acid repeats, such as, WD, YWTD, YVTN etc., are known to be present in tandem and are known to fold as  $\beta$ -propellers (Adindla et al., 2004; Chen et al., 2011). The proteins containing the  $\beta$ -propellers are associated with varied functions such as hydrolases, transferases, transport, cell surface proteins, cofactor binding proteins, lyases and isomerases (Chen et al., 2011). In few cases, the active site is located in the loops connecting the tandem blades. For instance, the active site in the crystal structure of influenza neuraminidase (PDB\_ID: 1BJI) is situated in the region that is connected by several loops (Taylor et al., 1998). The sequence alignment (~13% identity) corresponding to the PE protein (nearly 200 amino acids); WP\_023369269.1 and its template is shown in Figure 2.8.A. The structural comparison of the model and template is shown in Figure 2.8.B. In some PE proteins, for example, CCP43730.1, WP\_013988787.1 and WP\_013988789.1, the C-terminal region accommodates only 2.5 to 4 repeats that would fold to form only a partial  $\beta$ -propeller (Figure 2.8.C). We therefore hypothesize that the C-terminal regions in these proteins could form either a four bladed  $\beta$ -propeller or dimerise to form six-bladed  $\beta$ -propeller.

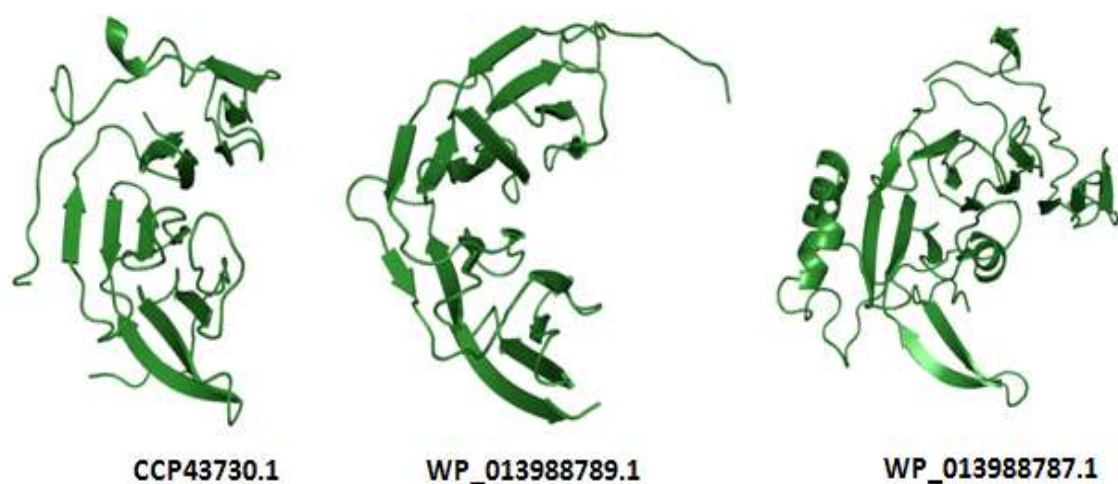


**Figure 2.8.A.** Sequence alignment of model WP\_023369269.1 and template PDB\_ID: 3JRO:A



**Figure 2.8.B.** Structure alignment of model WP\_023369269.1 (green) and template PDB\_ID: 1QNI:E (pink)





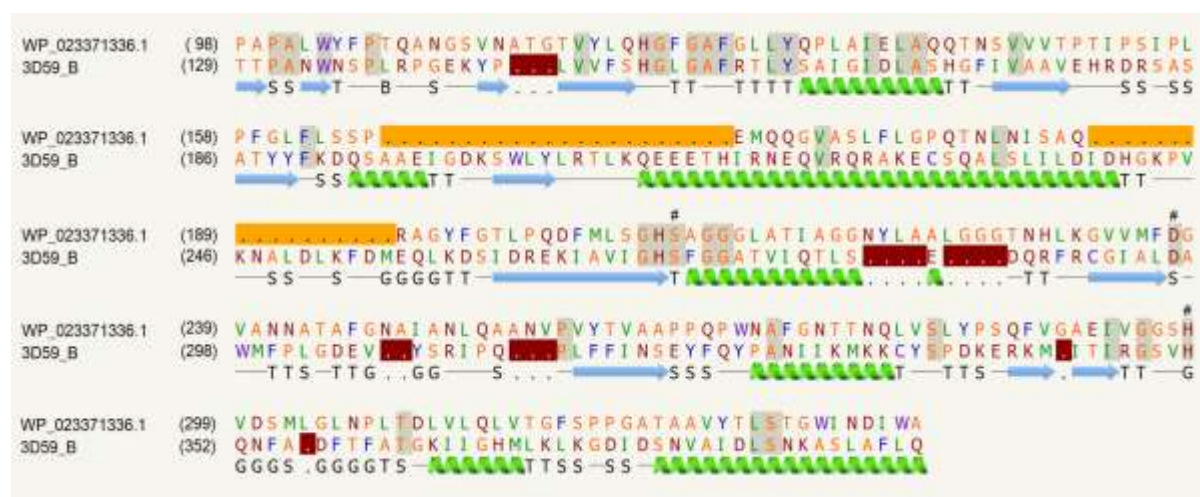
**Figure 2.8.C.** PE proteins models where only a part of the  $\beta$ -propeller are modeled, 2 – 4 blades that form a part of the  $\beta$ -propeller fold

### 2.3.2 Beta-Helix

Among various structural folds recognized, another interesting one is a  $\beta$ -helix identified by PHYRE2 in some of the PPE proteins. The  $\beta$ -helix predicted for the PPE proteins from mycobacterial species (*M. tuberculosis*, *M. kansasii*, *M. asiaticum*, *M. gastri*, *Mycobacterium sp. 012931*, *M. marinum*, *M. canettii*, *M. gordonae*, *M. ulcerans*, *M. bovis*, *M. orygis*, *M. liflandii*) are shown in the supplementary data (Appendix S9). Several PPE proteins are characterized by Gly-rich pentapeptide sequence repeats. Some of the PPE proteins (WP\_023366810.1, WP\_023368051.1 and WP\_012396836.1) were modeled on the N-terminal domain of a ubiquitin ligase (PDB \_ID: 3NB2:B) as the template. PDB \_ID: 3NB2:B is the crystal structure of *E. coli* o157:h7 which is an effector protein NleL. The crystal structure is composed of two structural domains: one at N-terminus with four stranded  $\beta$ -helix that is made up of pentapeptide sequence repeats and the other is located at the C-terminus with  $\alpha$ -helices due to the hydrogen bonds present between parallel  $\beta$ -sheet (Lin et al., 2012; Lin et al., 2011).  $\beta$ -helices were first described in the crystal structure of pectate lyase (Lietzke et al., 1994) and their functions, such as, cellulose or acid sugar binding lyases and polysaccharide lyases have been reviewed (Mitraki et al., 2002).  $\beta$ -helix can form two/three/four  $\beta$ -stranded helices that has a characteristic feature of the presence of Gly rich regions. The sequence alignment (~15% identity) predicted to mainly comprise the  $\beta$ -strands is shown in Figure 2.9.A and



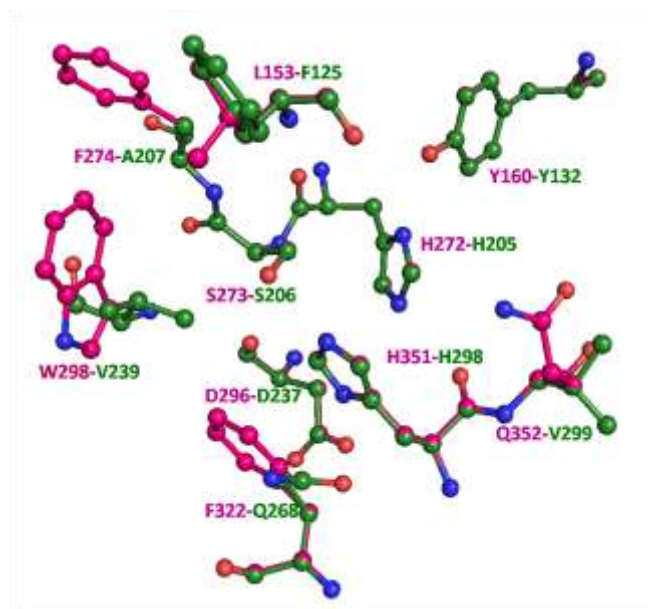
plasma platelet activating factor acetyl hydrolase (PDB\_IDs: 3D5E, 3D59:B) (Samanta and Bahnson, 2008). These structures have a typical  $\alpha/\beta$  hydrolase fold containing a catalytic triad of Ser, His and Asp. The alignment of the sequences (~15% identity) is shown in Figure 2.10.A. The structural comparison for the PE protein; WP\_023371336.1 with template is shown in Figure 2.10.B. The location of the catalytic residues Ser273, Asp296 and His351 (numbering according to PDB\_ID: 3D59) are shown in Figure 2.10.C. The homology model constructed superimposes on the template with low RMSD (0.17Å) that is indicative of high structural similarity.



**Figure 2.10.A.** Sequence alignment of model WP\_023371336.1 and template PDB\_ID: 3D59:B



**Figure 2.10.B.** Structure alignment of model WP\_023371336.1 (green) and template PDB\_ID: 3D59:B (pink)

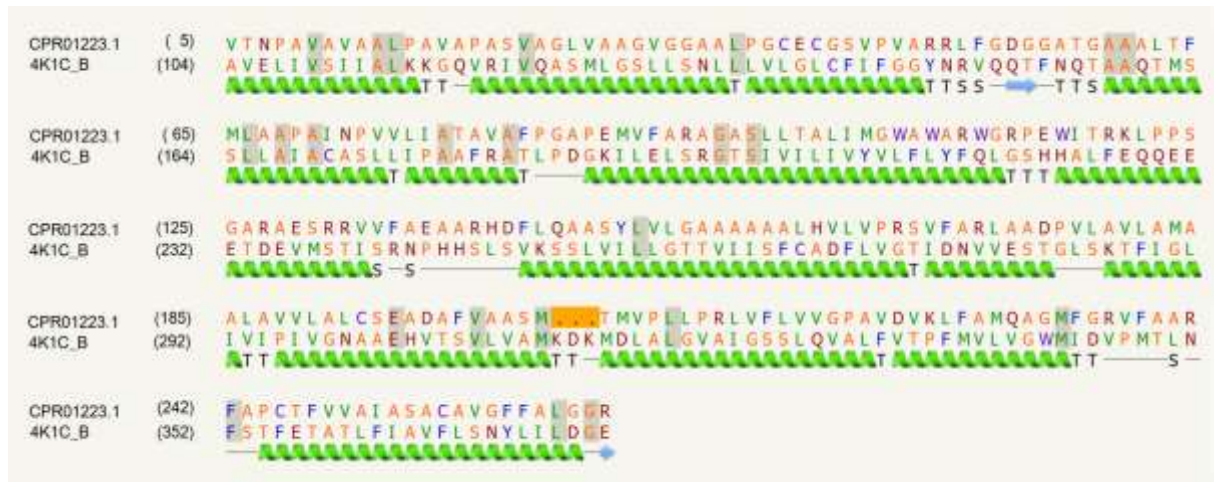


**Figure 2.10.C.** Superimposition of active site residues (ball and stick) in model WP\_023371336.1 (green) and template PDB\_ID: 3D59:B (pink)

### 2.3.2.2 Transmembrane Domain

Some PPE family proteins; WP\_036437315.1, WP\_036444818.1 and CPR01223.1 have been recognized as transmembrane helices with large stretches of intervening sequences suggesting these as membrane proteins. We modeled, one of the above proteins, CPR01223.1, a *M. bohemicum* protein on PDB\_ID: 4K1C, whose function is a eukaryotic calcium/proton exchanger (Waight et al., 2013). The sequence alignment that shares ~9% identity and predicted secondary structure is shown in Figure 2.11.A, suggesting that protein is mainly composed of  $\alpha$ -helices. The 3D model superimposed on the template is shown in Figure 2.11.B. From these results, it is believed that the PPE domain is situated in the cytoplasm, the role of helical segments is to traverse the lipid bilayer, thus making the protein as a transmembrane protein where the helical segments form a pore in the transmembrane that allows the passage of molecules. The proposed model of above mechanism of transmembrane PPE proteins is shown in Figure 2.11.C.





**Figure 2.11.A.** Sequence alignment of model CPR01223.1 and template PDB\_ID: 4K1C



**Figure 2.11.B.** Structure alignment of model CPR01223.1 (green) and template PDB\_ID: 4K1C (pink)

## 2.4 Conclusion

From our analysis of structural folds, we have identified certain well-characterized domains present in the PE and PPE proteins of *Mycobacterium*. These domains are known to be associated with a variety of functions, such as, hydrolysis of lipids ( $\alpha/\beta$  hydrolase fold and acetyl hydrolase), hydrolysis of carbohydrates (chitinase, endoglucanase and laminaripentaose-producing beta-1,3-glucanase domain) and hydrolysis of proteins (aspartic proteinase). Glucosyl-3-phosphoglycerate phosphatase plays an important role in the synthesis of mycobacterial cell wall components and cytochrome P450 domain is implicated in metabolic activity. The transmembrane domains,  $\beta$ -propellers, CBD and  $\beta$ -helices are regulators of protein function.

These results infer that some of the PE and PPE family proteins may be associated with enzymatic and regulatory roles. The  $\alpha/\beta$  hydrolase domain is present in both PE and PPE family proteins, whereas aspartic proteinase, chitinase, endoglucanase and  $\beta$ -propeller domains were only detected for PE proteins. The  $\beta$ -helix has been observed in PPE proteins. It was also observed that some domains such as  $\alpha/\beta$  hydrolase were present in all mycobacterial species, whereas some domains, such as, chitinase and endoglucanase were specific only to certain mycobacterial species. These observations suggest that the PE and PPE family proteins co-ordinate diverse roles that are mycobacterial species dependent.

During our work, we analyzed that although some structural folds were predicted with high confidence, few catalytic residues were not in conservation in the sequence alignment of some of the proteins. Therefore though the fold is conserved, the architecture is different that brings a variation in their structures. Out of several hundred diverse sequences that were analyzed, we were able to predict the fold with 'high' confidence for ~30% proteins. The structure and function predicted for the PE and PPE proteins discussed in this chapter provide the rationale for validation by experimental studies. Our work sheds new light on the structural and functional aspects of these important classes of mycobacterial.

\* The supplementary data of this chapter is added to the CD attached at the end of this thesis.

## Chapter 3

### **The PE-PPE Domain in *Mycobacterium* Reveals a Serine $\alpha/\beta$ Hydrolase Fold and Function: An *In Silico* Analysis**





### 3.1 Introduction

The whole genome sequence of the *Mtb* H37Rv strain revealed in 1998 (Cole et al., 1998), provided crucial information about the genes, physiology and pathogenesis responsible for highly infectious disease TB. The *mycobacterium* consists of a complex outer cell wall possessing an asymmetric lipid bilayer. The mycobacterial cell wall outer layer is composed of a layer of peptidoglycans that is linked through phosphodiester bond to the composite carbohydrates and arabinogalactans that are linked via high-molecular weight mycolic acids which forms a glycolipid (Rezwan et al., 2007). This tough cell envelope is often attributed to the impermeability of the drugs.

An important finding during this genome sequencing was the recognition of the PE and PPE gene families that comprise of about 10% of the whole genome (Cole et al., 1998). Subsequently, it was also identified that the PE and PPE gene families are *mycobacteria* specific. Almost 50% of these proteins contain only the characteristic homologous N-terminal domain, while the rest of the proteins comprise C-terminal extensions. Based on these C-terminal extension regions, the PE and PPE proteins were further categorized into various subfamilies (Cole et al., 1998). These variable C-terminal extensions appear to be the source of antigenic variation among various strains of this bacterium that guide to a speculation that these protein families could be pathogenic and immunologically important. Earlier studies on PE and PPE proteins from our lab by Adindla et al in 2003 reported a conserved domain comprising of 225 amino acid residues in some PE, PPE and hypothetical proteins of mycobacterial species. It was defined as PE-PPE domain (Pfam ID: PF08237) since it was commonly present at the C-terminus of both PE and PPE family of proteins (Adindla and Guruprasad, 2003).

Though the whole genome sequence information of *Mtb* is known for over 18 years, discovery of the specific function of all the PE and PPE proteins was not properly defined and remains a significant area of research aimed at the diagnosis and treatment of TB. The PE and PPE proteins have attracted much attention since they are cell wall associated and surface exposed proteins (Brennan et al., 2001; Delogu et al., 2004; Kaufmann and Helden, 2008; Sampson et al., 2001). The enzymatic functional characteristics of very few PE and PPE proteins have been reported so far. For example, the C-terminal domain of PE\_PGRS63 protein Rv3097c is homologous to the hormone-sensitive lipase family proteins with well defined conserved GDSAG motif and showed triacylglycerol hydrolase activity (Mishra et al., 2008). Rv3097c is induced under

starvation to hydrolyse the stored triacylglycerol suggesting its key role in nutrient deprived conditions of *Mtb* (Deb et al., 2006).

In *Mtb* genome, nearly 250 enzymes are reported as important in lipid metabolism and ~ 94 members are expected to comprise the  $\alpha/\beta$  hydrolase fold signifying the role of lipases/esterases/cutinases (Hotelier et al., 2004). Singh et al (Singh et al., 2010) provided an outlook of lipases and associated family members that are important in the virulence of *Mtb* and its pathogenicity. In *Mtb*, there are three well characterized members of antigen 85 complex Rv1886c, Rv3804c, and Rv0129c that are serine hydrolases and are vital for the transfer of mycolic acid in the cell wall biosynthesis (Ramulu et al., 2006). Rv2422c is a serine hydrolase and has been recognized as a cell wall associated virulence factor (Lun and Bishai, 2007) or important in triacylglycerol utilization during starvation (Deb et al., 2006). Further, seven cutinase-like proteins Rv1758, Rv1984c, Rv2301, Rv3451, Rv3452, Rv3724 and Rv3802c of *Mtb* strain H37Rv were identified using bioinformatics analysis. Various physiological functions and also immunological responses of cutinase like proteins were reported (West et al., 2008).

A fundamental and important aspect here is the presence of various lipid hydrolyzing enzymes in *Mycobacterium*. Nearly one third of the world population even today is infected with *Mtb* and is present in dormant form. Bacterium enters into the dormant form by first degrading the cell membrane of the host gathering the lipids required to resynthesize complex lipids for its survival in the host environment (Daniel et al., 2004). During dormancy condition, *mycobacteria* make use of fatty acids as a source of energy and stores fatty acids as triacylglycerol (Garton et al., 2002). Microscopic examination of the *mycobacteria* shows an intact cell envelope and huge lipid inclusion bodies present in the cytoplasm (Minnikin et al., 2002). Whenever there is a decrease in the immunity of the host, the pathogenic bacterium becomes virulent by entering into the reactivation phase. Thus, when the lipid inclusion bodies are hydrolysed, the infection develops beginning the onset of the disease. The proteins involved in the synthesis of complex cell wall of *Mtb* consisting of waxy outer layer, its infection in the host, survival in the dormant form and reactivation in host may be considered as some of the important drug targets.

The fusion of the PE or PPE proteins with PE-PPE domain is an example of gene fusion. Gene fusion during evolution means the occurrence of two genes, for example A and B with independent functions like a single bifunctional gene A–B in orthologs or paralogs. The gene fusion is indicative of the relatedness in the protein functions

(Overbeek et al., 1999). This is often observed in proteins of large size that are evolved as domains, so as to facilitate the protein folding and therefore to regulate the function. Domains usually comprise >60 amino acid residues that can have independent folding and functional units irrespective of their location along the protein sequence (Swathi Adindla, 2013). In this current chapter, we intended to find the characteristic structure and function of this PE-PPE domain using computational approaches like protein fold recognition methods.

## **3.2 Methods**

### **3.2.1 NCBI Protein Sequence Databank**

The amino acid sequence regions with respect to the PE-PPE domain from *Mtb* strain H37Rv were obtained from the NCBI protein sequence databank available at <http://www.ncbi.nlm.nih.gov/>. For instance, in the protein Rv1430, the amino acid sequence region between 108 to 337 is a PE-PPE domain. The PE-PPE domain regions were explored for sequence and structure analysis.

### **3.2.2 PSI-BLAST**

The PSI-BLAST method (Altschul et al., 1997) was used against non-redundant database at NCBI to identify all the proteins containing the PE-PPE domain region. To achieve this, the regions corresponding to PE-PPE domain in *Mtb* H37Rv proteins were used as query.

The PSI-BLAST search of PE-PPE domain region was carried out against PDB available at [www.rcsb.org/](http://www.rcsb.org/) containing sequences of known 3D structures, in order to select suitable templates for building the 3D structure models of the PE-PPE domain region using comparative modeling methods.

### **3.2.3 CLUSTALW**

CLUSTALW (Thompson et al., 1994) program was implemented for the multiple sequence alignment of the PE-PPE domain regions. This multiple sequence alignment helps to detect the extent of homology and variation between the new and existing families of sequences. CLUSTALW provides the consensus sequences and evolutionary history.

### **3.2.4 Signal Peptide Server**

To predict the presence and location of the signal sequence/signal peptide cleavage sites in the proteins comprising the PE-PPE domain, signal peptide server (Emanuelsson et al., 2007) was used. A signal sequence is generally a short amino acid sequence of nearly 5-30 amino acid residues length. It is believed that the proteins containing the signal peptides are driven towards the secretory pathway.

### **3.2.5 FUGUE**

To unveil the structure and function of the PE-PPE domain region, fold prediction method FUGUE (Shi et al., 2001) was used. This method recognizes the probable fold of the query sequence and also provides an alignment between the query protein and the likely template structures. The alignments generated by FUGUE are more specific and represents a better relatedness between the query protein sequence and the template structure. These alignments were further used in protein structure modeling steps.

### **3.2.6 Structure Modeling**

The alignments produced by FUGUE were used to construct the 3D structure models of the PE-PPE domain region using Homology module in InsightII (Accelrys Inc, USA) that is based on the methodology described in MODELLER (Sali and Blundell, 1993).

Homology module in InsightII automatically builds a model comprising non-hydrogen atoms by satisfying the spatial restraints that include especially important non-homologous loops. The stereochemistry of the model constructed is finally improved by energy optimization of the final model.

### **3.2.7 3D Model Structure Validation**

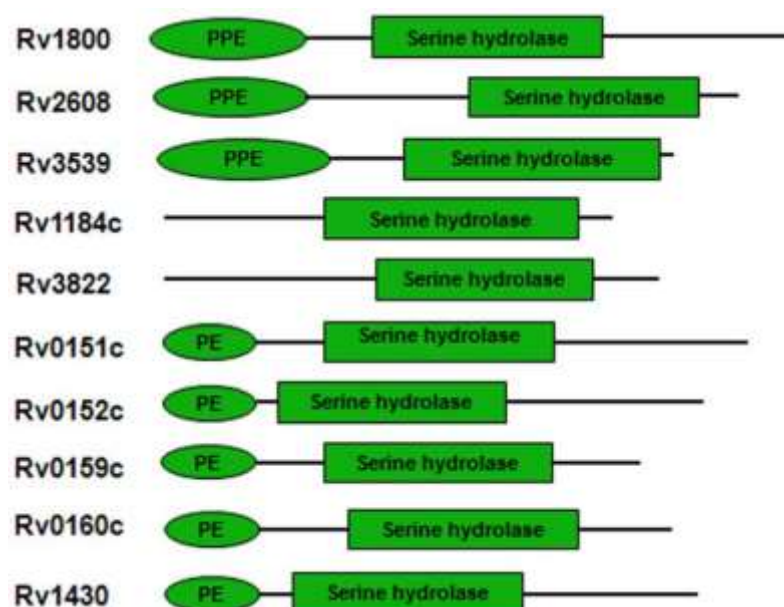
The reliability of the models constructed from above homology or comparative modeling methods were validated using PROCHECK (<http://nihserver.mbi.ucla.edu/SAVES/>) (Laskowski RA, 1993) and Verify\_3D (<http://nihserver.mbi.ucla.edu/SAVES/>) (Luthy et al., 1992). The stereochemical parameters of the protein structure model were measured using the automated PROCHECK method. Since the homology between the PE-PPE domain region and known 3D structures is quiet low, the 3D protein model structure was validated using a 3D profile that compares 3D protein model built with its own amino acid sequence, which is computed from the atomic coordinates of the structure.

### **3.2.8 MAPSCI**

The likely structural folds recognized by the FUGUE method and 3D model structures of the PE-PPE domains generated were superimposed using MAPSCI server (Ilinkin et al., 2010) (<http://www.geom-comp.umn.edu/mapsci/>). MAPSCI computes the multiple structure alignments by identifying the common substructures in a set of proteins that helps to evaluate the relatedness among the protein structure, function, and their evolutionary history.

### 3.3 Results and Discussion

Previously from our lab Adindla and Guruprasad 2003 reported that the PE-PPE domain is present in ten proteins in *Mtb* H37Rv strain. The PE-PPE domain architecture diagram of these ten proteins belonging to *Mtb* H37Rv strain is presented in Figure 3.1.



**Figure 3.1.** Domain architecture diagram of proteins in *Mtb* H37Rv genome comprising the PE-PPE domain

In the current work, the PSI-BLAST search results of the PE-PPE domain against non-redundant database at NCBI has shown the presence of the PE-PPE domain in all mycobacterial genomes. The detailed genome-wide sequence study of *mycobacteria* using PSI-BLAST searches showed that homologs of the PE and PPE proteins comprising the PE-PPE domain; Rv0151c, Rv0152c, Rv0159c, Rv0160c, Rv1430, Rv1800, Rv2608, Rv3539 are present in the following genomes of *mycobacteria*, namely *M.tuberculosis*, *M. bovis*, *M. kansasii*, *M. marinum* and *M. ulcerans*. While, the homologs of hypothetical proteins that is Rv1184c and Rv3822 comprising the PE-PPE domain only are present in the following genomes of *mycobacteria*, namely *M. tuberculosis*, *M. bovis*, *M. parascrofulaceum*, *M. smegmatis*, *M. vanbaalenii*, *M. abscessus*, *M. avium*, *M. leprae* and *Mycobacterium* sp. JLS, KMS and MCS. A representative list of the mycobacterial proteins that consists of the PE-PPE domain is shown in Table 3.1.

**Table 3.1.** A representative list of proteins comprising PE-PPE domain in mycobacterial genomes. NCBI\_IDs of the proteins are given.

Organism	Proteins with PE domain & PE-PPE domain	Proteins with PPE domain & PE-PPE domain	Hypothetical proteins with PE-PPE domain
<i>M. tuberculosis</i> CDC1551	NP_334571.1 NP_335924.1 NP_334570.1 NP_334576.1 NP_334577.1	NP_337185.1 NP_336306.1 NP_338188.1	NP_338483.1 NP_335664.1
<i>M. tuberculosis</i> H37Rv	YP_177696.1 YP_177810.1 YP_177695.1 YP_177697.1 YP_177698.1	YP_177893.1 YP_177839.1 YP_177987.1	NP_218339.1 NP_215700.1
<i>M. tuberculosis</i> H37Ra	YP_001282744.1 YP_001281438.1 YP_001281446.1 ZP_02552557.1 YP_001281447.1	YP_001283971.1 YP_001283129.1 YP_001284925.1	ZP_02548944.1 YP_001281439.1 ZP_02552547.1 ZP_02552214.1 YP_001285213.1 YP_001282494.1
<i>M. bovis</i> AF2122/97	NP_853823.1 NP_855117.1 NP_853822.1 NP_853831.1 NP_853830.1	NP_855481.1 NP_856286.1 NP_857208.1	NP_857489.1 NP_854870.1
<i>M. bovis</i> BCG str. Pasteur 1173P2	YP_977583.1 YP_976286.1 YP_976295.1 YP_976294.1	YP_977924.1 YP_978719.1 YP_979682.1	YP_976287.1 YP_979964.1 YP_977338.1
<i>M. bovis</i> BCG str. Tokyo 172	YP_002643224.1 YP_002644522.1 YP_002643223.1 YP_002643232.1 YP_002643231.1	YP_002644872.1 YP_002645676.1 YP_002646644.1	YP_002646925.1 YP_002644275.1
<i>M. kansasii</i> ATCC 12478	ZP_04750460 ZP_04746832 ZP_04746828 ZP_04746825 ZP_04746830 ZP_04746817	ZP_04751918 ZP_04751917 ZP_04747254	

	ZP_04746816 ZP_04746831 ZP_04751910 ZP_04746829 ZP_04746815 ZP_04750329		
<i>M. marinum</i> M	YP_001850539 YP_001848693 YP_001848692 YP_001848694 YP_001851707 YP_001848706 YP_001852582 YP_001852247 YP_001848705 YP_001848691 YP_001851742 YP_001848704 YP_001853198	YP_001849339 YP_001849803	
<i>M. ulcerans</i> Agy99	YP_905107.1 YP_905059.1	YP_906263.1	
<i>M. parascrofulaceum</i> ATCC BAA-614			ZP_06848426.1 ZP_06852104.1 ZP_06849370.1
<i>M. smegmatis</i> str. MC2 155			YP_889694.1 YP_887070.1 YP_888993.1 YP_890365.1 YP_884825.1 YP_889988.1 YP_889594.1 YP_885568.1
<i>Mycobacterium</i> sp. JLS			YP_001069995.1 YP_001068542.1 YP_001069270.1 YP_001072916.1 YP_001068711.1 YP_001068707.1 YP_001073586.1 YP_001071123.1 YP_001072947.1 YP_001072948.1
<i>Mycobacterium</i> sp. KMS			YP_937773.1 YP_936971.1



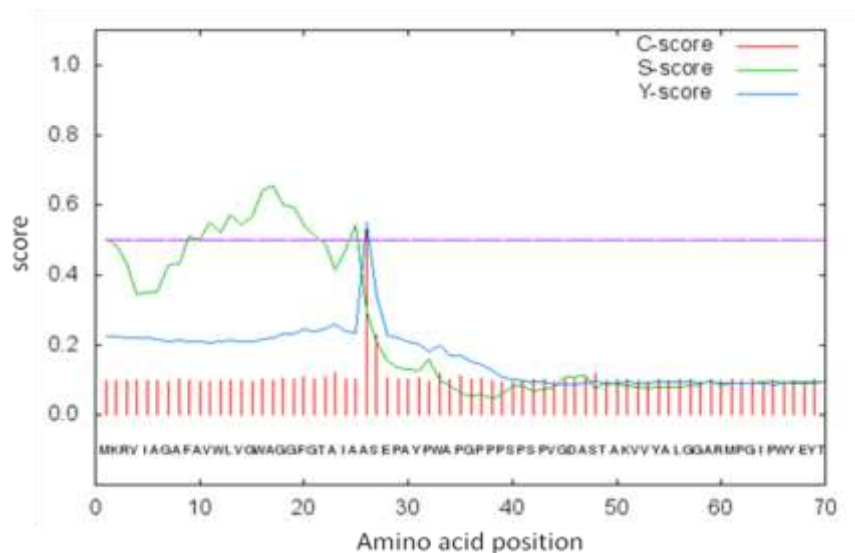
			YP_936265.1 YP_940348.1 YP_941018.1 YP_938853.1 YP_936435.1 YP_936436.1 YP_936431.1 YP_940378.1 YP_940379.1
<i>Mycobacterium</i> sp. MCS			YP_638900.1 YP_638120.1 YP_637425.1 YP_641442.1 YP_642111.1 YP_639988.1 YP_637596.1 YP_637597.1 YP_637592.1 YP_641472.1 YP_641473.1
<i>M. vanbaalenii</i> PYP-1			YP_952665.1 YP_953911.1 YP_953352.1 YP_955506.1 YP_952079.1 YP_951131.1 YP_953914.1 YP_953918.1 YP_953778.1 YP_955594.1 YP_953910.1 YP_951605.1 YP_950898.1 YP_951855.1 YP_952889.1 YP_955503.1 YP_950878.1 YP_956151.1 YP_955503.1 YP_950878.1 YP_952432.1 YP_954864.1
<i>M. abscessus</i> ATCC 19977			YP_001701682.1 YP_001704868.1
<i>M. avium</i> 104			YP_880985.1

<i>M. leprae</i> TN			NP_301893.1
<i>M. leprae</i> Br4923			YP_002503523.1

Our analysis from PSI-BLAST also revealed that homologs of the hypothetical proteins are present in some actinobacteria genomes like *Nocardia farcinica* (YP\_120155.1) and some species from *Rhodococcus* (YP\_002767666.1, YP\_002764179.1, YP\_002783136.1, YP\_002781243.1, YP\_705817.1, ZP\_04388076.1, ZP\_04387701.1, ADD80824.1) but these genomes do not contain the PE and PPE family proteins.

According to the results of signal peptide server, we identified that only the hypothetical protein Rv1184c among the ten proteins comprises the signal peptide cleavage site at the N-terminus between amino acids Ala26 and Ser27 as shown in the following Figure 3.2 and is therefore expected to be a secreted protein. The hypothetical protein Rv3822 was predicted to be a non-secretory protein by the signal peptide server.

According to the signal peptide server, we observed that Rv1184c protein is possibly exported protein, while Rv3822 is expected to be a non-secretory protein indicating that it is more likely cell wall associated. Secreted proteins play a major role in the pathogenicity of the bacterium, this is because many proteins are secreted during important stages of infection of the bacterium like macrophage phagosome and phagolysosomal fusion (Beatty and Russell, 2000). West et al., in 2009 reported seven cutinase like proteins (CULPs), out of which all are secreted proteins except CULP5. Nevertheless, CULP6 is shown as an essential cell wall associated protein in spite of comprising a predicted signal peptide (West et al., 2009) suggesting that experimental studies are necessary to decipher the localization of the hypothetical proteins Rv1184c and Rv3822 in *Mtb*.



**Figure 3.2.** Signal peptide cleavage of Rv1184c at the positions Ala26 and Ser27

where;

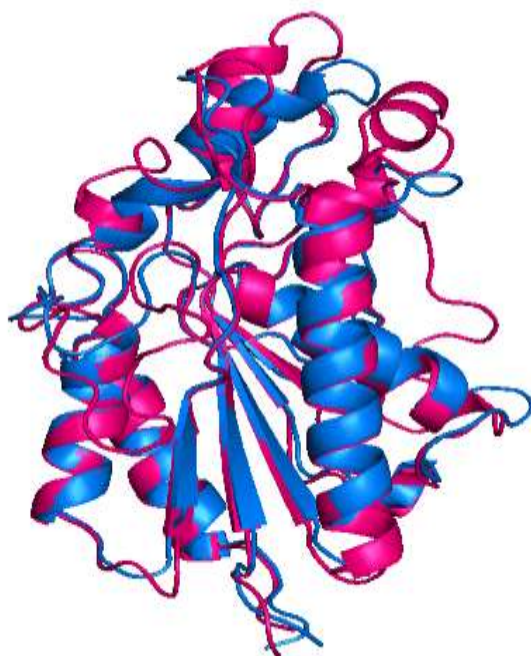
C score - raw cleavage site

S score - signal peptide score

Y score - combined cleavage site score

To find the probable fold and function of the PPE-PPE domain we have carried out the PSI-BLAST search against the PDB database. It identified the PDB\_ID: 1G29 with high E- value, 0.068. PDB\_ID: 1G29 is ATPase subunit, a member of the trehalose/maltose ABC transporter family that belongs to archaeon "*Thermococcus litoralis*". The high E-value indicated very low sequence homology of proteins that is between the PE-PPE domain and the identified 3D structure. Nonetheless, FUGUE program recognized PDB\_ID: 3AJA as the feasible fold with highest Z- score of 21.62. Further all the other proteins identified by FUGUE program as likely structures were PDB\_IDs: 1CEX, 1BS9, 2CZQ and 3HC7 with Z- scores >6.0 demonstrating the confidence in fold prediction analysis. The PDB\_ID: 3AJA is the crystal structure of a lipase that belongs to *M. smegmatis* strain MC2155, the PDB\_ID: 1CEX is a *Fusarium solani* cutinase, the PDB\_ID: 1BS9 is an acetylxyylan esterase that belongs to *Penicillium purpurogenum*, the PDB\_ID: 2CZQ is the crystal structure of cutinase like protein belonging to *Cryptococcus* sp. S-2, the PDB\_ID: 3HC7 is the crystal structure of mycobacteriophage esterase that catalyses the cleavage of the mycolylarabinogalactan covalent bond releasing free mycolic acids. The comparison of PSI-BLAST and FUGUE results clearly indicated that the fold based structure recognition methods have an advantage over the homology based sequence searches alone.

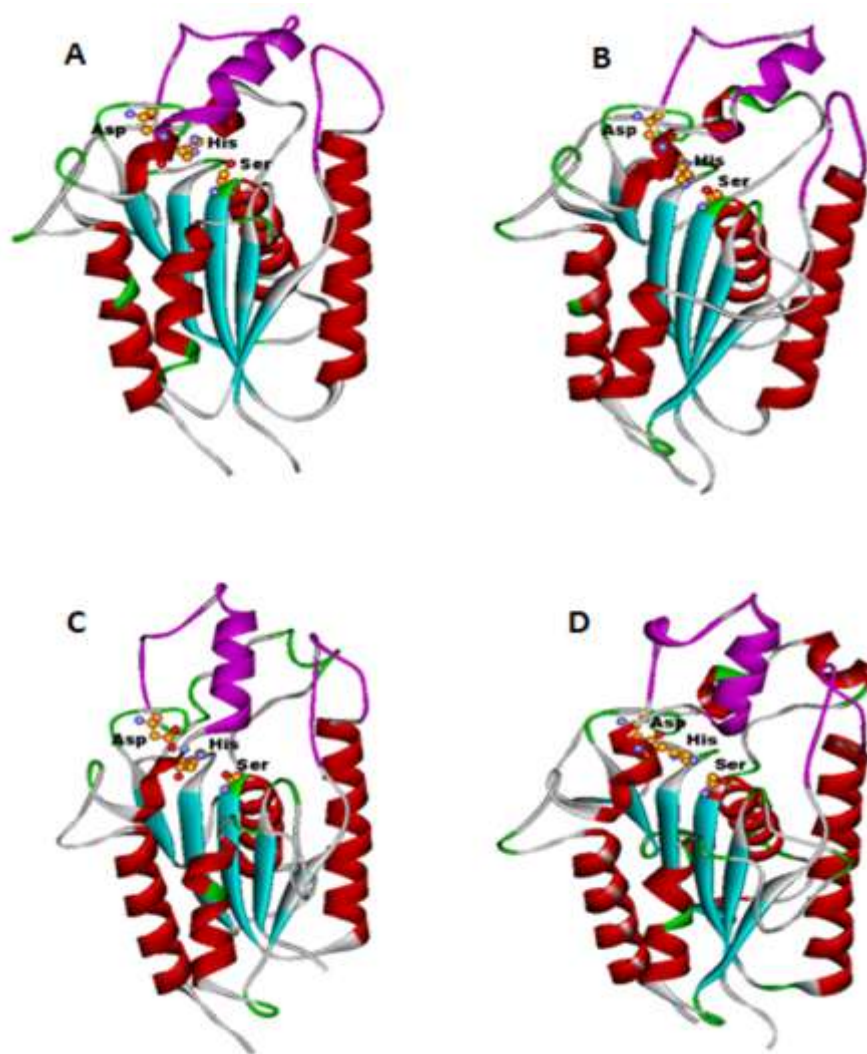
The sequence alignment generated by FUGUE program was used for the 3D structure modeling of the PE-PPE domains. Since FUGUE program recognized PDB\_ID: 3AJA with highest Z-score, the structure of the query protein was built on PDB\_ID: 3AJA using MODELLER. All homology models generated exhibited an overall canonical  $\alpha/\beta$  hydrolase fold with central  $\beta$ -strands which are flanked by  $\alpha$ -helices on either side as shown in Figure 3.3. This structural fold is well defined characteristic feature of serine  $\alpha/\beta$  hydrolase architecture. Structure alignment of model Rv1430 and template PDB \_ID: 3AJA is shown in Figure 3.3.



**Figure 3.3.** Structure alignment of model Rv1430 (blue) and template PDB \_ID: 3AJA (pink)

The structural validation using PROCHECK method indicated that greater than 85% of the residues were situated in the allowed region as observed from the Ramachandran plot (Ramachandran et al., 1963) and very few residues (less than 2%) were situated in the disallowed region that indicated satisfactory geometrical quality of PE-PPE models. Verify\_3D identified that the overall scores of the constructed structures were above 85 representing a good compatibility between the 3D constructed structure and its 1D amino acid sequence. The 3D structures of few representative models are shown in Figure 3.4.

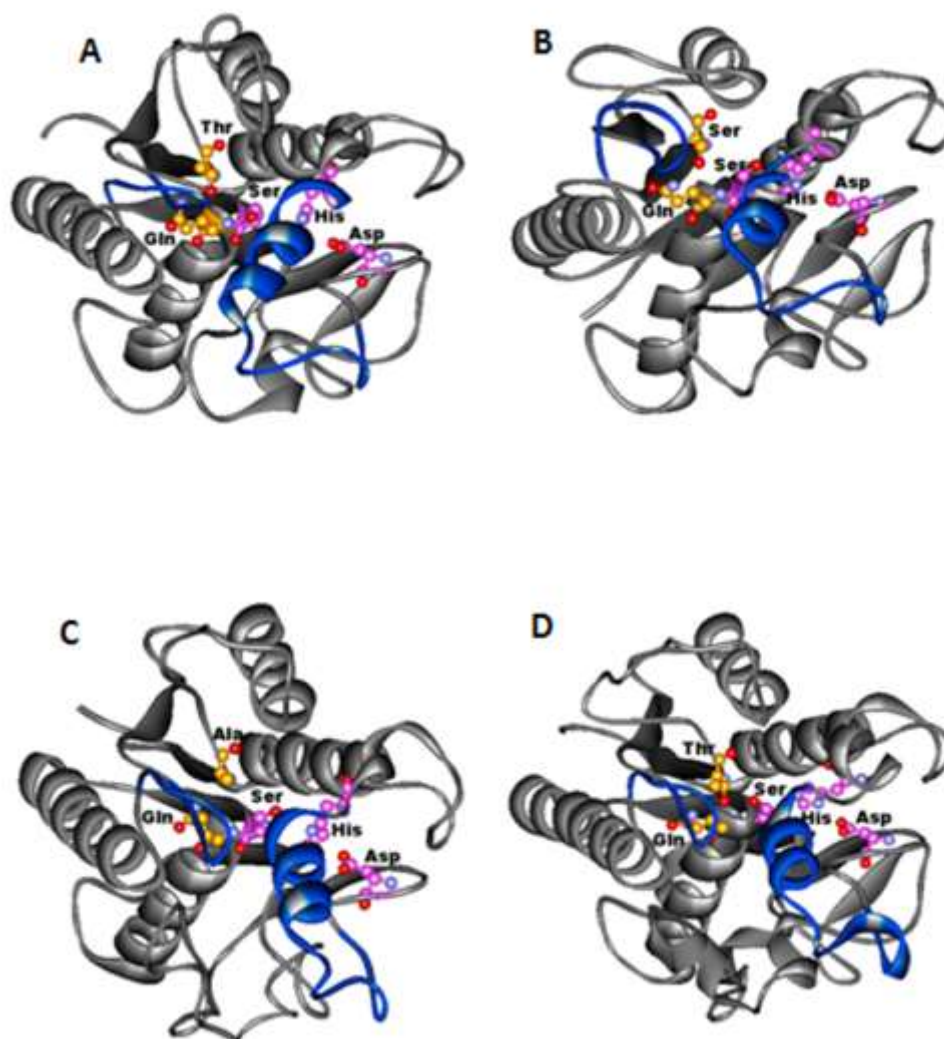
Finally the preliminary assessment of the 3D model was done by analyzing the location of amino acids in the catalytic triad in order to confirm the PE-PPE domain active site. The pentapeptide sequence motif position and the catalytic triad Ser, Asp and His are highly conserved in all the PE-PPE domain structures. The location of the catalytic triad in the constructed models is shown in Figure 4. Our results specify that for all the ten PE-PPE domain models constructed, properly positioned catalytic triad and pentapeptide sequence motif similar to the serine  $\alpha/\beta$  hydrolase fold was conserved.



**Figure 3.4.** The overall fold of the serine hydrolase models. (A) Rv1430 PE-PPE domain. (B) Rv1800 PE-PPE domain. (C) Rv1184c PE-PPE domain. (D) PDB\_ID: 3AJA used as template for homology modeling. The helices are represented in red, strands in blue, the lid insertion in pink. The sidechains of the amino acids in the catalytic triad are indicated in ball and stick

Further, the characteristic features required for serine hydrolases like lid insertion and the presence of the oxyanion hole has been identified. From a close observation of the

region from top of the structure around the active site, we observed a lid insertion region as well as solvent inaccessible catalytic residue Ser indicating the closed conformation assumed by these hydrolase structures (Figure 3.5).

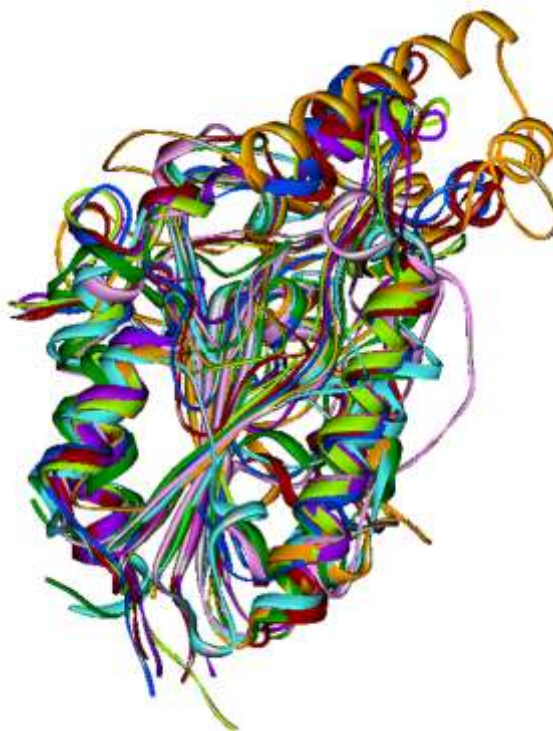


**Figure 3.5.** The top view of the serine hydrolase models indicating a closed conformation of the lid insertion on the active site. (A) Rv1430 PE-PPE domain. (B) Rv1800 PE-PPE domain. (C) Rv1184c PE-PPE domain. (D) PDB\_ID: 3AJA. The protein is represented in grey, the lid insertion in blue, the sidechains of the amino acids in the catalytic triad and oxyanion hole are indicated in ball and stick

For further understanding of this divergent family of proteins, we have generated multiple structure alignment of the PE-PPE domains of *Mtb* H37Rv and all the identified bacterial serine hydrolases by FUGUE using the program MAPSCI. These structures superimpose with root mean square deviation (RMSD) value 0.8 Å indicating high



structural similarity as shown in Figure 3.6, inspite of the high variance in the amino acid sequences.



**Figure 3.6.** The superposition of the crystal structures of bacterial serine hydrolases and the 3D structure models of the PE-PPE domain that encode serine hydrolases. PDB\_ID: 1BS9 (pink), PDB\_ID: 1CEX (cyan), PDB\_ID: 2CZQ (green), PDB\_ID: 3AJA (red), PDB\_ID: 3HC7 (orange), Rv1184c (blue), Rv1430 (purple) and Rv1800 (pale green). Only three PE-PPE homology model structures were included for the sake of clarity in structure superposition

The multiple sequence alignment obtained from structure superposition as shown in Figure 3.7, helps in understanding the relatedness between the amino acid sequences and structures. From the alignment we identified that the PE-PPE domain consists of the pentapeptide sequence motif GxSxG/S conserved in all the proteins and also we observed conserved amino acid residues Ser199, Asp276 and His302 (numbering according to Rv1430) forming a catalytic triad. Hydrophobic residues such as Leu/Tyr/Trp/Phe were observed at the second position of the conserved pentapeptide sequence motif with the exception of Rv2608, where it is substituted by Thr and the fourth position of the conserved pentapeptide sequence motif is occupied by Gln in all the proteins except Rv3539 where it is substituted by Met.

1CEX:A	RTTRDDLINGNSASC--ADVIFIYARGSTETG--N----	L----	GTLGPSIASNLE-SAF
2CZQ:A	-----ATSSAC---PQYVLINTRGTGEPQ--G----	Q----	SAGFRMTMNSQITAA-L
3HC7:A	-----SKPWLFTVHGTGQPD--P----	L---	GPGLPADTARDVL-D-I
1BS9:A	-----SC--PAIHVFGARETTASP--G----	Y-----	GSSSTVVNGVLSA-Y
3AJA:B	-----ADC--PDVMMVSIPGTWESSPTDDPFNPTQFPLSLMSNISKPLAEQ-F		
Rv0160c	-----LAAE--SPITALIMGGTNNP---LPDPE-----	Y	YVTDINKAFIQTL-F
Rv1800	-----AAA--AQTVGLVMGGSGTPIPSA---RY-----		VELANALYMS-G
Rv2608	-----NGG--PGVTALVMGGTDSL-PL----	P----	NIPLLEYAARFIT-P
Rv3539	-----TVPGASPVHAATLLPFIGRLLAARYAELNNTAIGTN-W		
Rv3822	-----DTALIVPGTAPS---P--YGPLRSLYHFNPMQPIGAN-Y		
Rv0151c	-----PMF--NQNTAIIMGGTGSPIPTP---SY-----		VNAITTLFID-P
Rv0152c	-----AMIPPPF--ANLTLFFGPTGIPL-PP----	P-----	SMLTPPIRCR-S
Rv0159c	-----LSG--NPLTALMGGTGEP---ILSDR-----		VLAIIDSAYIRPI-F
Rv1184c	-----TAK---VVYALGGARMP---G--IPWYEYTNQAGSQYFPNAKHD-L		
Rv1430	-----LGS--GGRTALILGSTGTP---RPPFD-----		YMQQVYDRIAPH-Y
	110	120	130 140

1CEX:A	GKD-GVWIQGVGGAYRATLG-----	DNALPRG--	TSSAAIREMLGL
2CZQ:A	S---GGTIYNTV--YTADF-----	SQ-N-----	SA-AGTADIIRR
3HC7:A	----YRWQPIG-NYPAAAF-----	PM-W---	PSVE-KGVAELILQ
1BS9:A	P-G--STAEAIN--YPACGG-----	QSSCGG--	ASY-S---SSVA-QGIAAVASA
3AJA:B	G-PDRLQVYTTP--YTAQFH-N--P-FAADK---	Q---MSY-N---	DSRA-EGMRTTVKA
Rv0160c	P-G--AVSQGLF--TPEQFW-PVTP-D--LG--N---	LTF-N---	QSVT-EGVALNNTA
Rv1800	S-VPGVIAQALF--TPQGLY-P--V-V-VIK---	N---LTF-D---	SSVA-QGAVILESA
Rv2608	V-HPGYTATFLE--TPSQFF-P--F-T-GLN---	S---LTY-D---	VSA-QGVTNLHTA
Rv3539	F-P-GTTPEVVS--YPATIG-V--L-SGSLG---	A---VDA-N---	QSIA-IGQQMLHNE
Rv3822	Y-NPTATRHVVS--YPGSEW-P--V-T-GLN---	S---PTV-G---	SSVS-AGTNNLDAA
Rv0151c	V-VSNPVVKALV--TPEELY-P--I-T-GVK---	S---LPF-Q---	TSVQ-LGLQILDGA
Rv0152c	V-R--RALQAVF--TPEELY-P--L-T-GVR---	S---LVL-N---	TSVE-EGLTILHDA
Rv0159c	G-PNNPVAQYT---P-EQWWP-----	FIGNLSL-D---	QSIA-QGVTLNNG
Rv1184c	I-DYPAGAAFSW--WPTMLL-P--PGS-HQD---	N---MTV-G---	VAVK-DGTNSLDNA
Rv1430	L-G--YAFSGLY--TPAQFQ-P--W-T-GIP---	S---LTY-D---	QSV-EGAGYLHTA
	150	160	170 180

		*#	
1CEX:A	FQQANTKCPDATLIAGGYSQGAALAAASIE-D-----		LDSAI-RDK
2CZQ:A	INSGLAANPNVCYILOGYSQGAATVVALQ-Q-----		LGT-SG---AA-FNA
3HC7:A	IELKLDADPYADFAMAGYSQGAIVVGQVLKHH-----		ILP-PTGRLHRF-LHR
1BS9:A	VNSFNSQCPSTKIVLVGYSQGEIMDVALC-GGGDPNQGYTNTAV-----		QLS-SSAVNM
3AJA:B	MTDMNDRCPPLTSYVIAGFSQGAVIAGDIAS-D-----		IGN-G-RGP-VD-EDL
Rv0160c	VNNQLA--LDNKVVAFGYSQSATIINNYIN-S-----		LMA-M-GSP-NP-D--
Rv1800	IRQQIA--AGNNVTVEFGYSQSATISSLVMA-N-----		LAA-S-ADP-PS-PDE
Rv2608	IMAQLA--AGNEVVVEGTSQSATIATFEMR-Y-----		LQSLP-AHL-RP-GLD
Rv3539	ILAATA--SGQPVTVAGLSMGSVIDRELA-Y-----		LAI-D-PNA-PP-SSA
Rv3822	IRSTD----GPIFVAGLSQCTLVLDREQA-R-----		LAN-D-PTA-PP-PGQ
Rv0151c	IWEQIN--AGNHVTVEFGYSQSAVIASLEMQ-H-----		LIS-L-GPN-AP-SPS
Rv0152c	IMVELA--TTGNAVTVFGWSQSAIIASLEMQ-R-----		FTA-M-GGA-AP-SAS
Rv0159c	INAELO--NGHDVVVEFGYSQSAAVATNEIR-A-----		LMALP-PGQ-AP-DPS
Rv1184c	IHHGT----DPAAAVGLSQGSLVLDQEQAR-----		LAN-D-PTA-PA-PDK
Rv1430	IMQQVA--AGNDVVVLGFSQGSASVATLEMR-H-----		LASLP-AGV-AP-SPD
	190	200	210 220



```

1CEX:A      IAGTVLFGYTKNLQNR-GR-----
2CZQ:A      VKGVFLIGNPDH--KSG LTC-NVDSNG-----GTTTRN-VNGLSV--A-----
3HC7:A      LKKVIFWGNPMR--QKG--F-AH---SDEWIHPV-AA-PDTLGIL---E-----
1BS9:A      VKAAIFMGDPMF--RAGLSY-EV-----G-TC-A-AGGFDQ-----
3AJA:B      VLGVTLIADGRR--QMGVGQ-DV---G-----P-NP-A-GQGAETTLHE-VPALSALGL
Rv0160c     DISFVMIGSGNN--PVGGLL-AR---F-----P-GF-Y-IPFLDVPFNG-A-----TP
Rv1800      L-SFTLIGNPNN--PNGGVA-TR---F-----P-GI-S-FPSLGV TATG-A-----
Rv2608      ELSFTLTGNPNR--PDGGIL-TR---F-----G-FS-I-PQLGFTLSGA-T-----
Rv3539      L-TFVELAGP-----ERGLAQ-TYLPVGT TI
Rv3822      L-TFIKAGDPNN--LLWRAF-RP---G-----T-HV-P-IIDYTV PAPA-E-----
Rv0151c     QLNFI LIGNEMN--PNGGIL-AR---I-----P-GL-N-VTTLGLPFY G-A-----
Rv0152c     DLNFVLVGNEMN--PNGGML-AR---F-----P-DL-T-LPTLDLT FYG-A-----T
Rv0159c     RLAFTLIGNINN--PNGGVL-ER---Y-----V-GL-Y-LPFLDMSFNG-A-----T
Rv1184c     L-QFTTFGDPTG--RHA FGAS-----FLARIFPPGS--HIPIPF
Rv1430      QLSFVLLGNPNN--PNGGIL-AR---F-----P-GL-Y-LQSLGLTFNG-A-----

```

230 240 250 260

\*

```

1CEX:A      -----I-PNY--PADRTKVFCNTGDLVCTG--S-----
2CZQ:A      -YQG--S-----VPSGW--VSKTLDVCAYGDGVC DT--AHGF-----
3HC7:A      ---D--R-----L-ENLEQYGF EVRDYAHDGDMYASI---K-EDDLHEY EVAIGRIVMKA
1BS9:A      -----RPAGFS---C-PSAAKIKSYCDASDPYCCNGSN-----
3AJA:B      TMTG-PR--PGG---FGALDNRTNQC SGDLICSA---P---E--Q-AFS-----
Rv0160c     ANSP--Y-----P-----THIYTAQYDGI AHA---P---Q--F-PLRI-----
Rv1800      TPHN--L-----Y---P-----TKIYTIEYDGVADF---P---R--Y-PLNF-----
Rv2608      PADA--Y-----P-----TVDYAFQYDGVNDF---P---K--Y-PLNV-----
Rv3539      PIAG-YT--VGN---APESQYNTSVVYSQYDIWADP---P---D--R-PWN-----
Rv3822      -SQY--D-----T-----INIVGQYDI FSDP---P---N--R-PGNL-----
Rv0151c     -TPD--N-----P---Y-----PTTTYTLEYDGFADF---P---R--Y-PLNV-----
Rv0152c     PSDT--I-----Y---P-----TAIYTLEYDGFADF---S---R--Y-PLNF-----
Rv0159c     PPDS--P-----Y---Q-----TYMYTGQYDGYAHN---P---Q--Y-PLNI-----
Rv1184c     IEYTMPQ--QVD---S---QYDTNHVV TAYDGFSDF---P---D--R-PDN-----
Rv1430      -TPD--T-----D---Y-----ATTIYTTQYDGFADF---P---K--Y-PLNI-----

```

270 280

\*

```

1CEX:A      -----LI-V--A-----APHLAYG---
2CZQ:A      -----G--I--N-----AQHLSYPSD-
3HC7:A      SGFIGGRDSVVAQLIELGQRPITEGIALAGAIIDA---LTFFARS RMGDKWPH-LY----
1BS9:A      -----A-----ATHQGYGS--
3AJA:B      -----V-FNLPKTL-ETLSGSAA-----GPVHALYNT PQ
Rv0160c     -----L--SDINAF-MGY-F--Y-----VHNTYPELM
Rv1800      -----V--STLNAI-AGT-Y--Y-----VHSNYFILT
Rv2608      -----F--ATANAI-AGI-L--F-----LHSGLI ALP
Rv3539      -----L-LAGANAL-MGA-A--Y-----FHDLTAYAA
Rv3822      -----L--ADLNAI-AAG-G--Y-----YGHSA TAFSD
Rv0151c     -----L--SDINAV-FGI-L--T-----VHTTYADLT
Rv0152c     -----I--SDLNAV-AGI-T--F-----VHTKYLDLT
Rv0159c     -----L--SDLNAF-MGI-R--W-----VHNAYPFTA
Rv1184c     -----L-LAVANAA-IGA-A--I-----AHTPIGFTG
Rv1430      -----L--ADVNAL-LGI-Y--Y-----SHSLYYGLT

```

290 300

```

1CEX:A      -----P----DARGPAPEFLIEKVRA--VRGS-----
2CZQ:A      -----QG----VQ-TMGYKFAVNKLGGSA-----
3HC7:A      -----N----R-YPAVEFLRQ-----
1BS9:A      -----E----YG-SQALAFVKSKL-G-----
3AJA:B      FWV-ENG----QT----AT-QWTLEWARNLVEN--AP--HP---
Rv0160c     ATQ-VDN----AVPLPTSP-GYTGNTQYYMFLT--QD--LP---
Rv1800      PEQ-IDA----AV----PL-TNTVGPTMTQYYI--IR--TENLP
Rv2608      PDL-ASG----VV----QP-VSSPDVLTTYILL--PS--QDLP-
Rv3539      PQQ-GIE----IA----AV-TSSLGGTTTTYMI--PS--GYS--
Rv3822      PAR-VAPRDITTT-----TN-SLGATTTTTYFIRT--DQ--LP---
Rv0151c     PAQ-IAS----AT----QL-PTQGTTSENTYYII--ET--EHL-
Rv0152c     PAQ-VEG----ATKLPTSP-GYTGVTDDYYIIRT--EN--RP---
Rv0159c     AEV-ANA----VP----LP-TSPGYTGNTTHYYM--FL--TQDLP
Rv1184c     PGDVPP----QN----IR-TTVNSRGATTTTTY--LV--PVN--
Rv1430      PEQ-VAS----GI----VL-PVSSPDNTTTYIL--LP--NED--
310              320              330

```

**Figure 3.7:** Multiple sequence alignment of the PE-PPE domain that encodes a serine hydrolase in *Mtb* H37Rv and some bacterial serine hydrolases of known structure, generated using the program MAPSCI. The conserved pentapeptide sequence motif is represented in square box. The amino acid residues in the catalytic triad are represented in \* and the residues in the oxyanion hole are represented in #

Most of the classical lipases such as PDB\_IDs: 2Z8X, 5TGL, 2VEO, 4TGL comprises a lid insertion region above the active site that makes the active site a closed conformation. This lid region is displaced during interfacial activation that allows the substrate to be activated by opening up and revealing the catalytic triad. There are many reports that explain these conformational states of the lipases (Cherukuvada et al., 2005; Grochulski et al., 1994). Another attractive feature of the serine hydrolases is the existence of an oxyanion hole that represents a part of the active site. In bacterial lipases such as PDB\_IDs: 1IVN, 1CRL the position of an oxyanion hole has been described. Based on the structure alignment between the crystal structure and the built homology structure models of the PE-PPE domain, we observed the position of the oxyanion hole in all our constructed model structures. For instance, the oxyanion hole in Rv1430 is formed by the residues Thr121 and Gln200 as indicated in Figure 3.5. Oxyanion hole is a tetrahedral intermediate formed by the amino acids that is stabilized by hydrogen bonding between the main chain N-H and the hydrolyzed substrate thus stabilizing the negative charge on the tetrahedral intermediate occurred by the nucleophilic attack of the catalytic Ser residue during activation (De Simone et al., 2004).

From these studies, using computational aids we observed that the PE-PPE domain is well characterized with a serine  $\alpha/\beta$  hydrolase fold which would have a typical esterase, lipase or cutinase activity. Cutinases are serine hydrolases that act on carboxyl ester bond cleaving cutin, a complex waxy glycolipid polymer consisting of hydroxy fatty

acids. Some examples of cutinases are from *Streptomyces scabies* (Lin TS, 1980), *Pseudomonas putida* (Sebastian et al., 1987; Sebastian and Kolattukudy, 1988) and *Thermobifida fusca* (Chen et al., 2008; Fett WF, 1999). Lipases belong to lipolytic hydrolases family that help the hydrolysis of carboxyl ester bonds of water insoluble substrates like mono-, di- and tri-glycerides to release fatty acids and alcohols in aqueous solutions. Some examples of bacterial lipases are from *Staphylococcus xylosus* (Brod et al., 2010a) and *Acinetobacter baumannii* BD5 (Park et al., 2009). While the lipases catalyze the hydrolysis of long chain acylglycerols, esterases perform their role by hydrolysing esters of short chain fatty acids. Few bacterial esterases are from *Lactobacillus plantarum* (Brod et al., 2010b) and *Thermus scotoductus* (Du Plessis et al., 2010). We therefore believe that high similarity of these structures within bacterial cutinases, esterases and lipases was observed despite of high divergence in the sequence homology.

From Figure 3.1, the occurrence of PE-PPE domain towards the C-terminus is an example of gene fusion. In the present context of *Mtb* strain H37Rv, the fusion of PE domain and the PE-PPE domain in some of the PE family proteins or else PPE domain and PE-PPE domain in some of the PPE family proteins is a classic example of gene fusion. One of the functions of the N-terminal PE or PPE domains possibly is to translocate the protein to the site of action and the role of the PE-PPE domain is to function as serine hydrolase.

Further, the ten distinct genes possessing the PE-PPE domain are present as paralogs in *Mtb* strain H37Rv this is an indication of gene duplication. During the evolution stages of an organism to survive in the selective pressure of the open environment or host, the preexisting ancestral genes undergo gene duplication that is followed by mutations as well as rearrangements to accommodate the new requirements rather than the synthesis of new genes. Hence, we consider that these ten genes comprising the PE-PPE domain would have closely related serine hydrolase function with distinct activities. In addition to this, we observed that the distribution of PE-PPE domain regions in proteins from a variety of mycobacterial genomes was not uniform (Table 3.1) and these variations may lead to the differences in pathogenesis as well as the virulence of the organism along with the nature of host.

Overall, the FUGUE method recognized template 3D structures belonging to cutinase, esterase and lipase family of proteins as likely structural folds of the PE-PPE domain. These enzyme structures belong to members of serine hydrolase superfamily

evolved from a common ancestor. The highly conserved active site residues forming the catalytic triad are suitably positioned to assist the hydrolysis activity. From structure analysis of the PE-PPE domain, the lid insertion and oxyanion hole are identified that regulate the enzyme activity stabilizing the intermediate formed during catalysis.

### 3.4 Conclusion

The work done in this chapter provided the structure and function of the PE-PPE domain from *mycobacteria*. Our analysis showed that the PE-PPE domain belongs to the serine  $\alpha/\beta$  hydrolase family of proteins. Serine hydrolases play an important role in the production of waxy lipid rich cell wall during dormancy and reactivation of *mycobacteria*. Therefore exploring these proteins may enhance the contribution towards drug targets and vaccine design. Herewith biochemical characterization would prove the defined function of these proteins.



## **Chapter 4**

**The PE16 (Rv1430) of *Mycobacterium tuberculosis*  
is an Esterase Belonging to Serine Hydrolase  
Superfamily of Proteins**





## 4.1 Introduction

Among the ~ 4000 genes in the entire *Mtb* H37Rv strain genome, 250 genes are a part of the fatty acid metabolism of the bacterium (Cole et al., 1998). Also the complex cell wall composition of *Mtb* that comprises peptidoglycan and lipids deserves special attention as it is unique cell wall structure among prokaryotes that stands for a determinant factor in the virulence of the bacterium (Brennan, 2003). Apart from these, two unique gene families, PE and PPE were identified in *Mtb* H37Rv genome comprising 10% of the coding regions (Cole et al., 1998).

Since their discovery, various studies have been performed that specify numerous physiological roles of these PE and PPE families as revealed below. PE/PPE transcriptomics revealed that these genes are expressed under diverse conditions with 128/169 PE and PPE genes found to be differentially regulated (Voskuil et al., 2004). Some PE proteins for example Rv0746, Rv1759c and Rv1818c play a key role in immune evasion of the host and antigenic variation (Banu et al., 2002; Brennan and Delogu, 2002; Cole et al., 1998; Delogu and Brennan, 2001). Some PE and PPE family of proteins like Rv1818c, Rv1917c and Rv3873 are found to be associated with the cell wall (Brennan, 2003; Delogu et al., 2004).

The PE family protein, Rv1818c was detected to be a surface exposed protein and influences the cellular innate response and macrophage function (Brennan et al., 2001). PPE proteins such as Rv2430c and Rv0256c induce a strong B-cell response (Abraham et al., 2014; Choudhary et al., 2003). Some PE and PPE family proteins such as, Rv1787, Rv2430c and Rv3018c are associated with virulence (Choudhary et al., 2003; Li et al., 2005b; Ramakrishnan et al., 2000). Some PE/PPE genes existence is identified as gene pairs that are coexpressed, co-regulated and interact functionally (Tundup et al., 2006). Several of these PE and PPE genes are involved during starvation and upregulated essentially during macrophage infection and within host granulomas signifying their role in virulence and pathogenesis of *mycobacteria* (Mohareer et al., 2011; Voskuil et al., 2004). There are several ORFs in PE/PPE family gene clusters that needs to be annotated with respect to their biochemical activity with few exceptions like Rv3097c (LipYtub). The C-terminal portion of Rv3097c is homologous to the hormone-sensitive lipase family protein and is defined by the characteristic conserved GDSAG sequence motif and was shown to be important for extracellular lipid hydrolysis (Cascioferro et al., 2007; Daleke et al., 2011; Deb et al., 2006; Mishra et al., 2008). Another interesting fact is that these

multigene PE and PPE families are unique to *Mycobacteria* and have evolutionarily, abundantly expanded most preferentially in pathogenic *Mycobacteria* with unknown functions (Gey van Pittius et al., 2006) and are absent in humans, that make them perfect for diagnosis, treatment and in new anti-TB drug development.

It is well known fact that the PE and PPE gene families are organized in a specific pattern of operonic arrangement. In this pattern, PE is followed by a PPE gene which are separated by nearly 90 bp and scattered all over the genome (Tundup et al., 2006). Some PE and PPE proteins such as Rv2430c, Rv2431c, Rv3018c, Rv3812 and Rv3873 mostly form inclusion bodies when they are overexpressed in *E. coli*. Some proteins like PE (Rv2431c) and PPE (Rv2430c) interact with each other and solubilize only when coexpressed (Chaitra et al., 2008; Choudhary et al., 2004; Okkels et al., 2003; Tundup et al., 2006).

In the previous work by Adindla et al, it was shown that some PE, PPE and hypothetical proteins of *Mycobacteria* comprises a 225 amino acid conserved domain at the C-terminus (Adindla and Guruprasad, 2003) and this was named as the PE-PPE domain (Pfam ID: PF08237). The PE-PPE domain region was observed among ten proteins of *Mtb* H37Rv strain, they are Rv0151c, Rv0152c, Rv0159c, Rv0160c, Rv1430, Rv1800, Rv2608, Rv3539, Rv1184c and Rv3822. Hitherto from our work on PE and PPE proteins using the fold prediction computational tools, we recognized that the PE-PPE domain region (Pfam: PF08237) has a typical “serine  $\alpha/\beta$  hydrolase” fold with the characteristic GxSxG/S pentapeptide sequence motif and conserved Ser, Asp and His catalytic region that is specific to lipases, esterases and cutinases as described in chapter 3.

Serine hydrolases play an important role in lipid metabolism. Lipid metabolism is one of the major pathways in *Mycobacteria*. Since the cell wall structure of *Mycobacteria* is extremely complex and unique in prokaryotes, it remains chief determinant of virulence in the bacterium. Almost 60% of the mycobacterial cell wall is made up of lipids. The lipid content of *Mtb* cell wall comprises of phthiocerol dimycocerosates, mycolic acid, glycolipids, polyketides and glycans, and are involved in the pathogenicity and virulence of *Mtb* (Daffe and Draper, 1998; Kremer et al., 2005; Minnikin et al., 2002). Further, another crucial function of lipids in *Mtb* is that they offer carbon source, support the growth of the bacteria at the time of chronic infection phase (Garton et al., 2002; Neyrolles et al., 2006). The required energy for future is preserved during the elongated periods of bacterial dormancy that serve as a storage house of fatty acids necessary for

membrane lipid formation (Daniel et al., 2004). This involves various serine hydrolases in *Mtb* that are essential for the bacterial survival and allow them to become accustomed to the surroundings provided by the host cells (Cotes et al., 2007). The achievement of *Mtb* survival in the human host is its distinctively complex lipid rich cell wall, therefore the proteins in the cell wall synthesis pathways are essential target areas for drug development (Barry, 2001; Brennan, 2003; Kusner, 2005; McKinney et al., 2000; Parker et al., 2009).

In order to prove our prediction that the PE-PPE domain region is indeed a serine hydrolase, we have carried out experimental studies. We have cloned both the full-length Rv1430 gene and its PE-PPE domain region into pET-28a vector, overexpressed the recombinant proteins in *E. coli*, purified the proteins to apparent homogeneity and carried out biochemical characteristic assays to analyze their enzyme activity.

## 4.2 Materials and Methods

### 4.2.1 Reagents Used

*Mtb* H37Rv genomic DNA was obtained from Blue Peter Research Centre, LEPR, Hyderabad, India, that was gifted by Colorado State University (Fort Collins, CO, USA). The Jumpstart accutag DNA polymerase was purchased from Fermentas and the pET-28a vector was purchased from Novagen. The *E. coli* DH5 $\alpha$  cells used for plasmid preparations during cloning and the *E. coli* BL21 cells used for protein expression were grown in Luria-Bertani (LB) broth or on LB agar. The isopropyl  $\beta$ -D -1-thiogalactopyranoside (IPTG) and antibiotic kanamycin were purchased from HiMedia and supplemented as necessary. The substrates, *p*-nitrophenyl-acetate (pNPC2), *p*-nitrophenyl-butyrate (pNPC4), *p*-nitrophenyl-caprylate (pNPC8), *p*-nitrophenyl-caprate (pNPC10), *p*-nitrophenyl-laurate (pNPC12), *p*-nitrophenyl-myristate (pNPC14) and *p*-nitrophenyl-palmitate (pNPC16), Triton X-100, Tween-20, gum arabic and phenylmethylsulfonyl fluoride (PMSF) were purchased from Sigma chemicals. The Cobalt metal affinity resin was purchased from Clontech and the substrate, *p*-nitrophenyl-caproate (pNPC6) was purchased from TCI chemicals, India.

### 4.2.2 Cloning, Expression and Purification of Recombinant Rv1430 and its PE-PPE Domain

The full-length Rv1430 (PE16 gene) about 1601 bp and its PE-PPE domain from 430-1107 bp about 678 bp of *Mtb* were PCR amplified from *Mtb* H37Rv genomic DNA using proof-reading thermostable Taq DNA polymerase. Primers used were

**Forward Primer** 5'ATGTCGTTTCGTTTTTCGCGGTGCCA3'

**Reverse Primer** 5'GGGGGCCTATACCGGAAAATCCTG3'

for the Rv1430.

**Forward Primer** 5'CACTACTTGGGCTATGCGTTTTCC3'

**Reverse Primer** 5'TCAGTCATAACCCAATTCGATGATCGC3'

for the PE-PPE domain.

*Bam*HI and *Hind*III were selected as the restriction sites. The PCR amplified products were directly cloned into the T7 promoter-based expression vector pET-28a which provided a hexa histidine tag by ligation. These recombinant plasmids (ligation product) were transformed into the extensively used *E. coli* DH5 $\alpha$  cells. The existence of

the inserts was confirmed by double digestion of the plasmids using *Bam*HI and *Hind*III restriction enzymes, followed by DNA sequencing.

For optimal conditions of protein expression, *E. coli* Rosetta gami cells were used to transform the pET-28a vector carrying insert Rv1430 gene and *E. coli* BL21 (DE3) CodonPlus-RIL cells were used to transform the pET-28a vector carrying the insert PE-PPE domain region. The *E. coli* Rosetta gami and *E. coli* BL21 (DE3) CodonPlus-RIL cells were grown in LB medium.

We allowed overnight expression of recombinant plasmids of *E. coli* BL-21 (DE3) and *E. coli* Rosetta gami and then these were diluted 10 times using fresh LB broth containing antibiotic, kanamycin. This mixture was allowed to grow at 37 °C for 3-4 h with vigorous shaking until the  $\lambda_{600}$  nm had reached with an absorbance of 0.6–0.8. A part of uninduced culture was reserved as a separate aliquot for control. The expression of both proteins were induced with IPTG (1 mM final concentration), the temperature was maintained at 18 °C and the cells were allowed to grow overnight. After 16 h of incubation, the cells were harvested by centrifugation at 6500 *g* for 20 min at 4 °C. After harvestation, the supernatant was removed and the pellet was resuspended in ice-cold (prechilled to 4 °C) lysis buffer (50 mM Tris/HCl at pH 7.4, 150 mM of mM NaCl, 0.25 mg/mL of lysozyme; 30 mL/litre of initial culture). For lysis of the cultured cells the mixture was incubated at 4 °C for 30 min, sonicated for 10 cycles at 4 °C, each cycle comprising of 30 s on and 60 s off time. Harvested the cells by centrifugation at 17000 *g* at 4 °C. The cell debris and the supernatant fractions were carefully separated. These fractions were used for the SDS-PAGE analysis and it was noticed that extremely low quantities of the protein was present in the supernatant and a large amount of it was present in the cell debris as inclusion bodies. In order to separate the protein from the inclusion bodies, cell debris were washed thrice in buffer A containing 50 mM Tris/HCl at pH 7.4, 150 of mM NaCl and 0.1% Triton X-100, followed by sonication for 10 cycles at 4 °C, each cycle comprising of 30 s on and 60 s off time and centrifuged at 17000 *g*. Later, the pellet was resuspended in solubilization buffer A containing 8 M urea, then incubated at room temperature for 2 h which was followed by centrifugation at 20,000 *g* by using a Sorvall Biofuge Stratos centrifuge (Thermo Scientific).

After solubilization, to proceed for immobilized metal ion affinity chromatography (IMAC) purification, the supernatant comprising desired proteins was applied on a Cobalt affinity column that was pre-equilibrated with buffer A containing 8 M urea. The column was first washed carefully with buffer A containing 8 M urea. Then

with wash buffer containing buffer A and 8 M urea along with 30 mM imidazole. The proteins were eluted with elution buffer containing buffer A, 8 M urea and 150 mM imidazole.

The proteins were finally further purified using gel filtration column chromatography technique to 99% purity by using Sephadex G-50 (Sigma) XK 16/100 column with buffer containing 10 mM Tris/HCl, 150 mM NaCl at pH 7.0 with a flow rate of 0.5 mL/min. The fractions obtained containing the desired proteins were checked on SDS-PAGE, then pooled and dialyzed using 20 mM Tris/HCl, pH 8.0 and 150 mM of NaCl. These purified folded proteins were used for enzyme activity studies. The concentrations of the proteins were evaluated by Bradford method (Bradford, 1976). Both proteins Rv1430 and its PE-PPE domain were purified and assayed with same procedures.

#### **4.2.3 Cloning, Expression and Purification of Recombinant Ser199Ala Mutant Rv1430 PE-PPE Domain**

To confirm the active site we have carried out the site directed mutagenesis of one of the active site residues of Rv1430 PE-PPE domain. For this purpose internal primers were designed and purchased from Bioserve (Bioserve Biotechnologies Pvt. Ltd, Hyderabad, India). PCR amplification of the PE-PPE domain with desired mutation was performed using overlap PCR method (Ho et al., 1989). Primers used were

**Upstream Primer** 5'GGTTTCGCGCAGGGCGCGTCGGTCGCC3'

**Downstream Primer** 5'GCCCTGCGCGAAACCCAACACCACAAC3', respectively.

The primers were designed to substitute the active site Ser199 with Ala residue and cloned with the restriction sites *Bam*HI and *Hind*III into the expression pET-28a vector containing T7 promoter and His-tag. The mutation of the Ser199 with Ala was confirmed by nucleotide sequencing from Macrogen sequencing service (Macrogen Inc., Seoul, Korea). After confirming the mutation, the vector containing the desired insert was transformed into the expression cells *E. coli* BL21 (DE3) CodonPlus-RIL. Similar to the wild proteins the expression and purification of Ser199Ala mutant Rv1430 PE-PPE domain was carried out.

#### **4.2.4 Western Blot**

The protein expressions of Rv1430 and its PE-PPE domain was confirmed by Western blot analysis using Amersham Wet blot apparatus TE 22. We have used anti-His

monoclonal antibody with 1:2000 dilutions as primary antibody and 1:2000 dilutions of peroxidase-conjugated goat anti-mouse IgG was used as secondary antibody. The protein samples were loaded and transferred to ethyl alcohol activated PVDF membrane (Millipore). Finally, the blot was developed using Supersignal w pico chemiluminiscent substrates from GeneX and the protein expression signals were visualized on a versa doc V5 (Bio-Rad).

#### 4.2.5 Enzymatic Assay

The enzymatic assays of both the proteins; full-length Rv1430 and its PE-PPE domain were carried out using different substrates, *p*-Nitrophenyl esters of different lengths like *p*-Nitrophenyl acetate (pNPC2), *p*-Nitrophenyl butyrate (pNPC4), *p*-Nitrophenyl caproate (pNPC6), *p*-Nitrophenyl caprylate (pNPC8), *p*-Nitrophenyl caprate (pNPC10), *p*-Nitrophenyl laurate (pNPC12), *p*-Nitrophenyl myristate (pNPC14) and *p*-Nitrophenyl palmitate (pNPC16). The substrates were prepared by dissolving in 10 mL isopropyl alcohol to a final concentration of 10 mM for working stock. The standard reaction mixture for enzymatic assay consists of (1 mL) 1 mM of *p*-nitrophenylester, 10 mM of phosphate buffer (pH 7.0) and 30 µg that is 11.63 µM of purified proteins. In some cases of esters like pNPC8 and longer carbon chain lengths, assay solution consists of 10 mM sodium phosphate buffer (pH 7.0) along with surfactant 0.4% Triton X-100 and 0.1% gum arabic to stabilize the enzyme substrate emulsion. Similar procedures were followed by Schue et al., (Schue et al., 2010) and West et al., (West et al., 2009) in the enzymatic assays of mycobacterial proteins Rv1984c, Rv3452 and also in the case of a family of seven cutinase-like proteins CULPs. For reference, all the components of the above mentioned standard reaction mixture except protein were taken in the cuvette. Since the enzymatic hydrolysis reaction product is *p*-nitrophenol for all the substrates, the absorbance of *p*-nitrophenol was used for quantification of the enzymatic assay by measuring the absorbance at 410 nm with a UV-1800 Shimadzu-UV spectrophotometer. For this purpose the molar extinction coefficients of *p*-nitrophenol was determined prior to the enzymatic reaction measurements under each condition. The enzyme activity was analyzed by measuring the initial reaction rate of hydrolysis of various substrates of *p*-nitrophenyl esters. One unit of enzyme activity is the amount of enzyme/protein essential to produce 1 µmol *p*-nitrophenol/min from *p*-nitrophenylester substrates. The enzyme activity of Ser199Ala mutated PE-PPE domain was examined on the most active *p*-nitrophenylester, pNPC6.

The lipolytic activity of the protein Rv1430 and its PE-PPE domain was analyzed using Tween-20, a polyoxyethylene sorbitan monolaurate substrate as reported previously (Pratt et al., 2000; West et al., 2009). The reaction mixture consists of 33 mM CaCl<sub>2</sub> and 0.33% Tween-20 in buffer containing 50 mM Tris–HCl for pH 7.0 and pH 8.0 and 50 mM MES for pH 6.0 to a final volume of 1 mL. The reaction mixture for each assay consists of 30 µg of the purified protein. The assay was performed at different pH by incubating at 37 °C for 1 h and the turbidity formed was measured at 405 nm.

For the cutinase activity, we have followed the protocol as reported earlier (Chen et al., 2008) and used apple cutin was used as a substrate. We have separated apple cutin from mature apples (Walton and Kolattukudy, 1972; West et al., 2009). The typical assay mixture consists of 1 mg of enzyme and substrate 100 mg of apple cutin and buffer containing 50 mM Tris–HCl at pH 7.5 to a final volume of 10 mL. The reaction mixture was shaken at 37 °C at 125 rpm for 18 h. After incubation, the left over cutin was separated by centrifugation. After centrifugation, the supernatant was carefully acidified using acetic acid and the produced monomers of cutin were extracted with chloroform. The organic soluble matter was separated and dried using a stream of nitrogen. Then the dried cutin monomers were transformed to their subsequent methyl esters and then these were silylated by bis-(trimethylsilyl) trifluoroacetamide (BSTFA). The silylated methyl esters were then dissolved in hexane solvent and measured by GC/MS QP2010 (Shimadzu, Japan) with temperature control program as follows: 125 °C for 5 min, 4 °C /min to 250 °C, 250 °C for 15 min.

#### **4.2.6 Circular Dichroism (CD) Spectroscopy Analysis of Rv1430 and its PE-PPE Domain**

To evaluate the secondary structure and proper folding of the proteins we have carried out the CD spectroscopy analysis of both Rv1430 and its PE-PPE domain. The CD spectra was recorded at 25 °C in 10 mM phosphate buffer at pH 7.0 using 0.125 mg/mL of Rv1430 and 0.15 mg/mL of PE-PPE domain on a Jasco J-810 spectropolarimeter in a quartz cell with a path length of 1 mm. Three scans were performed and accumulated at a scan speed of 50 nm/ min. The data was collected for every nm from 195 to 260 nm. To estimate the secondary structural elements of both the proteins CDNN 2.1 (Bohm et al., 1992) software was used. The CD data is measured in mean ellipticity values per each residue ( $[\theta]$ ).



#### **4.2.7 Enzyme Inhibition**

The inhibitory action of the chemical modifiers that are specific to certain amino acids in the catalytic active site region of the enzymes like phenylmethylsulfonyl (PMSF) to the amino acid residue Ser in this case was examined. The enzyme having 0.84  $\mu$ M concentration was incubated with different concentrations of PMSF such as from 0.5 mM to 10 mM for 15 min at 30 °C prior to the reaction. After incubation the reaction was started by the addition of the substrate that is most active; pNPC6 and the enzyme activity was recorded using the standard assay protocol as described above.

#### **4.2.8 Effect of pH and Temperature**

To study the effect of pH and temperature on the protein, the enzyme activities were measured as described above on the most suitable substrate, pNPC6. The reaction was carried out with 1 mM of pNPC6 at various pH (pH 4.0–pH 8.0) and at various temperatures (25 °C –45 °C). The buffers (10 mM) used were: sodium acetate for pH 4.0 and 5.0, sodium phosphate for pH 6.0 and, 7.0, Tris-HCl for pH 8.0 and 9.0 and glycine/NaOH for pH 10.0 and 11.0.

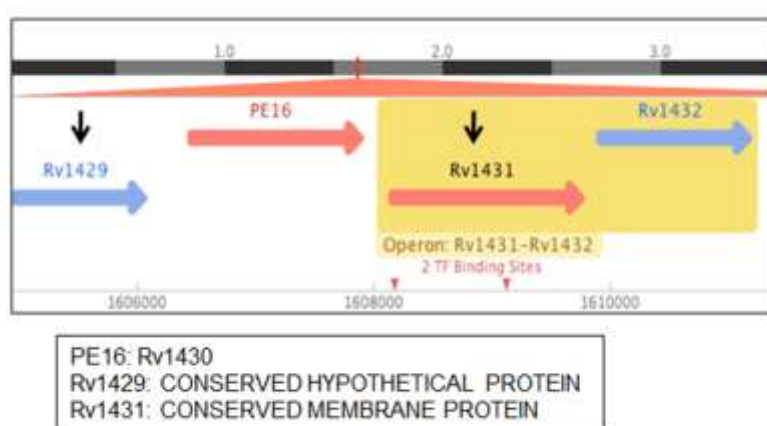
#### **4.2.9 Effect of Salt**

Compatibility of the protein function in the presence of inorganic salts like NaCl and KCl was studied at various salt concentrations such as 100 mM, 200 mM, 300 mM, 400 mM, 500 mM, 600 mM, 700 mM, 800 mM, 900 mM and 1 M in the assay mixture. The residual enzyme activities were recorded using the standard assay protocol with the most active substrate, pNPC6 as discussed above.

## 4.3 Results and Discussion

### 4.3.1 Cloning, Expression and Purification of Rv1430, PE-PPE Domain and Ser199Ala Mutant PE-PPE Domain

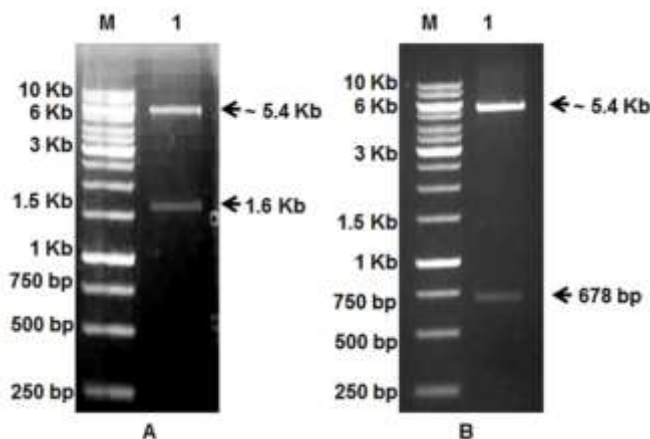
Among the ten proteins identified from our earlier studies, we have chosen to work with Rv1430 protein. As evident from the gene location in Figure 4.1, Rv1430 is flanked with conserved hypothetical proteins, therefore not organized in an operonic pattern with any PPE family of proteins (Figure 4.1, source: <http://www.tbdb.org/> and <http://genolist.pasteur.fr/TubercuList/>).



**Figure 4.1.** Genomic organization of Rv1430 (source: <http://www.tbdb.org/> and <http://genolist.pasteur.fr/TubercuList/>)

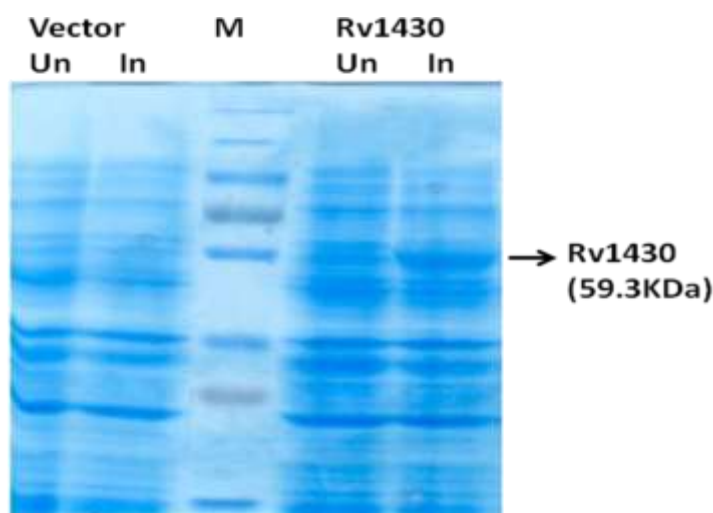
It is known that Rv1430 is a fusion protein with N-terminal PE domain (1-87 amino acids) and a linker region (88 -143 amino acids) linking the PE and C-terminal PE-PPE domains. Therefore, to estimate the consequence of these regions on the enzyme activity of Rv1430, besides the PE-PPE domain we have cloned and purified the full-length Rv1430.

The full-length Rv1430, its PE-PPE domain and mutant PE-PPE domain were separately cloned in pET-28a expression vector. The presence of the inserts were confirmed by performing restriction enzyme digestion of the recombinant plasmid. As shown in Figure 4.2, the DNA insert fragments were confirmed by agarose gel electrophoresis.

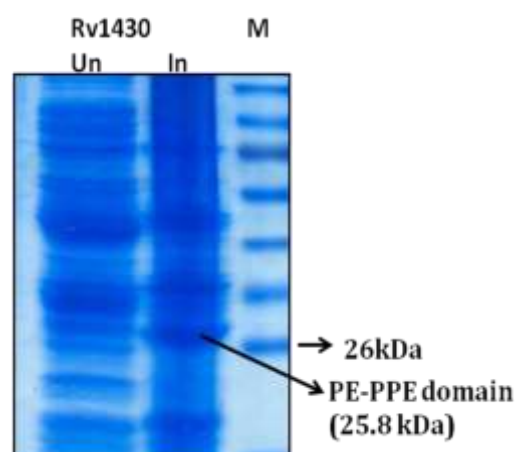


**Figure 4.2.** Cloning of Rv1430c (1601 bp) and its PE-PPE domain (678 bp) in vector pET28a. Agarose gel electrophoresis analysis after restriction digestion of the insert Rv1430 in pET28a (A) and its PE-PPE domain in pET28a (B). The vector and the insert bands are indicated by arrows. M: 1 Kb molecular weight marker

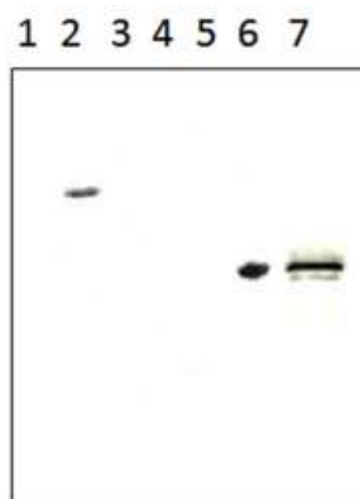
Further, precise sequence of the inserts, including mutant was confirmed by DNA sequencing. The cloning of genes Rv1430 and PE-PPE domain in pET-28a expression vector resulted in a hexahistidine tag towards the N-terminus of the proteins. The proteins were overexpressed using 1 mM IPTG and analysed on SDS-PAGE. The SDS-PAGE results showed a dense protein band corresponding to the expected molecular weight of both Rv1430 and PE-PPE domain as shown in the Figures 4.3.A and 4.3.B. Further the expressions of the proteins was confirmed through the Western blot method as shown in Figure 4.4.



**Figure 4.3.A.** SDS-12% polyacrylamide gel electrophoresis stained with Coomassie brilliant blue R250 of the pET-28a vector carrying Rv1430 gene transformed into *E.coli* Rosetta gami



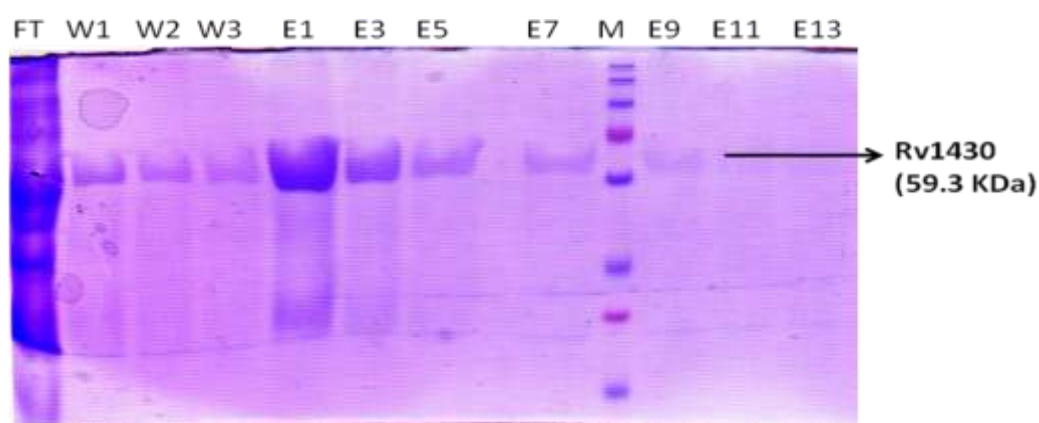
**Figure 4.3.B.** SDS-12% polyacrylamide gel electrophoresis stained with Coomassie brilliant blue R250 of the pET-28a vector carrying PE-PPE domain region transformed into the *E. coli* BL21 (DE3) CodonPlus-RIL



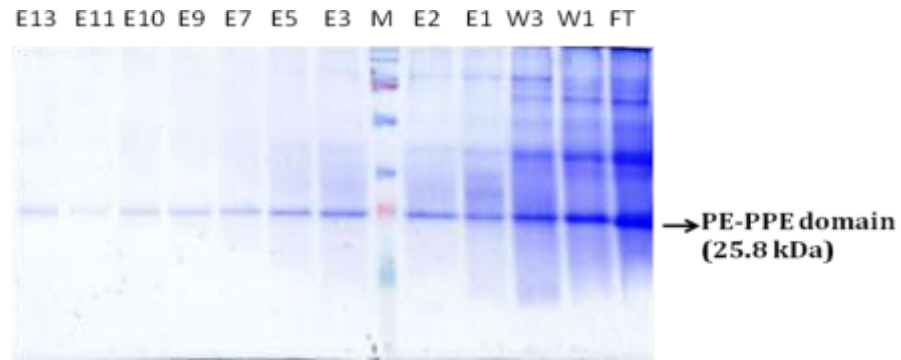
**Figure 4.4.** Western blot analysis of Rv1430 full-length protein and PE-PPE domain by anti-His antibody. Lane1, Uninduced *E. coli* BL21 (DE3) CodonPlus-RIL transformed with Rv1430 construct ; Lane 2, Induced sample - Rv1430 (59.3 kDa); lane 3, Protein Marker; Lane 4, lysate of *E. coli* Rosetta gami transformed with an empty vector; Lane 5, Uninduced Rv1430 PE-PPE domain sample; Lane 6, Induced sample- Rv1430 PE-PPE domain (25.8 kDa). Lane 7, Induced sample-Site directed mutagenesis of Ser199Ala mutated PE-PPE domain (25.8 kDa)

The full-length Rv1430 and its PE-PPE domain when separately overexpressed in bacterial strains, under all the conditions tested for the expression, the proteins were always found in inclusion bodies but not in the soluble fraction. Therefore, protein purification was carried out under denaturing conditions that was later refolded before enzyme activity analysis. This was achieved by using solubilizing agents such as urea;

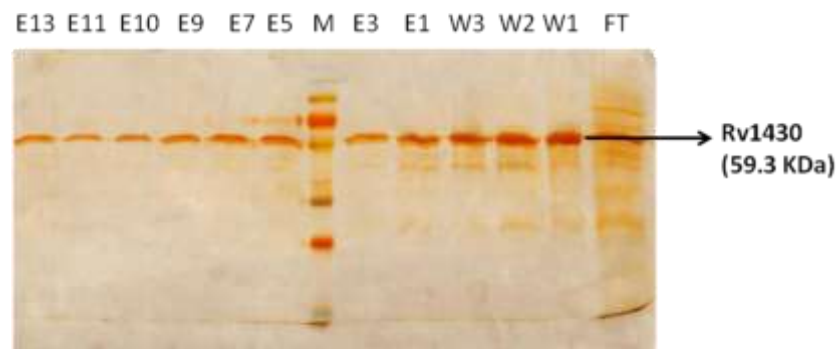
initially solubilizing the inclusion bodies containing the desired proteins in the presence of urea and then refolding the protein step-wise in the absence of urea to recover the active recombinant proteins. We purified the proteins after solubilizing in urea by adding the supernatant containing the desired proteins to the hexahistidine tag specific cobalt affinity chromatography. We collected all the fractions such as flow through (FT), washes (W) and eluents (E) of the proteins during the various purification stages and these were analysed on SDS–12% polyacrylamide gel electrophoresis stained with both coomassie brilliant blue R250 and by silver stained SDS–12% polyacrylamide gel electrophoresis as shown in Figures 4.5.1.A and 4.5.1.B, and Figures 4.5.2.A and 4.5.2.B. The migration of the purified Rv1430 to the expected molecular mass of ~59 kDa size and both PE-PPE domains wild and mutant to the expected molecular mass of ~26 kDa after second purification analyzed on SDS-PAGE (Figure 4.5.3.A, 4.5.3.B and 4.5.3.C) is in good agreement with the theoretical mass as calculated from the protein sequence. Further the purity of the proteins after performing the gel filtration is shown in Figure 4.5.4. We obtained nearly 5 mg of pure protein from 1 L of cell culture.



**Figure 4.5.1.A.** Affinity chromatographic purification of Rv1430 full-length protein (59.3 kDa), flow through (FT), washes (W) and eluted fractions (E) of the proteins recovered during the various purification steps were separated by SDS–12% polyacrylamide gel electrophoresis and stained with Coomassie brilliant blue R250



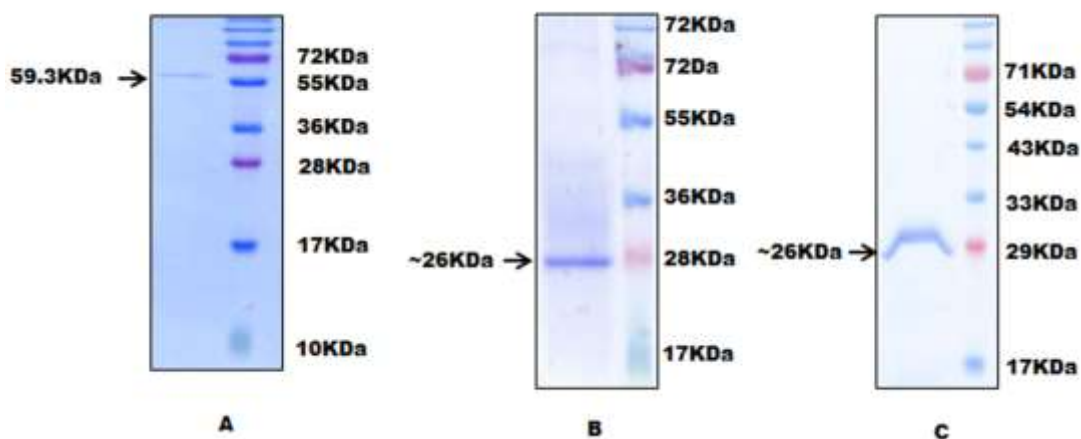
**Figure 4.5.1.B.** Affinity chromatographic purification of PE-PPE domain of Rv1430 (25.8 kDa), flow through (FT), washes (W) and eluted fractions (E) of the proteins recovered during the various purification steps were separated by SDS–12% polyacrylamide gel electrophoresis and stained with Coomassie brilliant blue R250



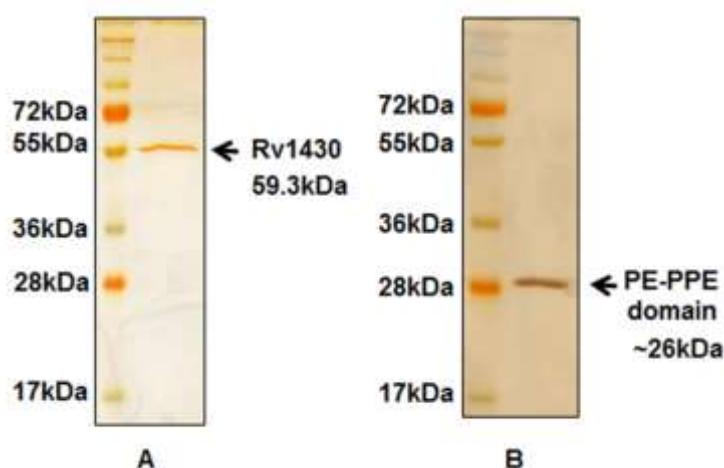
**Figure 4.5.2.A.** Affinity chromatographic purification of Rv1430 full-length protein (59.3 kDa), flow through (FT), washes (W) and eluted fractions (E) of the proteins recovered during the various purification steps were separated by Silver stained SDS–12% polyacrylamide gel electrophoresis



**Figure 4.5.2.B.** Affinity chromatographic purification of PE-PPE domain of Rv1430 (25.8 kDa), flow through (FT), washes (W) and eluted fractions (E) of the proteins recovered during the various purification steps were separated by Silver stained SDS–12% polyacrylamide gel electrophoresis



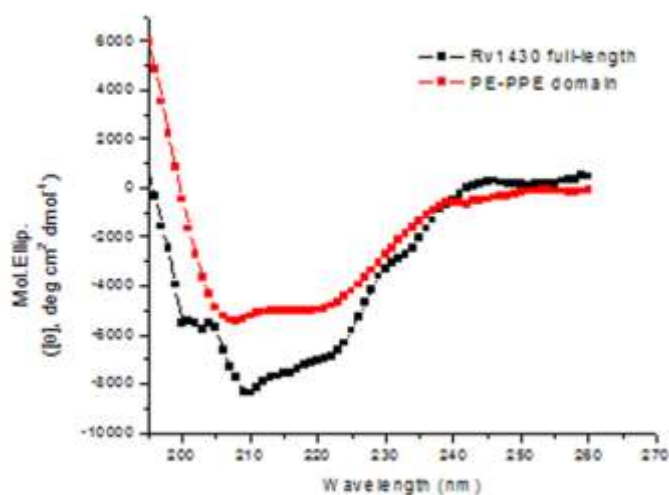
**Figure 4.5.3.** Affinity chromatographic purification of Rv1430 full-length protein, PE-PPE and Ser199Ala PE-PPE domains. Eluted fractions of the proteins recovered during the various purification steps were separated by SDS–12% polyacrylamide gel electrophoresis and stained with Coomassie brilliant blue R250. (A) Purified protein Rv1430 (59.3 kDa); (B) Purified PE-PPE domain of Rv1430 (25.8 kDa); (C) Purified mutated PE-PPE domain of Rv1430 (25.8 kDa)



**Figure 4.5.4.** Silver stained SDS–12% polyacrylamide gel electrophoresis of the purified protein (A) Rv1430 full-length and its (B) PE-PPE domain

### 4.3.2 CD Spectroscopy Data

To determine if Rv1430 and its PE-PPE domain were properly fold into native-like conformations we have examined their CD spectra as is shown in Figure 4.6. The results indicated that the purified proteins were properly refolded.



**Figure 4.6.** CD spectra of Rv1430 full-length (black) and its PE-PPE domain (red) measured in 10 mM Tris, at pH 7.0. Values represent the mean residue molar ellipticity. The concentrations of proteins are 0.125 mg/mL for Rv1430 full-length and 0.15 mg/mL for PE-PPE domain

#### 4.3.3 Rv1430 and its PE-PPE Domain have Esterase Activity

The enzymatic activity of both the proteins Rv1430 and the PE-PPE domain were performed by using the following substrates; pNPC2, pNPC4, pNPC6, pNPC8, pNPC10, pNPC12, pNPC14 and pNPC16 following the previous reported methods (Lopez-Lopez et al., 2003; Pencreach and Baratti, 1996). The enzyme activity data indicated that at pH 7.0 and 37 °C, both the proteins Rv1430 and its PE-PPE domain can hydrolyze a wide range of substrates of *p*-nitrophenyl esters (C4–C12), among which pNPC6 was found to be the most effectively hydrolyzed substrate. In comparison to this substrate, the hydrolysis of substrates pNPC4 and pNPC8 is less. The enzyme activities of both proteins were found to be quite similar as observed in Figure 4.7.1. From the Figure it can be clearly seen that both the proteins had no activity towards the substrates pNPC2, pNPC14 and pNPC16. Table 4.1 shows the relative enzyme activity of proteins Rv1430 and the PE-PPE domain towards the hydrolysis of various derivatives of pNPs.



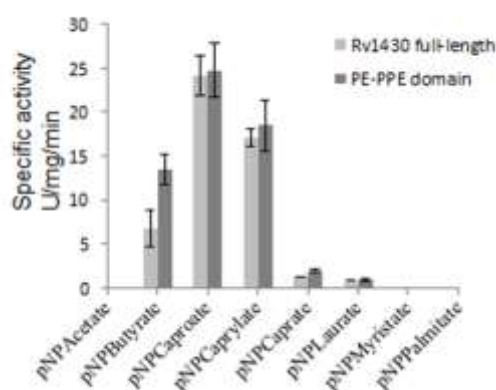
**Table 4.1.** Relative enzyme activity of Rv1430 and its PE-PPE domain towards *p*-nitrophenyl derivatives at pH 7.0 and 37 °C.

Substrate	Relative activity (%) of PE-PPE domain	Relative activity (%) of full-length Rv1430
<i>p</i> -nitrophenyl acetate (C2)	0	0
<i>p</i> -nitrophenyl butyrate (C4)	62	30
<i>p</i> -nitrophenyl caproate (C6)	100*	100*
<i>p</i> -nitrophenyl caprylate (C8)	79	71
<i>p</i> -nitrophenyl caprate (C10)	19	11
<i>p</i> -nitrophenyl laurate (C12)	4	2.5
<i>p</i> -nitrophenyl myristate (C14)	0	0
<i>p</i> -nitrophenyl palmitate (C16)	0	0

\*The specific activity of *p*-NPC6 is taken as 100% at pH 7 and 37 °C which is ~ 22.5 U/mg/min for both Rv1430 domain and Rv1430 protein.

For lipolytic activity, Tween-20 was used as substrate but no absorbance was observed at 405 nm indicating that Rv1430 and its PE-PPE domain does not show lipolytic activity. When cutin was used as substrate that has characteristic C16 and C18 hydroxyl functional group fatty acids, no hydrolysis products were observed indicating that Rv1430 and its PE-PPE domain does not have cutinase activity either. These indicated that the proteins do not hydrolyze higher chain length esters and that Rv1430 and the PE-PPE domain are not lipases or cutinases either.

Further, it was also observed from the data that with the C4 derivative of *p*-nitrophenylester substrate, the PE-PPE domain has almost double the activity when compared to the full-length Rv1430 protein during the hydrolysis of pNPC4 (Figure 4.7.1).



**Figure 4.7.1.** The specific enzyme activity of Rv1430 and PE-PPE domain towards the hydrolysis of various *p*-nitrophenyl ester derivatives at pH 7.0 and 37 °C

#### 4.3.4 Kinetic Properties

The kinetic constants such as  $K_m$ ,  $k_{cat}$  and  $k_{cat}/K_m$  determined from the initial rate of activity for both the proteins Rv1430 and its PE-PPE domain analysed against active substrates found from the above experiments pNPC4, pNPC6 and pNPC8, at pH 7.0 and temperature 37°C are shown in Table 4.2. For kinetic parameter calculations we have used the Lineweaver–Burk plot and the data indicated that pNPC6 is the most appropriate substrate for Rv1430. When kinetic parameters are compared, our observations showed that the most suitable substrate for Rv1430 was pNPC6, similar to the protein Rv0045c which is a characterized esterase of *Mtb* (Guo et al., 2010). Nevertheless, Rv1430 could catalyze the hydrolysis of substrates with chain lengths C4 and C8, while the protein Rv0045c showed better catalytic activities with C2 and C14 substrates also. This indicated that *mycobacterium* possesses an arsenal of lipase/esterases to hydrolyze lipids of varying chain lengths that depends on the accessibility of substrates under various conditions during *in vitro/in vivo* growth. Therefore, it would be interesting to learn about the expression of various esterases/lipases during diverse pathological conditions of bacterial infection.

**Table 4.2.** The kinetic parameters of Rv1430 and PE-PPE domain.

Enzyme	Substrate	$K_m$ (mM)	$K_{cat}$ (s <sup>-1</sup> )	$K_{cat}/K_m$ (M <sup>-1</sup> s <sup>-1</sup> )
Rv1430	pNPC4	5.2	341	6.55x10 <sup>4</sup>
	pNPC6	5.15	534	1.04 x10 <sup>5</sup>
	pNPC8	10	305	3.05 x10 <sup>4</sup>
PE-PPE domain	pNPC4	4.5	182	4.04 x10 <sup>4</sup>
	pNPC6	4.2	250	5.92 x10 <sup>4</sup>
	pNPC8	5.6	99	1.76 x10 <sup>4</sup>

#### 4.3.5 Temperature Dependence, pH Tolerance and Effect of Salt Concentration on the Esterase Activity of Rv1430 and the PE-PPE Domain

The enzyme activities of the purified proteins were studied at various pH within the range 4.0 to 11.0. We observed that at pH 4.0 and 5.0 there was no enzyme activity. At pH 9.0, 10.0 and 11.0, enzyme activity was found too low to be detected because the substrates instinctively decomposed resulting a deep background. Hence, the enzyme activity of both proteins Rv1430 and the PE-PPE domain were tested at mild pH conditions (pH 6.0 – 8.0) and in the temperature range (25 °C to 45 °C) by using the most active substrate, pNPC6. The Figure 4.7.2.A, and Figure 4.7.2.B shows the esterase activity of the PE-PPE domain with respect to pH and temperature. From this figure, the highest enzyme activity

was found between temperatures 37-38 °C and at both pH 7.0 and pH 8.0. At temperature 39 °C, 50% inactivation was observed and at temperature 40 °C, the PE-PPE domain was completely inactive. These results support the fact that, *Mtb* is a pathogen infecting the human or animal bodies and hence these proteins are active in the ambient conditions of the body temperature. It also showed that the highest activity of these proteins increased rapidly following increased pH. Consequently, it may be possible that Rv1430 like other serine hydrolases of *mycobacterium* is involved in ester metabolism of this pathogen under less favorable conditions of starvation. This supports the fact that *Mtb* becomes more pathogenic at alkaline conditions of pH as proposed by Guo et al (Guo et al., 2010).

Figure 4.7.2.A

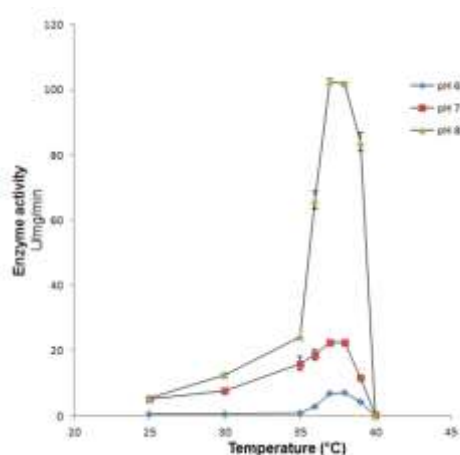
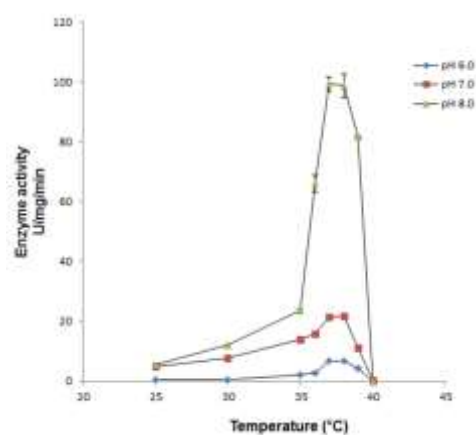
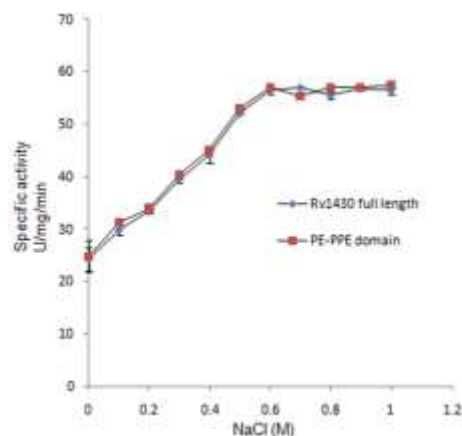


Figure 4.7.2.B



**Figure 4.7.2.A.** Effect of pH as a function of temperature on enzyme activity of PE-PPE domain. **Figure 4.7.2.B.** Effect of pH as a function of temperature on enzyme activity of Rv1430 full-length protein

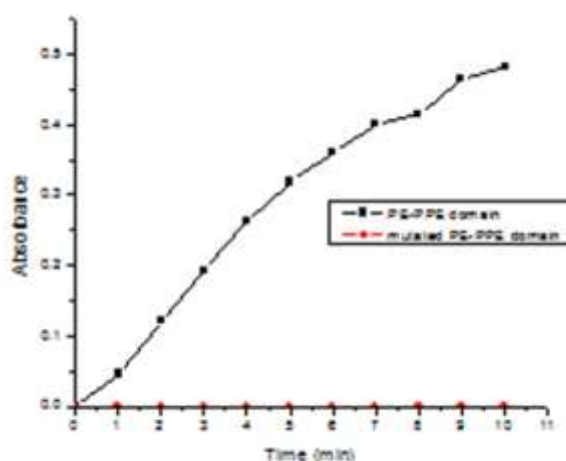
To analyze the effect of ionic interactions on the esterase activity of proteins Rv1430 and the PE-PPE domain, both the proteins were initially incubated with high ionic salt buffers, 15 min prior to the enzymatic assay. It was found that as salt concentration increased, the enzymatic activity of both the proteins Rv1430 and the PE-PPE domain also increased. Similar results were observed for NaCl (Figure 4.7.3) and KCl and the increase in activity was observed till 600 mM which later on remained constant up to 1 M salt concentration.



**Figure 4.7.3.** Effect of the NaCl concentration on the esterase activity of the Rv1430 and PE-PPE domain was determined by the hydrolysis of pNPC6

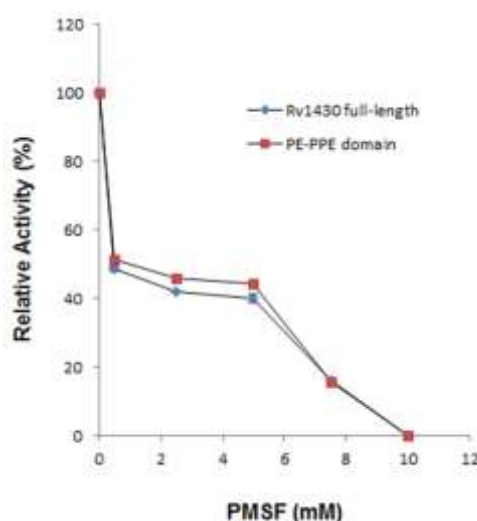
#### 4.3.6 Rv1430 belongs to Serine Hydrolase Family of Proteins

From earlier chapters on PE and PPE proteins it was observed that the PE-PPE domain displays a "serine  $\alpha/\beta$  hydrolase" fold with a characteristic conserved Ser, Asp and His catalytic residues forming a triad. To confirm that Ser199 is important in the catalytic triad, the Ser199Ala residue of the PE-PPE domain was mutated and purified, the esterase activity of the mutant protein was measured under identical conditions as followed for the PE-PPE domain by using most active substrate pNPC6. The mutant PE-PPE domain was observed to be enzymatically inactive (Figure 4.7.4) on the substrate pNPC6, which indicated that Ser199 residue is a part of the catalytic triad in the PE-PPE domain.



**Figure 4.7.4.** A time course of pNPC6 hydrolysis at pH7.0, 37°C with PE-PPE domain (black) and Ser199Ala mutated PE-PPE domain (red) of Rv1430 protein

For further confirmation of Ser199 as part of the catalytic triad, the enzyme activity was measured in the presence of different concentrations of PMSF by pre-incubating the enzyme prior to the enzymatic assay. It is well known that PMSF is a specific serine protease inhibitor that covalently binds to the active site amino acid serine. We observed that the enzyme activities of full-length Rv1430 and its PE-PPE domain were inhibited by nearly 50% even at lesser concentrations of PMSF such as 0.5 mM. The enzyme activities were well abolished at higher concentrations of PMSF when the enzyme was pre-incubated with PMSF for 15 min at 30 °C as shown in Figure 4.7.5. These results confirm the serine hydrolase activity of Rv1430.



**Figure 4.7.5.** Effect of PMSF on the hydrolysis of pNPC6 by Rv1430 and PE-PPE domain

From our experimental studies, we observed that the enzyme activities of full-length Rv1430 and its PE-PPE domain are similar and comparable. It implicates that the conserved N-terminal PE domain and its C-terminal PE-PPE domain of Rv1430 (PE16) protein have independent roles and so the PE-PPE domain is sufficient to display the serine hydrolase activity. The existence of two independent domains within a single protein is indicative of gene fusion (Snel et al., 2000). It was reported that PE and PPE proteins are accountable for surface antigenic variation (Cole et al., 1998) and some earlier reports indicated that few PE proteins are cell surface exposed (Adindla and Guruprasad, 2003; Kaufmann and Helden, 2008). Therefore, it can be believed that the PE-PPE domain with its serine hydrolase activity is situated on the cell surface and the N-terminal PE or PPE domains may function by translocating the PE-PPE domain to the site of action on the cell surface. The inhibition studies conducted with and without the PMSF

and the mutant Ser199Ala PE-PPE domain, confirm the serine hydrolase activity of the PE-PPE domain.

#### **4.4 Conclusion**

The Rv1430 gene and its PE-PPE domain region were effectively cloned, overexpressed and both the proteins were separately purified to homogeneity by using affinity and column chromatography. The CD spectral data provided evidence for the folded proteins following refolding and purification. Both the proteins do not exhibit lipase and cutinase activity, however the proteins hydrolyse short to medium chain lengths fatty acid esters with the maximum specific activity towards pNPC6 at 37 °C, 38 °C and at pH 7.0 and 8.0. These results are supported by the kinetic parameters. These proteins are efficiently inhibited by PMSF which is a serine protease inhibitor. The Ser199Ala mutant PE-PPE domain showed no esterase activity proving that Rv1430 is certainly an esterase. The esterase activity of the proteins, full-length Rv1430 and its PE-PPE domain are similar thus indicating that the role of the PE-PPE domain region is independent from the rest of the protein.





## **Chapter 5**

### **Inhibitor Binding Studies of Rv1430 PE-PPE Domain of *Mycobacterium tuberculosis*: Virtual Screening and Molecular Dynamic Simulations**



## 5.1 Introduction

*Mtb*, a member of the closely related group of slow growing pathogenic *mycobacteria* called MTBC is responsible for causing deadliest infectious disease TB (McEvoy et al., 2012; Zaman, 2010). TB remains a major health issue all over the world infecting over one third of the world's population. Although the BCG vaccination was introduced 90 years back, the heterogeneity in the efficacy of protective effect of BCG vaccine varies due to several factors such as the differences in exposure to a typical *mycobacteria* in the environment, the genetic susceptibility of the population, differences in the virulence affect of the *Mtb*, high risk of reinfection or reactivation, the availability of different BCG strains, nutritional differences and etc (Pereira et al., 2007; Sterne et al., 1998). The current first line drugs for anti-TB treatment are a multidrug course of therapy which consists of oral rifampicin, isoniazid, pyrazinamide and ethambutol. The second line of drugs consists of injectable aminoglycosides, injectable fluoroquinolones, *p*-aminosalicylic acid. The third line drugs consists of clofazimine, linezolid, amoxicillin, clavulanate, imipenem, cilastatin and clarithromycin. The basic treatment for TB involves the first-line anti-TB drugs. If these drugs are misused or poorly mismanaged, MDR-TB can develop. MDR-TB refers to the strains of the tuberculosis bacillus that are resistant to at least the two most powerful drugs of first line anti-TB drugs (rifampicin and isoniazid). Further, XDR-TB can develop with the poor management of these second line drugs or subset of MDR-TB that are resistant to nearly all current anti-TB drugs (Keshavjee et al., 2008). *Mtb* can also become drug resistant due to mutations in genes encoding drug targets (Madania et al., 2012). There is an increasing clinical occurrence of *Mtb* strains with XDR-TB where mortality rates are increasing (Berry and Kon, 2009). The problem with the present anti-TB drugs is their side-effects that mainly cause liver damage and other adverse reactions like nausea, vomiting and anorexia leading to hospitalization (Ormerod and Horsfield, 1996; Shang et al., 2011; Yee et al., 2003).

Therefore the hunt for novel TB drug targets is an ongoing pursuit to inhibit drug resistant bacteria, unique to pathogen and absent in human host. Lipoate protein ligase B (LipB), an enzyme involved in the biosynthesis of the lipoic acid cofactor, is a promising drug target in *Mtb*. By using virtual screening methods, it has been shown that NSC164080 (methyl 2-(2-(((benzyloxy)carbonyl)amino)propanamido)-3-(4-hydroxyphenyl)propanoate) is the best inhibitor (Junie B. Billones, 2013). For example, the MurE enzyme involved in Mur pathway of *Mtb* is an attractive drug target as it is unique to bacteria and is absent in human (Singh et al., 2014b). Therefore it is reasonable

to identify proteins that are specific to the *Mtb* and absent in human host, as ideal drug targets.

Two large gene families PE and PPE are highly expanded throughout the pathogenic *mycobacteria*, whereas nonpathogenic *mycobacteria* comprises relatively few PE and PPE genes (Ekiert and Cox, 2014). These gene families are specific to only *mycobacteria* and absent in the human host. These therefore become good candidates for drug design studies.

In our previous chapters, from PE and PPE protein sequence analysis, we have shown a common conserved domain present in some PE and PPE proteins in *mycobacteria*. This 225 amino acids domain, termed as PE-PPE domain (Adindla and Guruprasad, 2003) is present towards the C-terminus. Using the fold prediction methods and homology modeling, we proposed serine  $\alpha/\beta$  hydrolase fold with the pentapeptide sequence motif GxSxG/S and conserved Ser, Asp and His catalytic triad characteristic of lipase, esterase and cutinase activities as described in the previous chapter (Sultana et al., 2011). Biochemical and mutational studies on the Rv1430c (PE16) PE-PPE domain confirmed its esterase function as described in the previous chapter.

Esterases or lipases are hydrolases required for lipid metabolism in both prokaryotes and eukaryotes. It was reported that there are at least 250 enzymes related to lipid metabolism which includes extracellular secreted enzymes, integrated cell wall enzymes and intracellular esterases/lipases in *Mtb* (Camus et al., 2002; Cole et al., 1998). Further, it has been reported that most of the mycobacterial genes involved in lipid metabolism, cell division chromosomal partitioning and secretion are required during infection in mouse model (Lamichhane et al., 2005; Sasseti et al., 2003).

The Rv3802c is a secreted protein and participates in cell wall metabolism and is absent in the human host, suggesting its virulence role in pathogenesis of *Mtb*. Based on the analysis, using sequence and structure prediction, and comparative docking studies of Rv3802c potential inhibitors effective towards mycobacterial proteins were identified (Saravanan et al., 2012). The Rv3203, another lipase from *Mtb* H37Rv is known to upregulate during acidic stress (Singh et al., 2014a). The Rv2224c, another gene of H37Rv belonging to the microbial esterase/lipase family was identified as a virulence gene, required for bacterial survival in mouse model and full virulence of *Mtb* (Lun and Bishai, 2007). Rv1399c and Rv0045c are novel proteins which have putative hydrolase function, probably involved in the ester/lipid metabolism of *Mtb* (Canaan et al., 2004; Guo et al., 2010).

The importance of PE and PPE proteins in *Mtb* compounded with the significant role played by cutinases/esterases/lipases in the cell wall and lipid metabolism of pathogenic *mycobacteria* strongly implicates that the PE-PPE domain (Rv1430 esterase domain) is a biological receptor to design new anti-TB potential drug candidates with high affinity and selectivity using various computer aided drug design tools to combat TB. We have therefore used *in silico* screening of large databases of molecules and molecular docking techniques to select new and potent Rv1430 esterase domain inhibitors for better drug candidates. To confirm the stability of their binding and energetics, we have carried out molecular dynamics simulations of the protein-inhibitor complexes. These results explain the basis for inhibitor binding to Rv1430 esterase domain and provide more precise directions in the design of new inhibitors.

## 5. 2 Methods

### 5.2.1 Virtual Screening

For virtual screening we have used DrugBank approved molecules from ZINC database, which is freely available at <http://zinc.docking.org> and contains commercially available molecules in 3D formats for structure/target based virtual screening (Irwin and Shoichet, 2005; Irwin et al., 2012). Initially all the 6447 drug molecules were converted into pdbqt format using MGLTools-1.5.6 for the virtual screening into the Rv1430 esterase domain using AutoDock Vina. Hydrogen atoms were added to Rv1430 model followed by the addition of Gasteiger–Marsili charges. The non-polar hydrogens were merged onto their respective heavy atoms and atom types were fixed using AutoDock Tools (Morris et al., 2009). The active site of the Rv1430 esterase domain was covered by a grid box with 20 x 22 x 22 points and 1 Å was given for grid spacing with 47 x 1 x 30.8 centre of the grid box. AutoDock Vina software uses gradient optimization method and multithreading with local optimization (Trott and Olson, 2010). The number of orientations for each molecule was set to 10. The best conformers with highest binding energies to the Rv1430 esterase domain were chosen on the basis of affinity score. These protein ligand docking conformations were analyzed using Discovery Studio 2.5 visualizer to understand the inhibitor binding interactions. The top hit molecules with highest affinity towards the esterase domain of Rv1430 were subjected to MD simulations and binding energy calculations were carried out subsequently.

### 5.2.2 MD Simulations

The best orientated and high affinity Rv1430 esterase domain-inhibitor complexes were used for further MD simulations studies. These Rv1430 esterase domain-inhibitor complexes were placed in a 10 Å cubic water box with TIP3P water molecules. MD simulations studies were carried out using GROMACS 4.5.5 simulation package (Hess et al., 2008; Van Der Spoel et al., 2005) with the AMBER99SB as force field for protein. The inhibitor parameter files are generated with GAFF force fields in antechamber (Wang et al., 2006; Wang et al., 2004) module using ACPYPE script (Sousa da Silva and Vranken, 2012). Then the Rv1430 esterase domain-inhibitor complexes were allowed for energy minimization for 2000 steps of steepest descent, 2000 steps of conjugate gradient, and 1 ns position-restrained dynamics for the distribution of the water molecules throughout the protein-inhibitor complexes. MD simulations were performed on the whole system for about 25 ns, using 0.002 ps time step. The particle-mesh Ewald (PME)

summation method (Darden T, 1993; Ulrich Essmann, 1995) was employed for the electrostatic calculation, with a real space cut-off of 10 Å, PME order of 6, and a relative tolerance between long- and short-range energies of  $10^{-6}$ . To evaluate the short-range interactions, a neighbour list of 10 Å was updated for every 10 steps, Lennard–Jones (LJ) interactions and the real space electrostatic interactions were truncated at 9 Å. Isothermal–isobaric ensemble (NPT) of 298 K and 1 atmosphere were used along with the periodic boundary conditions. The V-rescale thermostat method for temperature (Bussi et al., 2007) and the Parrinello–Rahman algorithm (Parrinello M, 1981) for pressure maintenance were used, and the hydrogen bonds were constrained using LINCS algorithm (Hess et al., 1997). Binding free energy calculations were performed by using the trajectory file obtained from MD simulations. The RMSD of protein Cα atoms was calculated from g\_rms program of GROMACS by least-square fitting the structure to the reference structure.

### 5.2.3 Binding Free Energies Calculations

The binding free energies of Rv1430 esterase domain with potential inhibitors were calculated using Solvated Interaction Energies (SIE) method (Naim et al., 2007; Sulea et al., 2011). Sietraj ([http://www2.bri.nrc.ca/ccb/pub/sietraj\\_main.php](http://www2.bri.nrc.ca/ccb/pub/sietraj_main.php)) is an alternative to the MM-PBSA software provided by the AMBER distribution (Cui et al., 2008; Naim et al., 2007). This method was successfully utilized in earlier work (Tanneeru et al., 2015; Tanneeru and Guruprasad, 2013). The binding free energies ( $\Delta G$ ) of the Rv1430 esterase domain-inhibitor complexes were calculated for the snapshot structures obtained from the MD trajectory of the systems. The  $\Delta G$  is the sum of Coulomb interactions, intermolecular van der Waals (vdw), the change in reaction-field energy (obtained from solving the Poisson–Boltzmann equation) and nonpolar solvation energy (that is proportional to the solvent accessible surface area) (Naim et al., 2007). Similar to MM-PBSA/GBSA, SIE method treats the protein–ligand system in atomistic detail and implicit solvent model. The free energy of binding of the protein-inhibitor complex was calculated using the equation below:

$$\Delta G_{\text{bind}}(\rho, D_{\text{in}}, \alpha, \gamma, C) = \alpha [\Delta E_{\text{vdW}} + \Delta E_{\text{Coul}}(D_{\text{in}}) + \Delta G_{\text{RF}}(\rho, D_{\text{in}}) + \gamma \Delta \text{SA}(\rho)] + C$$

Here  $\Delta E_{\text{vdW}}$  and  $\Delta E_{\text{Coul}}$  are the intermolecular vdW and Coulomb interaction energies between protein and inhibitor. The  $\Delta G_{\text{RF}}(\rho, D_{\text{in}})$  is the difference between the reaction-field energy of the bound and free state of the protein–inhibitor complex, calculated by solving the Poisson equation with BRIBEM software (Purisima, 1998;

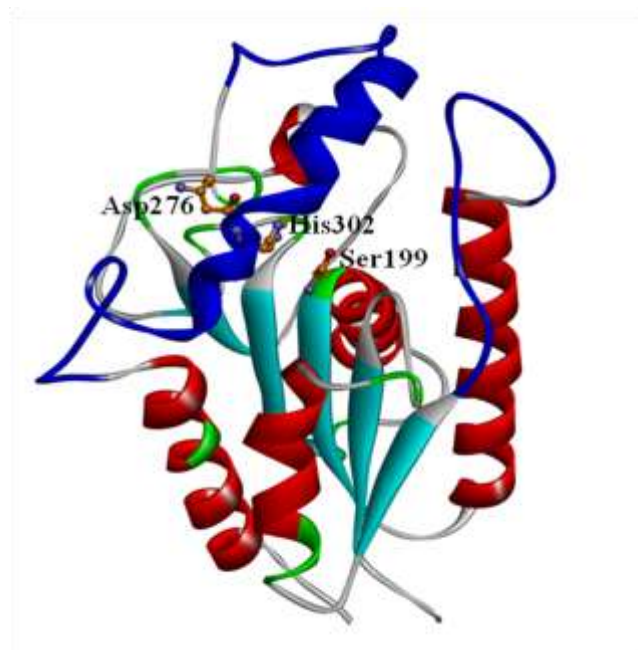
Purísima and Nilar, 1995). The difference in molecular surface area between the bound and free state of the protein is  $\Delta SA(\rho)$ . The cavity energy is the change in the molecular surface area ( $\Delta SA$ ) which is calculated from  $\gamma \Delta SA(\rho)$ . The linear scaling factor  $\rho$  (1.1) is the vdW radii of the AMBER99 force field, and  $D_{in}$  (2.25) is the solute interior dielectric constant. The prefactor  $\alpha$  (0.104758) implicitly quantifies the loss of entropy upon binding, also known as entropy–enthalpy compensation. The coefficient  $\gamma$  (0.012894) is the molecular surface tension coefficient which describes the nonpolar component of solvation free energy, and a constant  $C$  (−2.89) includes protein-dependent contributions not explicitly modeled by the SIE methodology. The scaling can be considered as a simple treatment of entropy–enthalpy compensation containing the caveats of implicit solvation and neglecting the vibrational entropy (Chen et al., 2004; Naim et al., 2007). Here, we have estimated the  $\Delta G$  of structures selected for 100 snapshots from the MD trajectory and averaging over the resulting free energies obtained from each snapshot. We have calculated the contribution of each active site residue to the binding free energy of the inhibitors.



## 5.3 Results and Discussion:

### 5.3.1 Homology Modeling and Virtual Screening

The Rv1430 esterase domain model was constructed and reported in the earlier chapter. Briefly, the FUGUE (Shi et al., 2001) method identified the probable fold of the query protein (Rv1430 esterase domain), as the PDB\_ID: 3AJA corresponding to a lipase from *M. smegmatis* strain MC2155 with highest Z- score of 21.62. From the sequence alignment we observed that GxSxG/S sequence motif and the catalytic triad amino acid residues Ser199, Asp276 and His302 of Rv1430 esterase domain are conserved in both template and query proteins. The 3D structure model of the esterase domain was generated by using MODELLER (Sali and Blundell, 1993) and exhibited an overall  $\alpha/\beta$  hydrolase fold with central  $\beta$ -sheet, flanked by  $\alpha$ -helices on either side as shown in Figure 5.1. A lid region covers the active site from the top of the catalytic site to render the catalytic Ser inaccessible to the solvent. The lid insertion region of the closed conformation is displaced to open and expose the catalytic triad during interfacial activation that allows the substrate to be activated. Therefore, the lid region will alter the conformation of the protein with opening and closing of the active site and is therefore responsible for the activity of the protein. There are many reports which explain similar mechanism of lid movement in the lipase interfacial activation (Cherukuvada et al., 2005; James et al., 2003; James et al., 2007). The serine hydrolases also have an oxyanion hole that constitutes a part of the active site and the amino acid residues of the oxyanion hole mainchain nitrogens participate in the hydrogen bonding as donors atoms to the hydrolysed substrate, stabilizing the negative charge on the tetrahedral intermediate occurring from the nucleophilic attack of the catalytic Ser during activation (De Simone et al., 2004). The location of the lid insertion region and the oxyanion hole are mapped on the Rv1430 esterase domain model structure. The model quality validated using Verify 3D showed higher than 85 3D-1D correlation indicating accuracy in the model constructed. In the close vicinity of the catalytic triad, amino acids, Ser120, Thr121, Tyr130, Met131, Phe157, Gln158, Pro159, Trp160, Thr161, Tyr168, Phe198, Gln200, Phe278, Ala294, Leu295, Leu296, Ile298, Tyr299 and Ser303 are located thus identifying the larger substrate/inhibitor binding cleft. In the Rv1430 esterase domain, amino acid residues (285-315) are identified as the lid insertion region and the oxyanion hole is formed by the amino acid residues Thr121 and Gln200.



**Figure 5.1.** Homology model of Rv1430 esterase domain constructed using MODELLER. The catalytic triad is shown in ball and stick model (colour according to atom type)

### 5.3.2 Virtual Screening

All the Drug Bank approved molecules (6447 molecules) were downloaded from ZINC database and were chosen for the virtual screening of the PE-PPE domain. Virtual screening of the database was performed with an efficient docking program Autodock Vina, which identified molecules that interacted well with the protein. Recently many researchers successfully utilized this software for the docking and virtual screening of the databases (Anand, 2016; Fonseca et al., 2016; Saravanan et al., 2012). We identified the top 100 drug molecules with over -9.0 kcal/mol Autodock free energy of binding. These molecules were further validated and confirmed by using computer graphics on DS visualizer 2.5. We observed from the literature that, in lipase/esterase like hydrolase proteins, the catalytic serine residue forms the covalent bond with many irreversible drug molecules, for example, JZL184 (Long et al., 2009a; Long et al., 2009b) and tetrahydrolipstatin (Saravanan et al., 2012). By considering this concept of this catalytic serine - ligand interaction, we observed the orientation of the drug molecule in the active site along with the hydrophilic interaction with catalytic serine. From this aspect, we shortlisted 10 molecules and these molecules were again subjected to secondary screening by increasing the number of GA runs to 20 and we finally identified five molecules (ZINC13681668, ZINC16052749, ZINC01547088, ZINC16052239, ZINC16052883) that

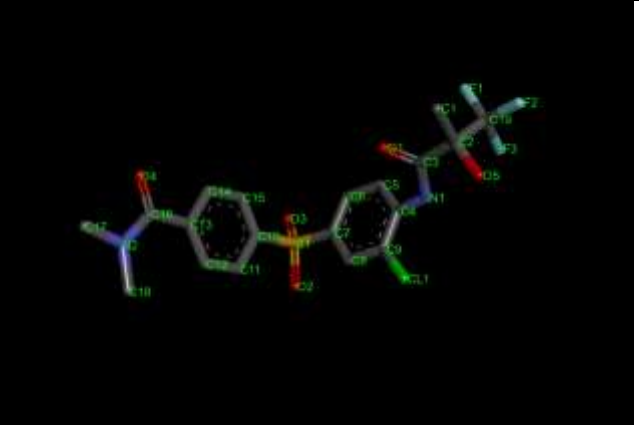
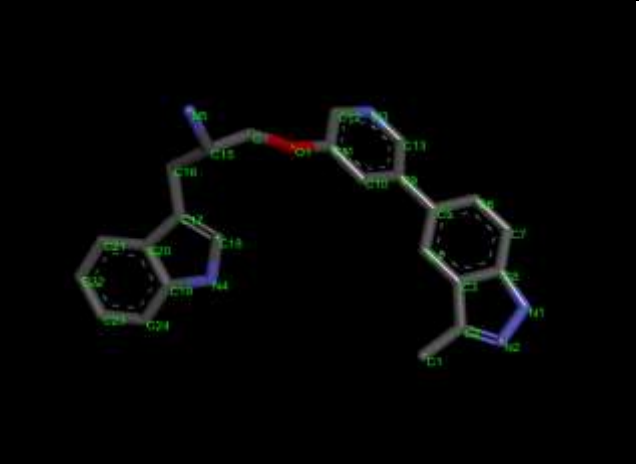
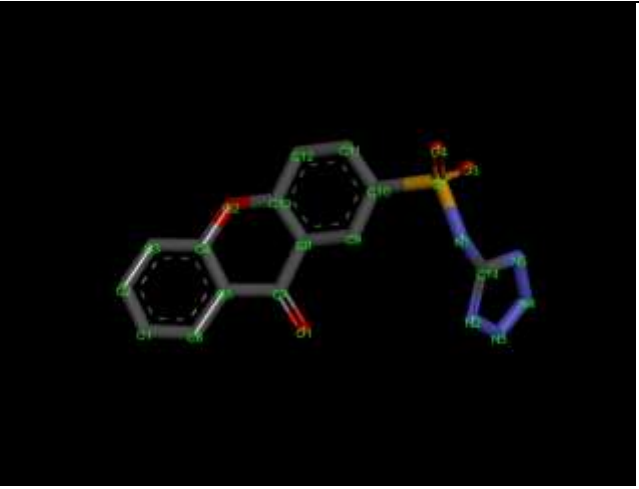
show good interactions with the protein. We found the binding free energy values and the conformation of bindings are same as previously identified from virtual screening of the entire database. These selected 5 molecules were also docked in to the binding site of the human esterase D (PDB\_ID: 3FCX\_A) using Autodock vina with 20 GA runs. The binding energy values of top 5 hits are shown in Table 5.1. The docking results assured us that the binding of these molecules to Rv1430 is specific and have only weak binding to the human esterase.

From the docking results, five energy-minimized inhibitors that display good binding to the enzyme active site are i) ZINC13681668, the substituted xanthane molecule crystal structure observed in the PDB databank complexed with type II dehydroquinase from *H. pylori* (PDB\_ID: 2C4W) ( $K_d = 20000$  nM); ii) ZINC16052749, a kinase inhibitor also observed in the crystal structure in complex with tyrosine kinase domain of the hepatocyte growth factor receptor c-met (3CTJ) ( $IC_{50} = 100$  nM); iii) ZINC01547088, known as glucose lowering drug is observed in crystal structure complex of pyruvate dehydrogenase kinase (2Q8G) ( $IC_{50} = 87$  nM). iv) ZINC16052239 identified as inhibitor in few kinase crystal structures (2JDR, 2JDS, 2JDV etc) and this molecule is known as a Ser/Thr kinase inhibitor ( $IC_{50} = 0.5-30$  nM) v) ZINC16052883, a benzimidazole molecule present in the crystal structure complexed with VEGFR2 kinase domain (PDB\_ID: 2QU5) as an inhibitor ( $IC_{50} = 9$  nM). From the molecular docking using graphical analysis in DS 2.5 Visualizer, we observed that the molecules/inhibitors were well resided in the active site of the protein. The molecular structures of these inhibitors is shown in Table 5.2.

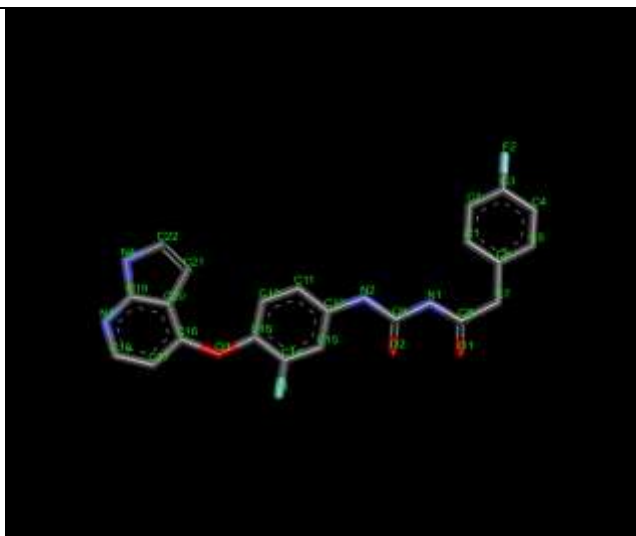
**Table 5.1.** Free energy of binding of top 5 hits of inhibitor molecules from ZINC database selected by virtual screening performed using Autodock Vina docking program into the active site of Rv1430 esterase domain.

S.No.	Molecule	Free energy of binding (kcal/mol)		$\Delta G$ difference	SIE of Rv1430
		Rv1430 $\Delta G$	3FCX:A $\Delta G$		
1	ZINC01547088	-9.7	-6.6	-3.1	$-7.39 \pm 0.65$
2	ZINC13681668	-9.9	-7.2	-2.7	$-7.51 \pm 0.47$
3	ZINC16052239	-10.0	-7.4	-2.6	$-8.61 \pm 0.40$
4	ZINC16052749	-10.2	-7.1	-3.1	$-9.16 \pm 0.44$
5	ZINC16052883	-10.0	-7.4	-2.6	$-7.93 \pm 0.49$

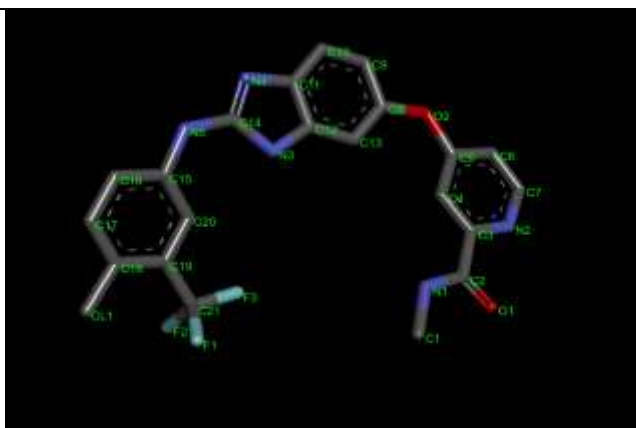
**Table 5.2.**The molecular structures of the top 5 inhibitor molecules.

Ligand	Molecular Structures
ZINC01547088	
ZINC16052239	
ZINC13681668	

ZINC16052749

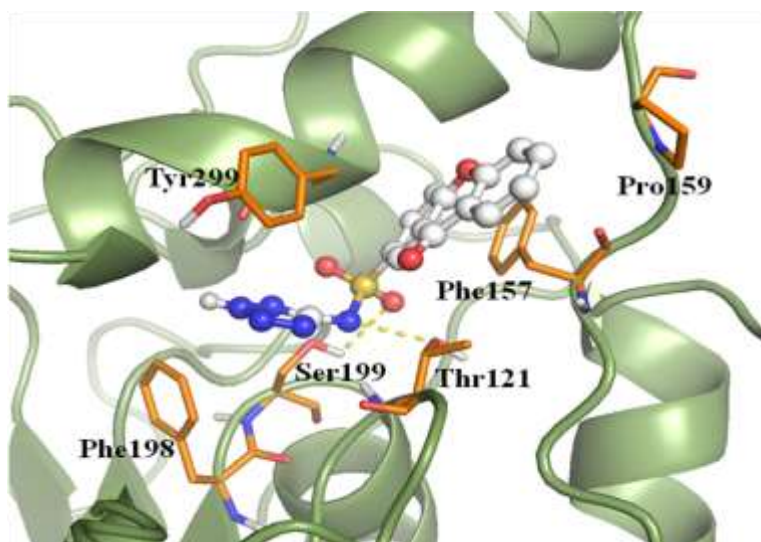


ZINC16052883



### 5.3.3 Molecular Docking

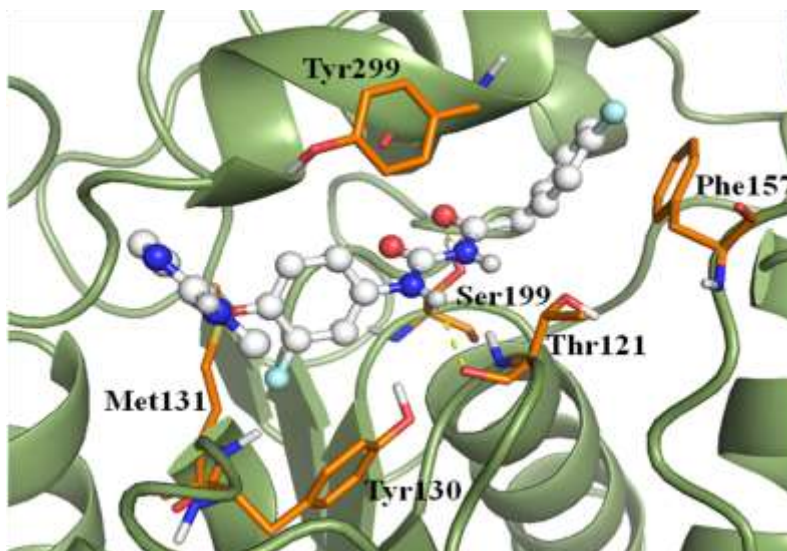
**ZINC13681668:** The docking results of ZINC13681668 in the active site of Rv1430 esterase domain were analyzed. The inhibitor was located in the active site and displays close interaction with catalytic Ser199 of the Rv1430 as shown in Figure 5.2. From our docking studies we observed that SO<sub>2</sub> group oxygen forms hydrogen bond with hydroxy group of Ser199 (OH...O4, 2.725 Å). The tetrazine N5H of the inhibitor molecule forms hydrogen bond with hydroxy oxygen of Ser199 (HO.... HN5, 2.79 Å). The Phe157 sidechain forms pi-pi stacking interaction with the inhibitor aromatic ring in the active site. The sidechain of Tyr299 on the lid region shields the tetrazole ring from the solvent accessible area to hold the molecule in the active site.



**Figure 5.2.** The interaction of the top hit ZINC13681668 with the active site of the Rv1430 esterase domain. Hydrogen bonding interactions are indicated in yellow dashed lines

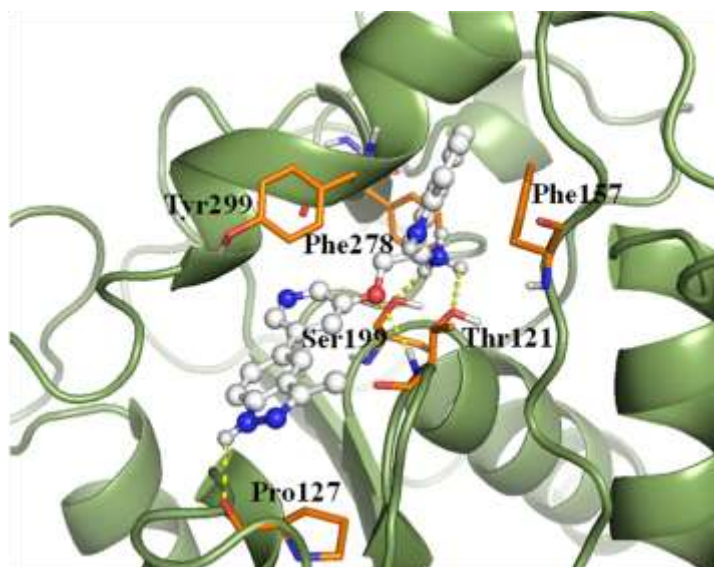
**ZINC16052749:** The docking conformation of the molecule reveals the non bonding interaction with Rv1430 esterase domain. The carbonyl oxygen of the inhibitor forms hydrogen bond with the sidechain hydroxy group of the Ser199 (OH...O1, 2.96 Å). The mainchain carbonyl oxygen of Thr121 forms hydrogen bond with N<sub>1</sub>H of the molecule (CO...HN1, 3.0 Å). The substituted flourine forms hydrogen bond with mainchain NH of the Met131 (NH...F1, 2.9 Å). The *p*-flouro substituted phenyl ring of the molecule forms pi-pi stacking interaction with the sidechain of Phe157. The inhibitor is located close to the lid region residues such as Tyr299 as shown in Figure 5.3.





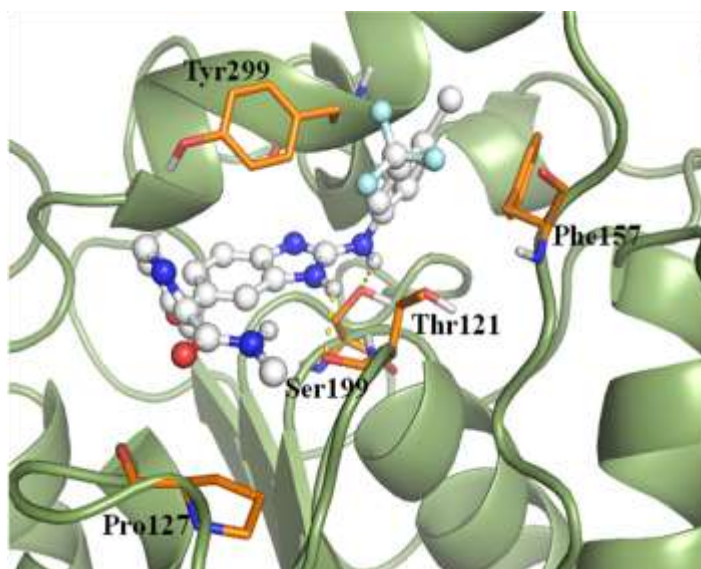
**Figure 5.3.** The interaction of the top hit ZINC16052749 with the active site of the Rv1430 esterase domain. Hydrogen bonding interactions are indicated in yellow dashed lines

**ZINC16052239:** The docking conformation of this molecule into Rv1430 showed the primary amine group of inhibitor has hydrogen bond with Ser199 (HO...HN5, 2.48 Å). The same amino group also forms hydrogen bond with sidechain hydroxy group of Thr121 (HO...HN5, 3.0 Å). The NH of the indazole group forms hydrogen bond with mainchain carbonyl oxygen of Pro127 (CO..HN1, 2.267 Å) as shown in Figure 5.4.



**Figure 5.4.** The interaction of the top hit ZINC16052239 with the active site of the Rv1430 esterase domain. Hydrogen bonding interactions are indicated in yellow dashed lines

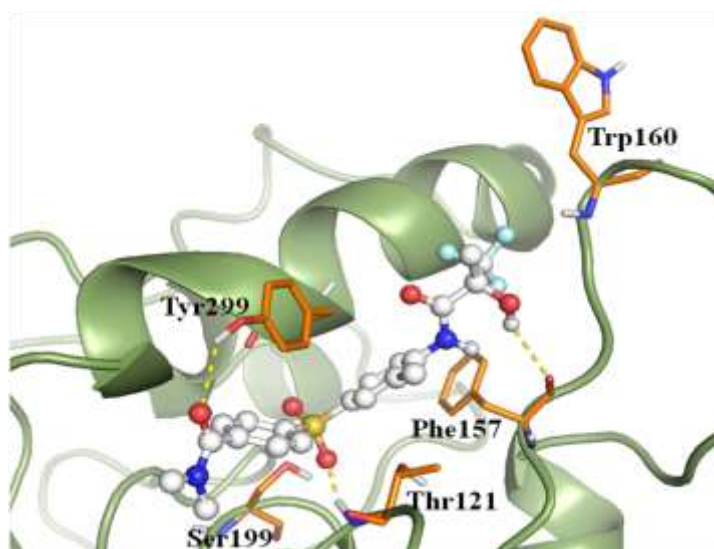
**ZINC16052883:** From the docking conformation, we observed that the NH of the molecule forms hydrogen bond with the Ser199 (HO..HN5, 2.43 Å). The NH of the benzimidazole forms hydrogen bond with sidechain hydroxy group of Thr121 (HO..HN3, 2.47 Å). The substituted phenyl ring of the inhibitor forms pi-pi interactions with the sidechain phenyl ring of Phe157 as shown in Figure 5.5.



**Figure 5.5.** The interaction of the top hit ZINC16052883 with the active site of the Rv1430 esterase domain. Hydrogen bonding interactions are indicated in yellow dashed lines

**ZINC01547088:** From the docking conformation, we observed that one of the oxygens on the sulfate group of the inhibitor forms trifurcated hydrogen bonds with sidechain hydroxy group of the Ser199 (OH...O2, 2.7 Å), sidechain hydroxy group of Thr121 (OH...O2, 2.75 Å) and mainchain NH group of Thr121 (NH...O2, 2.85 Å). The sidechain hydroxy group of the inhibitor forms hydrogen bond with the mainchain carbonyl oxygen of Phe157 (CO...HO5, 2.60 Å). One of the fluorine atoms of CF<sub>3</sub> group forms hydrogen bonding with the mainchain NH of the Trp160 (NH...F2., 3.2 Å) as shown in Figure 5.6.





**Figure 5.6.** The interaction of the top hit ZINC01547088 with the active site of the Rv1430 esterase domain. Hydrogen bonding interactions are indicated in yellow dashed lines

#### 5.3.4 Molecular Dynamic Simulations

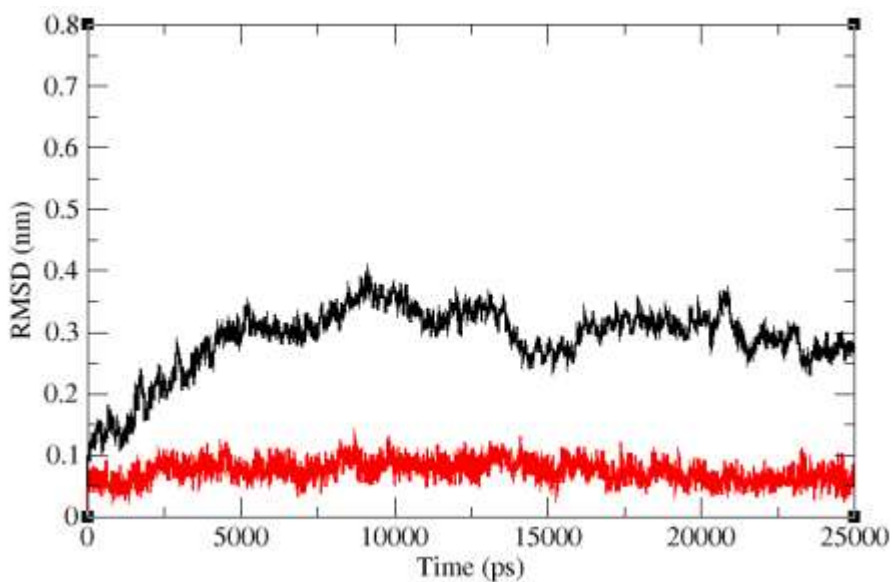
From molecular docking screening, we identified five molecules that make good non-bonding interactions with RV1430 esterase domain. These protein-inhibitor complexes were subjected up to 25 ns of MD simulations to check their structural stability and binding. Among the five initial molecules, three comparably showed weak interaction with active site residues during the MD simulations. Here the lid region of the protein is actively involved in the movement of the inhibitor. The two molecules ZINC16052749 and ZINC13681668, showed better interactions with the active site of Rv1430 esterase domain and retain non-bonding interactions with the enzyme, and displayed continuous hydrogen bond with Ser199. From the trajectory of the MD simulations, we observed the protein-inhibitor complexes and analyzed the movement of the inhibitor at the active site. Low RMSD values of these inhibitors and minor variations in the active site residues of the Rv1430 esterase domain were observed. These interactions were analyzed along with docking results. The results of the MD simulations and binding free energies of the two protein-inhibitor complexes are described below.

**ZINC16052749:** From the trajectory analyses, we have found low RMSD of the ligand in the active site and continuous hydrogen bond with Ser199. The protein RMSD fluctuated around 4 Å during the initial MD simulations but during the last 5 ns it converged near 3

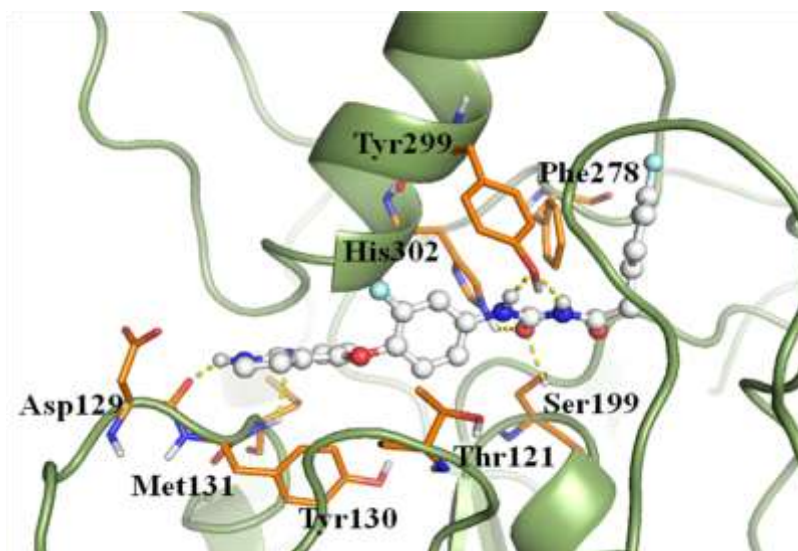
Å. The RMSD of the ligand initially converged and continued throughout the MD simulations at 1 Å of RMSD as shown in Figure 5.7.1. The sidechain hydroxy group of Ser199 forms a bifurcated hydrogen bond with two carbonyl oxygens of the inhibitor (OH...O1, 3.3 Å ; OH...O2, 2.7 Å) as shown in Figure 5.7.2. From the SIE free energy calculations as indicated in Table 5.3 the catalytic Ser199 has good Columbic interaction ( $-3.96 \pm 0.92$  kcal/mol) and less vdW interaction energy ( $-0.46 \pm 1.04$  kcal/mol) indicated the existence of continuous hydrogen bond. The NH of His302 forms a new hydrogen bond with the carbonyl oxygen of the inhibitor (NH...O1, 2.7 Å). The Columbic interactions ( $-2.47 \pm 0.99$  kcal/mol) and high vdW interaction ( $-4.19 \pm 0.92$  kcal/mol) energies indicated that the molecule aromatic ring is involved in the hydrophobic interactions. The 1H-Pyrrolo[2,3-b]pyridine NH of inhibitor forms bifurcated hydrogen bonds with the mainchain carbonyl oxygen (CO...N4, 1.8 Å) and sidechain carbonyl oxygen (CO...N4, 3.0 Å) of the Asp129. The sidechain hydroxy group of Tyr299 forms hydrogen bond with the NH of the inhibitor (HO...HN1, 2.4 Å). The Tyr299 residue showed high vdW ( $-4.75 \pm 1.24$  kcal/mol) and Columbic ( $-1.24 \pm 0.93$  kcal/mol) interaction energy values, and these values indicated that the phenyl ring of the Tyr299 stabilizes the molecule in the active site. Based on the SIE calculations, the contribution from Phe157 to Coloumbic and vdW energies are low, indicating its less participation, this is supported by the structural trajectories that indicated the loss of pi-pi stacking interactions from Phe157. The high vdW contribution from the residues Thr121 ( $-3.79 \pm 0.79$  kcal/mol), Tyr130 ( $-3.84 \pm 0.68$  kcal/mol) and Phe278 ( $-3.05 \pm 0.66$  kcal/mol) is an indicative that greater hydrophobic interactions stabilize the inhibitor binding in this location.

**Table 5.3.** The results of the SIE free energy (kcal/mol) calculations of ZINC16052749 in the active site of the Rv1430 esterase domain in terms of columbic interaction vdW interaction energies.

Amino acid	Inter VdW+ stdev (kcal/mol)	Inter Coulmb + stdev (kcal/mol)
Phe157	$-0.51 \pm 0.31$	$0.31 \pm 0.11$
Phe198	$-2.84 \pm 0.62$	$-0.41 \pm 0.27$
Ser199	$-0.46 \pm 1.04$	$-3.96 \pm 0.92$
Gln200	$-2.52 \pm 0.72$	$-2.28 \pm 0.52$
Tyr299	$-4.75 \pm 1.24$	$-1.24 \pm 0.93$
Thr121	$-3.79 \pm 0.79$	$-0.04 \pm 0.52$
Tyr130	$-3.84 \pm 0.68$	$-1.25 \pm 0.28$
His302	$-4.19 \pm 0.92$	$-2.47 \pm 0.99$
Pro262	$-1.38 \pm 0.39$	$-0.18 \pm 0.10$
Phe278	$-3.05 \pm 0.66$	$-0.47 \pm 0.29$
<b>Total protein with ZINC16052749</b>	<b><math>-54.10 \pm 3.11</math></b>	<b><math>-17.57 \pm 2.19</math></b>



**Figure 5.7.1.** The RMSD plot of Rv1430 esterase domain (black) in complex with the inhibitor molecule ZINC16052749 (red)



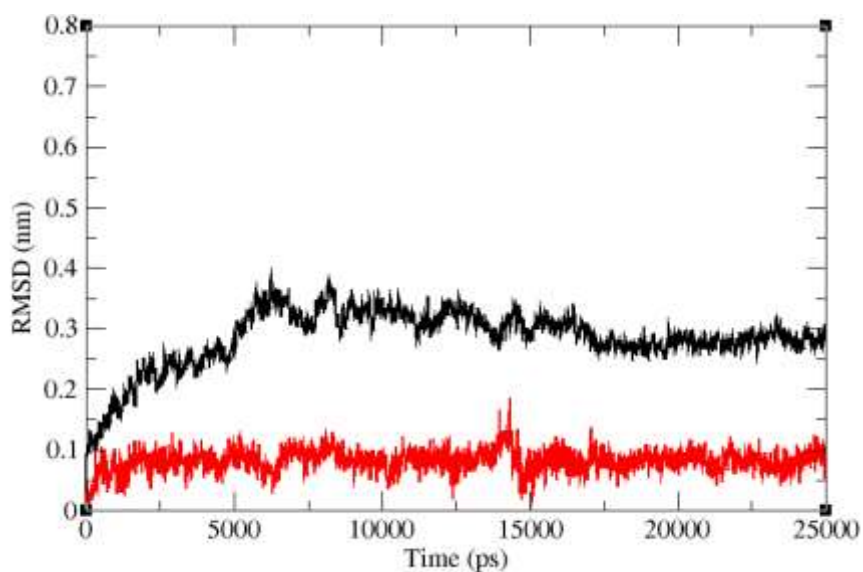
**Figure 5.7.2.** The interaction of the best hit ZINC16052749 with the active site of the Rv1430 esterase domain from MD simulations. Hydrogen bonding interactions are indicated in yellow dashed lines

**ZINC13681668:** We observed that the inhibitor is stabilized in the active site throughout the MD simulations. The RMSD of the protein converged to 3 Å from 16 ns to the end of the simulations, and the ligand RMSD is converged at 1 Å throughout the simulations. A small movement of the inhibitor is made to strengthen the bonding with protein and more number of hydrogen bonds were observed as shown in Figure 5.8.1. The tetrazine nitrogen forms new bifurcated hydrogen bonds with sidechain OH (OH...N2, 3.1 Å) and mainchain NH (NH...N2, 3.2 Å) of Thr121 as shown in Figure 5.8.2. We observed the molecule forming hydrogen bond with Ser199 is disturbed some times due to formation of the hydrogen bond with Thr121.. The SIE free energy calculations also indicated that Thr121 has high Coloumbic energy value ( $-4.05 \pm 0.86$  kcal/mol) whereas the Ser199 showed less Coloumbic interactions ( $0.68 \pm 0.25$  kcal/mol). One of the oxygens of inhibitor forms hydrogen bond with the sidechain OH of Tyr168 (OH...O3, 2.9 Å). One of the tetrazine nitrogen forms hydrogen bond with the sidechain nitrogen of His302 (N...HN4, 2.9 Å). The sidechains of Leu295 and Leu296 also exhibit good hydrophobic interactions and vdW contribution ( $-3.01 \pm 0.67$  kcal/mol and  $-3.43 \pm 0.88$  kcal/mol respectively ) to the free energy of binding. The Tyr299 and Phe157 form pi-pi stacking interactions with the aromatic rings of the inhibitors to hold the molecule in the active site. The high vdW values of Phe157 ( $-3.43 \pm 1.77$  kcal/mol) and Tyr299 ( $-8.51 \pm 1.28$

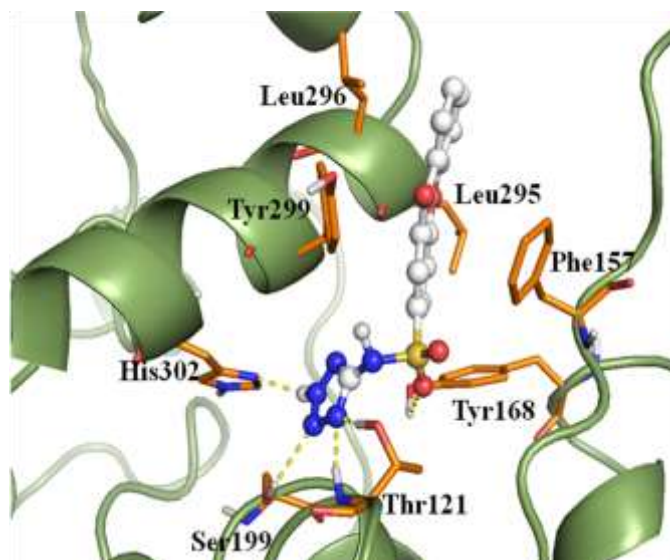
kcal/mol) also indicated the hydrophobic nature of interactions in stabilizing the protein-inhibitor complex as shown in Table 5.4.

**Table 5.4.** The results of the SIE free energy (kcal/mol) calculations of ZINC13681668 in the active site of the Rv1430 esterase domain in terms of columbic interaction vdW interaction energies.

Amino acid	Inter VdW+ stdev (kcal/mol)	Inter Coulmb + stdev (kcal/mol)
Phe157	$-3.43 \pm 1.77$	$0.06 \pm 0.13$
Ser120	$-1.75 \pm 0.34$	$-1.02 \pm 0.26$
Ser199	$-1.14 \pm 0.21$	$0.68 \pm 0.25$
Gln200	$-2.23 \pm 0.68$	$-0.91 \pm 0.71$
Tyr299	$-8.51 \pm 1.28$	$-0.18 \pm 0.32$
Thr121	$-1.90 \pm 1.51$	$-4.05 \pm 0.86$
Tyr168	$-1.59 \pm 1.20$	$-2.74 \pm 1.37$
His302	$-1.03 \pm 1.00$	$-2.77 \pm 1.67$
Leu295	$-3.01 \pm 0.67$	$0.08 \pm 0.31$
Leu296	$-3.43 \pm 0.88$	$0.15 \pm 0.11$
<b>Total protein with ZINC13681668</b>	<b><math>-36.18 \pm 3.87</math></b>	<b><math>-11.46 \pm 2.43</math></b>



**Figure 5.8.1.** The RMSD plot of Rv1430 esterase domain (black) in complex with the inhibitor molecule ZINC13681668 (red)



**Figure 5.8.2.** The interaction of the best hit ZINC13681668 with the active site of the Rv1430 esterase domain from MD simulations. Hydrogen bonding interactions are indicated in yellow dashed lines

The importance of PE and PPE proteins in *Mtb* along with the significant role played by cutinases/esterases/lipases in the cell wall lipid metabolism of pathogenic *mycobacteria* strongly implicates that the PE-PPE domain is a biological receptor to design new anti-TB potential drug candidates. We have therefore used *in silico* screening of large databases of molecules and molecular docking techniques to select new and potent Rv1430 PE-PPE domain inhibitors for better drug candidates. The inhibitors from ZINC database were docked into the model structure of Rv1430 PE-PPE domain, which resulted in five best ligands ZINC13681668, ZINC16052749, ZINC01547088, ZINC16052239 and ZINC16052883 that have good binding affinity and interactions in the active site region of the protein. To confirm the stability of their binding and energetics, we have carried out MD simulations of the protein-inhibitor complexes. MD simulations of the above five protein-ligand complexes resulted in the identification of two molecules, ZINC16052749 and ZINC13681668 that are stable and have better interactions with the active site of Rv1430 esterase. These results provide more insights for the design of better inhibitors to Rv1430 PE-PPE domain with increased binding affinities there by increasing the inhibitory activity. We therefore propose ZINC16052749 and ZINC13681668 as good inhibitors of this TB drug target. Using MD simulations it was observed that protein with inhibitors ZINC16052749 and ZINC13681668 stabilized in the active site. Since hydrolases are required for lipid metabolism, ZINC16052749 and ZINC13681668 can be used for inhibiting the lipid metabolism pathway of *Mtb*.

## 5.4 Conclusions

The present chapter extends our work on Rv1430 esterase domain from earlier chapters using the computational approaches. The inhibitors from ZINC database were docked into the Rv1430 esterase domain of Rv1430 protein, which resulted in five best inhibitors ZINC13681668, ZINC16052749, ZINC01547088, ZINC16052239, ZINC16052883 that have good interaction and binding into the active site region of the protein. Further MD simulations of the above five protein-inhibitors complexes resulted in two molecules ZINC16052749 and ZINC13681668 that showed better interactions with the active site of Rv1430 esterase domain. These results provide more insights for the design of better inhibitors to Rv1430 esterase domain that has increased the binding affinities there by increasing the inhibitory activity.





## References

- Abdallah, A.M., Verboom, T., Weerdenburg, E.M., Gey van Pittius, N.C., Mahasha, P.W., Jimenez, C., Parra, M., Cadieux, N., Brennan, M.J., Appelmelk, B.J., Bitter, W., 2009. PPE and PE\_PGRS proteins of *Mycobacterium marinum* are transported via the type VII secretion system ESX-5. *Mol Microbiol* 73, 329-340.
- Abraham, P.R., Latha, G.S., Valluri, V.L., Mukhopadhyay, S., 2014. *Mycobacterium tuberculosis* PPE protein Rv0256c induces strong B cell response in tuberculosis patients. *Infect Genet Evol* 22, 244-249.
- Adindla, S., Guruprasad, L., 2003. Sequence analysis corresponding to the PPE and PE proteins in *Mycobacterium tuberculosis* and other genomes. *J Biosci* 28, 169-179.
- Adindla, S., Inampudi, K.K., Guruprasad, K., Guruprasad, L., 2004. Identification and analysis of novel tandem repeats in the cell surface proteins of archaeal and bacterial genomes using computational tools. *Comp Funct Genomics* 5, 2-16.
- Adindla, S., Inampudi, K.K., Guruprasad, L., 2007. Cell surface proteins in archaeal and bacterial genomes comprising "LVIVD", "RIVW" and "LGxL" tandem sequence repeats are predicted to fold as beta-propeller. *International journal of biological macromolecules* 41, 454-468.
- Akhter, Y., Ehebauer, M.T., Mukhopadhyay, S., Hasnain, S.E., 2012. The PE/PPE multigene family codes for virulence factors and is a possible source of mycobacterial antigenic variation: perhaps more? *Biochimie* 94, 110-116.
- Alonso, H., Bliznyuk, A.A., Gready, J.E., 2006. Combining docking and molecular dynamic simulations in drug design. *Med Res Rev* 26, 531-568.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J Mol Biol* 215, 403-410.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-3402.
- Anand, P., Sankaran, S., Mukherjee, S., Yeturu, K., Laskowski, R., Bhardwaj, A., Bhagavat, R., Brahmachari, S.K., Chandra, N., 2011. Structural annotation of *Mycobacterium tuberculosis* proteome. *PLoS One* 6, e27044.
- Anand, R., 2016. Identification of Potential Antituberculosis Drugs Through Docking and Virtual Screening. *Interdisciplinary sciences, computational life sciences*.
- Andersen, P., Doherty, T.M., 2005. The success and failure of BCG - implications for a novel tuberculosis vaccine. *Nat Rev Microbiol* 3, 656-662.
- Arora, A., Chandra, N.R., Das, A., Gopal, B., Mande, S.C., Prakash, B., Ramachandran, R., Sankaranarayanan, R., Sekar, K., Suguna, K., Tyagi, A.K., Vijayan, M., 2011. Structural biology of *Mycobacterium tuberculosis* proteins: the Indian efforts. *Tuberculosis (Edinb)* 91, 456-468.
- Balaji, K.N., Goyal, G., Narayana, Y., Srinivas, M., Chaturvedi, R., Mohammad, S., 2007. Apoptosis triggered by Rv1818c, a PE family gene from *Mycobacterium tuberculosis* is regulated by mitochondrial intermediates in T cells. *Microbes Infect* 9, 271-281.
- Banu, S., Honore, N., Saint-Joanis, B., Philpott, D., Prevost, M.C., Cole, S.T., 2002. Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens? *Mol Microbiol* 44, 9-19.
- Barathy, D.V., Suguna, K., 2013. Crystal structure of a putative aspartic proteinase domain of the *Mycobacterium tuberculosis* cell surface antigen PE\_PGRS16. *FEBS Open Bio* 3, 256-262.
- Barry, C.E., 3rd, 2001. Interpreting cell wall 'virulence factors' of *Mycobacterium tuberculosis*. *Trends Microbiol* 9, 237-241.
- Beatty, W.L., Russell, D.G., 2000. Identification of mycobacterial surface proteins released into subcellular compartments of infected macrophages. *Infect Immun* 68, 6997-7002.
- Belisle, J.T., Vissa, V.D., Sievert, T., Takayama, K., Brennan, P.J., Besra, G.S., 1997. Role of the major antigen of *Mycobacterium tuberculosis* in cell wall biogenesis. *Science* 276, 1420-1422.

Bentley, S.D., Comas, I., Bryant, J.M., Walker, D., Smith, N.H., Harris, S.R., Thurston, S., Gagneux, S., Wood, J., Antonio, M., Quail, M.A., Gehre, F., Adegbola, R.A., Parkhill, J., de Jong, B.C., 2012. The genome of *Mycobacterium africanum* West African 2 reveals a lineage-specific locus and genome erosion common to the *M. tuberculosis* complex. *PLoS Negl Trop Dis* 6, e1552.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. *Nucleic Acids Res* 28, 235-242.

Bermudez, L.E., Goodman, J., 1996. *Mycobacterium tuberculosis* invades and replicates within type II alveolar cells. *Infect Immun* 64, 1400-1406.

Berry, M., Kon, O.M., 2009. Multidrug- and extensively drug-resistant tuberculosis: an emerging threat. *Eur Respir Rev* 18, 195-197.

Berry, M.P., Blankley, S., Graham, C.M., Bloom, C.I., O'Garra, A., 2013. Systems approaches to studying the immune response in tuberculosis. *Curr Opin Immunol* 25, 579-587.

Blundell, T.L., Sibanda, B.L., Sternberg, M.J., Thornton, J.M., 1987. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 326, 347-352.

Bohm, G., Muhr, R., Jaenicke, R., 1992. Quantitative analysis of protein far UV circular dichroism spectra by neural networks. *Protein Eng* 5, 191-195.

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

Bradford, M.M., 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* 72, 248-254.

Brennan, M.J., Delogu, G., 2002. The PE multigene family: a 'molecular mantra' for mycobacteria. *Trends Microbiol* 10, 246-249.

Brennan, M.J., Delogu, G., Chen, Y., Bardarov, S., Kriakov, J., Alavi, M., Jacobs, W.R., Jr., 2001. Evidence that mycobacterial PE\_PGRS proteins are cell surface constituents that influence interactions with other cells. *Infect Immun* 69, 7326-7333.

Brennan, P.J., 2003. Structure, function, and biogenesis of the cell wall of *Mycobacterium tuberculosis*. *Tuberculosis (Edinb)* 83, 91-97.

Brennan, P.J.D., P, 1994. Ultrastructure of *Mycobacterium tuberculosis*. In Bloom, B.R. (ed.), *Tuberculosis: pathogenesis, protection, and control*. American Society for Microbiology, Washington DC, pp. 271-284.

Brod, F.C., Pelisser, M.R., Bertoldo, J.B., Vernal, J., Bloch, C., Jr., Terenzi, H., Arisi, A.C., 2010a. Heterologous Expression and Purification of a Heat-Tolerant *Staphylococcus xylosus* Lipase. *Mol Biotechnol* 44, 110-119.

Brod, F.C., Vernal, J., Bertoldo, J.B., Terenzi, H., Arisi, A.C., 2010b. Cloning, expression, purification, and characterization of a novel esterase from *Lactobacillus plantarum*. *Mol Biotechnol* 44, 242-249.

Brohawn, S.G., Schwartz, T.U., 2009. Molecular architecture of the Nup84-Nup145C-Sec13 edge element in the nuclear pore complex lattice. *Nature structural & molecular biology* 16, 1173-1177.

Brosch, R., Gordon, S.V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., Parsons, L.M., Pym, A.S., Samper, S., van Soolingen, D., Cole, S.T., 2002. A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proceedings of the National Academy of Sciences of the United States of America* 99, 3684-3689.

Brown, K., Djinić-Carugo, K., Haltia, T., Cabrito, I., Saraste, M., Moura, J.J., Moura, I., Tegoni, M., Cambillau, C., 2000a. Revisiting the catalytic CuZ cluster of nitrous oxide (N<sub>2</sub>O) reductase. Evidence of a bridging inorganic sulfur. *J Biol Chem* 275, 41133-41136.

Brown, K., Tegoni, M., Prudencio, M., Pereira, A.S., Besson, S., Moura, J.J., Moura, I., Cambillau, C., 2000b. A novel type of catalytic copper cluster in nitrous oxide reductase. *Nat Struct Biol* 7, 191-195.

Browne, W.J., North, A.C., Phillips, D.C., Brew, K., Vanaman, T.C., Hill, R.L., 1969. A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. *J Mol Biol* 42, 65-86.

Bussi, G., Donadio, D., Parrinello, M., 2007. Canonical sampling through velocity rescaling. *J Chem Phys* 126, 014101.

Calmette, A., 1922. L'infection bacillaire et la tuberculose chez l'homme et chez les animaux.

Camacho, L.R., Ensergueix, D., Perez, E., Gicquel, B., Guilhot, C., 1999. Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. *Mol Microbiol* 34, 257-267.

Camus, J.C., Pryor, M.J., Medigue, C., Cole, S.T., 2002. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology* 148, 2967-2973.

Canaan, S., Maurin, D., Chahinian, H., Pouilly, B., Dourousseau, C., Frassinetti, F., Scappuccini-Calvo, L., Cambillau, C., Bourne, Y., 2004. Expression and characterization of the protein Rv1399c from *Mycobacterium tuberculosis*. A novel carboxyl esterase structurally related to the HSL family. *Eur J Biochem* 271, 3953-3961.

Cascioferro, A., Daleke, M.H., Ventura, M., Dona, V., Delogu, G., Palu, G., Bitter, W., Manganelli, R., 2011. Functional dissection of the PE domain responsible for translocation of PE\_PGRS33 across the mycobacterial cell wall. *PLoS One* 6, e27713.

Cascioferro, A., Delogu, G., Colone, M., Sali, M., Stringaro, A., Arancia, G., Fadda, G., Palu, G., Manganelli, R., 2007. PE is a functional domain responsible for protein translocation and localization on mycobacterial cell wall. *Mol Microbiol* 66, 1536-1547.

Chaitra, M.G., Shaila, M.S., Nayak, R., 2008. Characterization of T-cell immunogenicity of two PE/PPE proteins of *Mycobacterium tuberculosis*. *J Med Microbiol* 57, 1079-1086.

Chan, J., Fan, X.D., Hunter, S.W., Brennan, P.J., Bloom, B.R., 1991. Lipoarabinomannan, a possible virulence factor involved in persistence of *Mycobacterium tuberculosis* within macrophages. *Infect Immun* 59, 1755-1761.

Chatterjee, D., Hunter, S.W., McNeil, M., Brennan, P.J., 1992. Lipoarabinomannan. Multiglycosylated form of the mycobacterial mannosylphosphatidylinositols. *J Biol Chem* 267, 6228-6233.

Chen, C.K., Chan, N.L., Wang, A.H., 2011. The many blades of the beta-propeller proteins: conserved but versatile. *Trends in biochemical sciences* 36, 553-561.

Chen, S., Tong, X., Woodard, R.W., Du, G., Wu, J., Chen, J., 2008. Identification and characterization of bacterial cutinase. *J Biol Chem* 283, 25854-25862.

Chen, W., Chang, C.E., Gilson, M.K., 2004. Calculation of cyclodextrin binding affinities: energy, entropy, and implications for drug design. *Biophys J* 87, 3035-3049.

Cherukuvada, S.L., Seshasayee, A.S., Raghunathan, K., Anishetty, S., Pennathur, G., 2005. Evidence of a double-lid movement in *Pseudomonas aeruginosa* lipase: insights from molecular dynamics simulations. *PLoS Comput Biol* 1, e28.

Choudhary, R.K., Mukhopadhyay, S., Chakhaiyar, P., Sharma, N., Murthy, K.J., Katoch, V.M., Hasnain, S.E., 2003. PPE antigen Rv2430c of *Mycobacterium tuberculosis* induces a strong B-cell response. *Infect Immun* 71, 6338-6343.

Choudhary, R.K., Pullakhandam, R., Ehtesham, N.Z., Hasnain, S.E., 2004. Expression and characterization of Rv2430c, a novel immunodominant antigen of *Mycobacterium tuberculosis*. *Protein Expr Purif* 36, 249-253.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E., 3rd, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M.A., Rajandream, M.A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J.E., Taylor, K., Whitehead, S., Barrell, B.G., 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537-544.

Cole, S.T., Telenti, A., 1995. Drug resistance in *Mycobacterium tuberculosis*. *Eur Respir J Suppl* 20, 701s-713s.

Colman, P.M., 1994. Influenza virus neuraminidase: structure, antibodies, and inhibitors. *Protein Sci* 3, 1687-1696.

Cooper, J.B., 2002. Aspartic proteinases in disease: a structural perspective. *Current drug targets* 3, 155-173.

Cotes, K., Dhoub, R., Douchet, I., Chahinian, H., de Caro, A., Carriere, F., Canaan, S., 2007. Characterization of an exported monoglyceride lipase from *Mycobacterium tuberculosis* possibly involved in the metabolism of host cell membrane lipids. *Biochem J* 408, 417-427.

Cruz-Knight, W., Blake-Gumbs, L., 2013. Tuberculosis: an overview. *Prim Care* 40, 743-756.

Cryle, M.J., 2011. Carrier protein substrates in cytochrome P450-catalysed oxidation. *Metallomics : integrated biometal science* 3, 323-326.

Cui, Q., Sulea, T., Schrag, J.D., Munger, C., Hung, M.N., Naim, M., Cygler, M., Purisima, E.O., 2008. Molecular dynamics-solvated interaction energy studies of protein-protein interactions: the MP1-p14 scaffolding complex. *J Mol Biol* 379, 787-802.

Daffe, M., Draper, P., 1998. The envelope layers of mycobacteria with reference to their pathogenicity. *Adv Microb Physiol* 39, 131-203.

Daleke, M.H., Cascioferro, A., de Punder, K., Ummels, R., Abdallah, A.M., van der Wel, N., Peters, P.J., Luirink, J., Manganelli, R., Bitter, W., 2011. Conserved Pro-Glu (PE) and Pro-Pro-Glu (PPE) protein domains target LipY lipases of pathogenic mycobacteria to the cell surface via the ESX-5 pathway. *J Biol Chem* 286, 19024-19034.

Daniel, J., Deb, C., Dubey, V.S., Sirakova, T.D., Abomoelak, B., Morbidoni, H.R., Kolattukudy, P.E., 2004. Induction of a novel class of diacylglycerol acyltransferases and triacylglycerol accumulation in *Mycobacterium tuberculosis* as it goes into a dormancy-like state in culture. *J Bacteriol* 186, 5017-5030.

Danielson, P.B., 2002. The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. *Current drug metabolism* 3, 561-597.

Darden T, Y.D., Pedersen L, 1993. Particle mesh Ewald: an Nlog (N) method for Ewald sums in large systems *J Chem Phys* 98, 10089–10092.

Davies, G., Henrissat, B., 1995. Structures and mechanisms of glycosyl hydrolases. *Structure* 3, 853-859.

Davies, G., Sinnott, M. L., and Withers, S. G., 1997. *Comprehensive Biological Catalysis*. Sinnott, M. L., Ed Academic Press, London.

De Simone, G., Mandrich, L., Menchise, V., Giordano, V., Febbraio, F., Rossi, M., Pedone, C., Manco, G., 2004. A substrate-induced switch in the reaction mechanism of a thermophilic esterase: kinetic evidences and structural basis. *J Biol Chem* 279, 6815-6823.

Deb, C., Daniel, J., Sirakova, T.D., Abomoelak, B., Dubey, V.S., Kolattukudy, P.E., 2006. A novel lipase belonging to the hormone-sensitive lipase family induced under starvation to utilize stored triacylglycerol in *Mycobacterium tuberculosis*. *J Biol Chem* 281, 3866-3875.

Delogu, G., Brennan, M.J., 2001. Comparative immune response to PE and PE\_PGRS antigens of *Mycobacterium tuberculosis*. *Infect Immun* 69, 5606-5611.

Delogu, G., Pusceddu, C., Bua, A., Fadda, G., Brennan, M.J., Zanetti, S., 2004. Rv1818c-encoded PE\_PGRS protein of *Mycobacterium tuberculosis* is surface exposed and influences bacterial cell structure. *Mol Microbiol* 52, 725-733.

Deng, W., Li, W., Zeng, J., Zhao, Q., Li, C., Zhao, Y., Xie, J., 2014. *Mycobacterium tuberculosis* PPE family protein Rv1808 manipulates cytokines profile via co-activation of MAPK and NF-kappaB signaling pathways. *Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology* 33, 273-288.

Dheenadhayalan, V., Delogu, G., Brennan, M.J., 2006. Expression of the PE\_PGRS 33 protein in *Mycobacterium smegmatis* triggers necrosis in macrophages and enhanced mycobacterial survival. *Microbes Infect* 8, 262-272.

Din, N., Damude, H.G., Gilkes, N.R., Miller, R.C., Jr., Warren, R.A., Kilburn, D.G., 1994. C1-Cx revisited: intramolecular synergism in a cellulase. *Proceedings of the National Academy of Sciences of the United States of America* 91, 11383-11387.

Din N, G.N., Tekant B, Miller RC Jr, Warren RAJ & Kilburn DG, 1991. Non-Hydrolytic Disruption of Cellulose Fibres by the Binding Domain of a Bacterial Cellulase. *Nature Biotechnology* 9, 1096-1099.

Doxey, A.C., Cheng, Z., Moffatt, B.A., McConkey, B.J., 2010. Structural motif screening reveals a novel, conserved carbohydrate-binding surface in the pathogenesis-related protein PR-5d. *BMC structural biology* 10, 23.

Du Plessis, E.M., Berger, E., Stark, T., Louw, M.E., Visser, D., 2010. Characterization of a novel thermostable esterase from *Thermus scotoductus* SA-01: evidence of a new family of lipolytic esterases. *Curr Microbiol* 60, 248-253.

Dubnau, E., Fontan, P., Manganelli, R., Soares-Appel, S., Smith, I., 2002. *Mycobacterium tuberculosis* genes induced during infection of human macrophages. *Infect Immun* 70, 2787-2795.

Dunn, B., 1989. "Determination of protease mechanism". *Proteolytic Enzymes a practical approach*. eds. R. Beynon and J. Bond. Oxford University Press, oxford.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32, 1792-1797.

Efron, B., Halloran, E., Holmes, S., 1996. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences of the United States of America* 93, 13429-13434.

Ekiert, D.C., Cox, J.S., 2014. Structure of a PE-PPE-EspG complex from *Mycobacterium tuberculosis* reveals molecular specificity of ESX protein secretion. *Proceedings of the National Academy of Sciences of the United States of America* 111, 14758-14763.

Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H., 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* 2, 953-971.

Felsenstein, J., 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Society for the Study of Evolution* 39, 783-791.

Fett WF, W.C., Moreau RA, Osman SF 1999. Production of cutinases by *Thermomonospora fusca* ATCC27730. *J Appl Microbiol* 86, 561-568.

Fishbein, S., van Wyk, N., Warren, R.M., Sampson, S.L., 2015. Phylogeny to function: PE/PPE protein evolution and impact on *Mycobacterium tuberculosis* pathogenicity. *Mol Microbiol* 96, 901-916.

Fogel, N., 2015. *Tuberculosis: A disease without boundaries*. Tuberculosis (Edinb).

Fonseca, A.L., Nunes, R.R., Braga, V.M., Comar, M., Jr., Alves, R.J., Varotti, F.P., Taranto, A.G., 2016. Docking, QM/MM, and molecular dynamics simulations of the hexose transporter from *Plasmodium falciparum* (PfHT). *J Mol Graph Model* 66, 174-186.

Garnier, T., Eiglmeier, K., Camus, J.C., Medina, N., Mansoor, H., Pryor, M., Duthoy, S., Grondin, S., Lacroix, C., Monsempe, C., Simon, S., Harris, B., Atkin, R., Doggett, J., Mayes, R., Keating, L., Wheeler, P.R., Parkhill, J., Barrell, B.G., Cole, S.T., Gordon, S.V., Hewinson, R.G., 2003. The complete genome sequence of *Mycobacterium bovis*. *Proceedings of the National Academy of Sciences of the United States of America* 100, 7877-7882.

Garton, N.J., Christensen, H., Minnikin, D.E., Adegbola, R.A., Barer, M.R., 2002. Intracellular lipophilic inclusions of mycobacteria in vitro and in sputum. *Microbiology* 148, 2951-2958.

Gengenbacher, M., Kaufmann, S.H., 2012. *Mycobacterium tuberculosis*: success through dormancy. *FEMS Microbiol Rev* 36, 514-532.

Gey van Pittius, N.C., Sampson, S.L., Lee, H., Kim, Y., van Helden, P.D., Warren, R.M., 2006. Evolution and expansion of the *Mycobacterium tuberculosis* PE and PPE multigene families and their association with the duplication of the ESAT-6 (*esx*) gene cluster regions. *BMC Evol Biol* 6, 95.

Glickman, M.S., Jacobs, W.R., Jr., 2001. Microbial pathogenesis of *Mycobacterium tuberculosis*: dawn of a discipline. *Cell* 104, 477-485.

Goldstone, R.M., Goonesekera, S.D., Bloom, B.R., Sampson, S.L., 2009. The transcriptional regulator Rv0485 modulates the expression of a *pe* and *ppe* gene pair and is required for *Mycobacterium tuberculosis* virulence. *Infect Immun* 77, 4654-4667.

Gordon, E.H., Sjogren, T., Lofqvist, M., Richter, C.D., Allen, J.W., Higham, C.W., Hajdu, J., Fulop, V., Ferguson, S.J., 2003. Structure and kinetic properties of *Paracoccus pantotrophus* cytochrome *cd1* nitrite reductase with the *d1* heme active site ligand tyrosine 25 replaced by serine. *J Biol Chem* 278, 11773-11781.

Gordon, S.V., Eiglmeier, K., Garnier, T., Brosch, R., Parkhill, J., Barrell, B., Cole, S.T., Hewinson, R.G., 2001. Genomics of *Mycobacterium bovis*. *Tuberculosis (Edinb)* 81, 157-163.

Grochulski, P., Li, Y., Schrag, J.D., Cygler, M., 1994. Two conformational states of *Candida rugosa* lipase. *Protein Sci* 3, 82-91.

Guo, J., Zheng, X., Xu, L., Liu, Z., Xu, K., Li, S., Wen, T., Liu, S., Pang, H., 2010. Characterization of a novel esterase Rv0045c from *Mycobacterium tuberculosis*. *PLoS One* 5.

Henrissat, B., 1991. A Classification of Glycosyl Hydrolases Based on Amino-Acid-Sequence Similarities. *Biochemical Journal* 280, 309-316.

Henrissat, B., Bairoch, A., 1996. Updating the sequence-based classification of glycosyl hydrolases. *Biochem J* 316 ( Pt 2), 695-696.

Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J.P., Davies, G., 1995. Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. *Proceedings of the National Academy of Sciences of the United States of America* 92, 7090-7094.

Henrissat, B., Davies, G., 1997. Structural and sequence-based classification of glycoside hydrolases. *Current opinion in structural biology* 7, 637-644.

Hermans, P.W., van Soolingen, D., van Embden, J.D., 1992. Characterization of a major polymorphic tandem repeat in *Mycobacterium tuberculosis* and its potential use in the epidemiology of *Mycobacterium kansasii* and *Mycobacterium gordonae*. *J Bacteriol* 174, 4157-4165.

Hess, B., Bekker, H., Berendsen, H.J.C., Fraaije, J.G.E.M., 1997. LINCS: A linear constraint solver for molecular simulations. *J Comput Chem* 18, 1463-1472.

Hess, B., Kutzner, C., van der Spoel, D., Lindahl, E., 2008. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* 4, 435-447.

Ho, S.N., Hunt, H.D., Horton, R.M., Pullen, J.K., Pease, L.R., 1989. Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene* 77, 51-59.

Hotelier, T., Renault, L., Cousin, X., Negre, V., Marchot, P., Chatonnet, A., 2004. ESTHER, the database of the alpha/beta-hydrolase fold superfamily of proteins. *Nucleic Acids Res* 32, D145-147.

Huggins, D.J., Venkitaraman, A.R., Spring, D.R., 2011. Rational methods for the selection of diverse screening compounds. *ACS chemical biology* 6, 208-217.

Hur, G.H., Vickery, C.R., Burkart, M.D., 2012. Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. *Natural product reports* 29, 1074-1098.

Ilinkin, I., Ye, J., Janardan, R., 2010. Multiple structure alignment and consensus identification for proteins. *BMC Bioinformatics* 11, 71.

Irwin, J.J., Shoichet, B.K., 2005. ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 45, 177-182.

Irwin, J.J., Sterling, T., Mysinger, M.M., Bolstad, E.S., Coleman, R.G., 2012. ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* 52, 1757-1768.

Iseman, M.D., 2002. Tuberculosis therapy: past, present and future. *Eur Respir J Suppl* 36, 87s-94s.

Ishikawa, J., Yamashita, A., Mikami, Y., Hoshino, Y., Kurita, H., Hotta, K., Shiba, T., Hattori, M., 2004. The complete genomic sequence of *Nocardia farcinica* IFM 10152. *Proceedings of the National Academy of Sciences of the United States of America* 101, 14925-14930.

James, J.J., Lakshmi, B.S., Raviprasad, V., Ananth, M.J., Kanguane, P., Gautam, P., 2003. Insights from molecular dynamics simulations into pH-dependent enantioselective hydrolysis of ibuprofen esters by *Candida rugosa* lipase. *Protein Eng* 16, 1017-1024.

James, J.J., Lakshmi, B.S., Seshasayee, A.S., Gautam, P., 2007. Activation of *Candida rugosa* lipase at alkane-aqueous interfaces: a molecular dynamics study. *FEBS Lett* 581, 4377-4383.

Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G., Gibson, T.J., 1998. Multiple sequence alignment with Clustal X. *Trends in biochemical sciences* 23, 403-405.

Johnson, P.E., Joshi, M.D., Tomme, P., Kilburn, D.G., McIntosh, L.P., 1996. Structure of the N-terminal cellulose-binding domain of *Cellulomonas fimi* CenC determined by nuclear magnetic resonance spectroscopy. *Biochemistry* 35, 14381-14394.

Junie B. Billones, M.C.O.C., Voltaire G. Organo, Stephani Joy Y. Macalino, Inno A. Emnacen and Jamie Bernadette A. Sy, 2013. Virtual Screening Against *Mycobacterium tuberculosis* Druggable Targets and In Silico ADMET Evaluation of Top Hits. *ORIENTAL JOURNAL OF CHEMISTRY* 29, 1457-1468.

Karplus, M., McCammon, J.A., 2002. Molecular dynamics simulations of biomolecules. *Nat Struct Biol* 9, 646-652.

Kaufmann, S.H.E., Helden, P.V., 2008. *Handbook of Tuberculosis: Clinics, Diagnostics, Therapy, and Epidemiology*. Wiley-Blackwell.

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., Sternberg, M.J., 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 10, 845-858.

Kendall, B.A., Varley, C.D., Choi, D., Cassidy, P.M., Hedberg, K., Ware, M.A., Winthrop, K.L., 2011. Distinguishing tuberculosis from nontuberculous mycobacteria lung disease, Oregon, USA. *Emerg Infect Dis* 17, 506-509.

Keshavjee, S., Farmer, P.E., 2012. Tuberculosis, drug resistance, and the history of modern medicine. *N Engl J Med* 367, 931-936.

Keshavjee, S., Gelmanova, I.Y., Farmer, P.E., Mishustin, S.P., Strelis, A.K., Andreev, Y.G., Pasechnikov, A.D., Atwood, S., Mukherjee, J.S., Rich, M.L., Furin, J.J., Nardell, E.A., Kim, J.Y., Shin, S.S., 2008. Treatment of extensively drug-resistant tuberculosis in Tomsk, Russia: a retrospective cohort study. *Lancet* 372, 1403-1409.

Khan, N., Alam, K., Nair, S., Valluri, V.L., Murthy, K.J., Mukhopadhyay, S., 2008. Association of strong immune responses to PPE protein Rv1168c with active tuberculosis. *Clin Vaccine Immunol* 15, 974-980.

Kim, B.J., Choi, B.S., Lim, J.S., Choi, I.Y., Lee, J.H., Chun, J., Kook, Y.H., 2012. Complete genome sequence of *Mycobacterium intracellulare* strain ATCC 13950(T). *J Bacteriol* 194, 2750.

Kohli, S., Singh, Y., Sharma, K., Mittal, A., Ehtesham, N.Z., Hasnain, S.E., 2012. Comparative genomic and proteomic analyses of PE/PPE multigene family of *Mycobacterium tuberculosis* H(3)(7)Rv and H(3)(7)Ra reveal novel and interesting differences with implications in virulence. *Nucleic Acids Res* 40, 7113-7122.

Korotkova, N., Freire, D., Phan, T.H., Ummels, R., Creekmore, C.C., Evans, T.J., Wilmanns, M., Bitter, W., Parret, A.H., Houben, E.N., Korotkov, K.V., 2014. Structure of the *Mycobacterium tuberculosis* type VII secretion system chaperone EspG5 in complex with PE25-PPE41 dimer. *Mol Microbiol* 94, 367-382.

Koshland, D.E., 1953. Stereochemistry and the mechanism of enzymatic reactions. *Biol. Rev. Camb. Philos. Soc* 28, 416-436.

Koul, A., Herget, T., Klebl, B., Ullrich, A., 2004. Interplay between mycobacteria and host signalling pathways. *Nat Rev Microbiol* 2, 189-202.

Kremer, L., de Chastellier, C., Dobson, G., Gibson, K.J., Bifani, P., Balor, S., Gorvel, J.P., Locht, C., Minnikin, D.E., Besra, G.S., 2005. Identification and structural characterization of an unusual mycobacterial monomeromycetyl-diacylglycerol. *Mol Microbiol* 57, 1113-1126.

Kroemer, R.T., 2007. Structure-based drug design: docking and scoring. *Curr Protein Pept Sci* 8, 312-328.

Kruh, N.A., Troudt, J., Izzo, A., Prenni, J., Dobos, K.M., 2010. Portrait of a pathogen: the *Mycobacterium tuberculosis* proteome in vivo. *PLoS One* 5, e13938.

Kumar, S., Tamura, K., Nei, M., 1994. MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. *Comput Appl Biosci* 10, 189-191.

Kusner, D.J., 2005. Mechanisms of mycobacterial persistence in tuberculosis. *Clin Immunol* 114, 239-247.

Lamichhane, G., Tyagi, S., Bishai, W.R., 2005. Designer arrays for defined mutant analysis to detect genes essential for survival of *Mycobacterium tuberculosis* in mouse lungs. *Infect Immun* 73, 2533-2540.

Laskowski RA, M.M., Moss DS, Thornton JM 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *J App Cryst* 26, 283-291.

Li, L., Bannantine, J.P., Zhang, Q., Amonsin, A., May, B.J., Alt, D., Banerji, N., Kanjilal, S., Kapur, V., 2005a. The complete genome sequence of *Mycobacterium avium* subspecies paratuberculosis. *Proceedings of the National Academy of Sciences of the United States of America* 102, 12344-12349.

Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658-1659.

Li, Y., Miltner, E., Wu, M., Petrofsky, M., Bermudez, L.E., 2005b. A *Mycobacterium avium* PPE gene is associated with the ability of the bacterium to grow in macrophages and virulence in mice. *Cell Microbiol* 7, 539-548.

Lietzke, S.E., Yoder, M.D., Keen, N.T., Jurnak, F., 1994. The Three-Dimensional Structure of Pectate Lyase E, a Plant Virulence Factor from *Erwinia chrysanthemi*. *Plant physiology* 106, 849-862.

Lin, D.Y., Diao, J., Chen, J., 2012. Crystal structures of two bacterial HECT-like E3 ligases in complex with a human E2 reveal atomic details of pathogen-host interactions. *Proceedings of the National Academy of Sciences of the United States of America* 109, 1925-1930.

Lin, D.Y., Diao, J., Zhou, D., Chen, J., 2011. Biochemical and structural studies of a HECT-like ubiquitin ligase from *Escherichia coli* O157:H7. *J Biol Chem* 286, 441-449.

Lin TS, K.P., 1980. Isolation and characterization of a cuticular polyester (cutin) hydrolyzing enzyme from phytopathogenic fungi. *Physiol Plant Pathol* 17, 1-15.

Long, J.Z., Li, W., Booker, L., Burston, J.J., Kinsey, S.G., Schlosburg, J.E., Pavon, F.J., Serrano, A.M., Selley, D.E., Parsons, L.H., Lichtman, A.H., Cravatt, B.F., 2009a. Selective blockade of 2-arachidonoylglycerol hydrolysis produces cannabinoid behavioral effects. *Nat Chem Biol* 5, 37-44.

Long, J.Z., Nomura, D.K., Cravatt, B.F., 2009b. Characterization of monoacylglycerol lipase inhibition reveals differences in central and peripheral endocannabinoid metabolism. *Chem Biol* 16, 744-753.

Lopez-Lopez, S., Nolasco, H., Vega-Villasante, F., 2003. Characterization of digestive gland esterase-lipase activity of juvenile redclaw crayfish *Cherax quadricarinatus*. *Comp Biochem Physiol B Biochem Mol Biol* 135, 337-347.

Luca, S., Mihaescu, T., 2013. History of BCG Vaccine. *Maedica (Buchar)* 8, 53-58.

Lun, S., Bishai, W.R., 2007. Characterization of a novel cell wall-anchored protein with carboxylesterase activity required for virulence in *Mycobacterium tuberculosis*. *J Biol Chem* 282, 18348-18356.

Luthy, R., Bowie, J.U., Eisenberg, D., 1992. Assessment of protein models with three-dimensional profiles. *Nature* 356, 83-85.



Mackenzie, N., Alexander, D.C., Turenne, C.Y., Behr, M.A., De Buck, J.M., 2009. Genomic comparison of PE and PPE genes in the *Mycobacterium avium* complex. *Journal of clinical microbiology* 47, 1002-1011.

Madania, A., Habous, M., Zarzour, H., Ghoury, I., Hebbo, B., 2012. Characterization of mutations causing rifampicin and isoniazid resistance of *Mycobacterium tuberculosis* in Syria. *Pol J Microbiol* 61, 23-32.

Mahajan, R., 2013. Bedaquiline: First FDA-approved tuberculosis drug in 40 years. *International journal of applied & basic medical research* 3, 1-2.

Mani, V., Wang, S., Inci, F., De Libero, G., Singhal, A., Demirci, U., 2014. Emerging technologies for monitoring drug-resistant tuberculosis at the point-of-care. *Adv Drug Deliv Rev* 78, 105-117.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., Sali, A., 2000. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 29, 291-325.

McEvoy, C.R., Cloete, R., Muller, B., Schurch, A.C., van Helden, P.D., Gagneux, S., Warren, R.M., Gey van Pittius, N.C., 2012. Comparative analysis of *Mycobacterium tuberculosis* pe and ppe genes reveals high sequence variation and an apparent absence of selective constraints. *PLoS One* 7, e30593.

McKinney, J.D., Honer zu Bentrup, K., Munoz-Elias, E.J., Miczak, A., Chen, B., Chan, W.T., Swenson, D., Sacchettini, J.C., Jacobs, W.R., Jr., Russell, D.G., 2000. Persistence of *Mycobacterium tuberculosis* in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase. *Nature* 406, 735-738.

Meena, L.S., Rajni, 2010. Survival mechanisms of pathogenic *Mycobacterium tuberculosis* H37Rv. *Febs J* 277, 2416-2427.

Minnikin, D.E., Kremer, L., Dover, L.G., Besra, G.S., 2002. The methyl-branched fortifications of *Mycobacterium tuberculosis*. *Chem Biol* 9, 545-553.

Mishra, K.C., de Chastellier, C., Narayana, Y., Bifani, P., Brown, A.K., Besra, G.S., Katoch, V.M., Joshi, B., Balaji, K.N., Kremer, L., 2008. Functional role of the PE domain and immunogenicity of the *Mycobacterium tuberculosis* triacylglycerol hydrolase LipY. *Infect Immun* 76, 127-140.

Mitraki, A., Miller, S., van Raaij, M.J., 2002. Review: conformation and folding of novel beta-structural elements in viral fiber proteins: the triple beta-spiral and triple beta-helix. *Journal of structural biology* 137, 236-247.

Mohareer, K., Tundup, S., Hasnain, S.E., 2011. Transcriptional regulation of *Mycobacterium tuberculosis* PE/PPE genes: a molecular switch to virulence? *J Mol Microbiol Biotechnol* 21, 97-109.

Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., Olson, A.J., 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem* 30, 2785-2791.

Mukhopadhyay, S., Balaji, K.N., 2011. The PE and PPE proteins of *Mycobacterium tuberculosis*. *Tuberculosis (Edinb)* 91, 441-447.

Naim, M., Bhat, S., Rankin, K.N., Dennis, S., Chowdhury, S.F., Siddiqi, I., Drabik, P., Sulea, T., Bayly, C.I., Jakalian, A., Purisima, E.O., 2007. Solvated interaction energy (SIE) for scoring protein-ligand binding affinities. 1. Exploring the parameter space. *J Chem Inf Model* 47, 122-133.

Nair, S., 2014. Immunomodulatory Role of *Mycobacterial* PE/PPE Family of Proteins. *Proc Indian Natn Sci Acad* 80, 1055-1072.

Neyrolles, O., Hernandez-Pando, R., Pietri-Rouxel, F., Fornes, P., Tailleux, L., Barrios Payan, J.A., Pivert, E., Bordat, Y., Aguilar, D., Prevost, M.C., Petit, C., Gicquel, B., 2006. Is adipose tissue a place for *Mycobacterium tuberculosis* persistence? *PLoS One* 1, e43.

Okkels, L.M., Brock, I., Follmann, F., Agger, E.M., Arend, S.M., Ottenhoff, T.H., Oftung, F., Rosenkrands, I., Andersen, P., 2003. PPE protein (Rv3873) from DNA segment RD1 of *Mycobacterium tuberculosis*: strong recognition of both specific T-cell epitopes and epitopes conserved within the PPE family. *Infect Immun* 71, 6116-6123.

Ormerod, L.P., Horsfield, N., 1996. Frequency and type of reactions to antituberculosis drugs: observations in routine treatment. *Tuber Lung Dis* 77, 37-42.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N., 1999. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America* 96, 2896-2901.

Park, I.H., Kim, S.H., Lee, Y.S., Lee, S.C., Zhou, Y., Kim, C.M., Ahn, S.C., Choi, Y.L., 2009. Gene cloning, purification, and characterization of a cold-adapted lipase produced by *Acinetobacter baumannii* BD5. *J Microbiol Biotechnol* 19, 128-135.

Parker, S.K., Barkley, R.M., Rino, J.G., Vasil, M.L., 2009. *Mycobacterium tuberculosis* Rv3802c encodes a phospholipase/thioesterase and is inhibited by the antimycobacterial agent tetrahydrolipstatin. *PLoS One* 4, e4281.

Parrinello M, R.A., 1981. Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys* 52, 7182–7190.

Pearson, W.R., 2013. An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics / editorial board, Andreas D. Baxevanis ... [et al.] Chapter 3, Unit 3.1.*

Pencreach, G., Baratti, J.C., 1996. Hydrolysis of p-nitrophenyl palmitate in n-heptane by the *Pseudomonas cepacia* lipase: A simple test for the determination of lipase activity in organic media. *Enzyme Microb Tech* 18, 417-422.

Pereira, S.M., Dantas, O.M., Ximenes, R., Barreto, M.L., 2007. [BCG vaccine against tuberculosis: its protective effect and vaccination policies]. *Rev Saude Publica* 41 Suppl 1, 59-66.

Phulera, S., Mande, S.C., 2013. The crystal structure of *Mycobacterium tuberculosis* NrdH at 0.87 Å suggests a possible mode of its activity. *Biochemistry* 52, 4056-4065.

Pohle, S., Appelt, C., Roux, M., Fiedler, H.P., Sussmuth, R.D., 2011. Biosynthetic gene cluster of the non-ribosomally synthesized cyclodepsipeptide skyllamycin: deciphering unprecedented ways of unusual hydroxylation reactions. *J Am Chem Soc* 133, 6194-6205.

Poulet, S., Cole, S.T., 1995. Characterization of the highly abundant polymorphic GC-rich-repetitive sequence (PGRS) present in *Mycobacterium tuberculosis*. *Arch Microbiol* 163, 87-95.

Poulsen, K., Haber, E., Burton, J., 1976. On the specificity of human renin. Studies with peptide inhibitors. *Biochimica et biophysica acta* 452, 533-537.

Pratt, J., Cooley, J.D., Purdy, C.W., Straus, D.C., 2000. Lipase activity from strains of *Pasteurella multocida*. *Curr Microbiol* 40, 306-309.

Purisma, E.O., 1998. Fast summation boundary element method for calculating solvation free energies of macromolecules. *J Comput Chem* 19, 1494-1504.

Purisma, E.O., Nilar, S.H., 1995. A Simple yet Accurate Boundary-Element Method for Continuum Dielectric Calculations. *J Comput Chem* 16, 681-689.

Ramachandran, G.N., Ramakrishnan, C., Sasisekharan, V., 1963. Stereochemistry of polypeptide chain configurations. *J Mol Biol* 7, 95-99.

Ramakrishnan, G., Ochoa-Montano, B., Raghavender, U.S., Mudgal, R., Joshi, A.G., Chandra, N.R., Sowdhamini, R., Blundell, T.L., Srinivasan, N., 2015. Enriching the annotation of *Mycobacterium tuberculosis* H37Rv proteome using remote homology detection approaches: insights into structure and function. *Tuberculosis (Edinb)* 95, 14-25.

Ramakrishnan, L., Federspiel, N.A., Falkow, S., 2000. Granuloma-specific expression of *Mycobacterium* virulence proteins from the glycine-rich PE-PGRS family. *Science* 288, 1436-1439.

Ramulu, H.G., Adindla, S., Guruprasad, L., 2006. Analysis and modeling of mycolyl-transferases in the CMN group. *Bioinformation* 1, 161-169.

Ratledge, C., 1982. Nutrition, growth and metabolism. In: Ratledge S, Stanford J L, editors. *Biology of the mycobacteria*. London, United Kingdom: Academic Press Inc., Ltd 1, 186–212.

Rezwani, M., Laneelle, M.A., Sander, P., Daffe, M., 2007. Breaking down the wall: fractionation of mycobacteria. *J Microbiol Methods* 68, 32-39.

Saikrishnan, K., Kalapala, S.K., Varshney, U., Vijayan, M., 2005. X-ray structural studies of *Mycobacterium tuberculosis* RRF and a comparative study of RRFs of known structure. Molecular plasticity and biological implications. *J Mol Biol* 345, 29-38.

Sali, A., Blundell, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234, 779-815.

Samanta, U., Bahnson, B.J., 2008. Crystal structure of human plasma platelet-activating factor acetylhydrolase: structural implication to lipoprotein binding and catalysis. *J Biol Chem* 283, 31617-31624.

Sampson, S.L., 2011. Mycobacterial PE/PPE proteins at the host-pathogen interface. *Clinical & developmental immunology* 2011, 497203.

Sampson, S.L., Lukey, P., Warren, R.M., van Helden, P.D., Richardson, M., Everett, M.J., 2001. Expression, characterization and subcellular localization of the *Mycobacterium tuberculosis* PPE gene Rv1917c. *Tuberculosis (Edinb)* 81, 305-317.

Saravanan, P., Avinash, H., Dubey, V.K., Patra, S., 2012. Targeting essential cell wall lipase Rv3802c for potential therapeutics against tuberculosis. *J Mol Graph Model* 38, 235-242.

Sasseti, C.M., Boyd, D.H., Rubin, E.J., 2003. Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol* 48, 77-84.

Sasseti, C.M., Rubin, E.J., 2003. Genetic requirements for mycobacterial survival during infection. *Proceedings of the National Academy of Sciences of the United States of America* 100, 12989-12994.

Sayes, F., Sun, L., Di Luca, M., Simeone, R., Degaiffier, N., Fiette, L., Esin, S., Brosch, R., Bottai, D., Leclerc, C., Majlessi, L., 2012. Strong immunogenicity and cross-reactivity of *Mycobacterium tuberculosis* ESX-5 type VII secretion: encoded PE-PPE proteins predicts vaccine potential. *Cell Host Microbe* 11, 352-363.

Schneider, G., Bohm, H.J., 2002. Virtual screening and fast automated docking methods. *Drug Discov Today* 7, 64-70.

Schue, M., Maurin, D., Dhouib, R., Bakala N'Goma, J.C., Delorme, V., Lambeau, G., Carriere, F., Canaan, S., 2010. Two cutinase-like proteins secreted by *Mycobacterium tuberculosis* show very different lipolytic activities reflecting their physiological function. *Faseb J* 24, 1893-1903.

Sebastian, J., Chandra, A.K., Kolattukudy, P.E., 1987. Discovery of a cutinase-producing *Pseudomonas* sp. cohabiting with an apparently nitrogen-fixing *Corynebacterium* sp. in the phyllosphere. *J Bacteriol* 169, 131-136.

Sebastian, J., Kolattukudy, P.E., 1988. Purification and characterization of cutinase from a fluorescent *Pseudomonas putida* bacterial strain isolated from phyllosphere. *Arch Biochem Biophys* 263, 77-85.

Shang, P., Xia, Y., Liu, F., Wang, X., Yuan, Y., Hu, D., Tu, D., Chen, Y., Deng, P., Cheng, S., Zhou, L., Ma, Y., Zhu, L., Gao, W., Wang, H., Chen, D., Yang, L., He, P., Wu, S., Tang, S., Lv, X., Shu, Z., Zhang, Y., Yang, Z., Li, N., Sun, F., Li, X., He, Y., Garner, P., Zhan, S., 2011. Incidence, clinical features and impact on anti-tuberculosis treatment of anti-tuberculosis drug induced liver injury (ATLI) in China. *PLoS One* 6, e21836.

Shi, J., Blundell, T.L., Mizuguchi, K., 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol* 310, 243-257.

Shrivastava, T., Ramachandran, R., 2007. Mechanistic insights from the crystal structures of a feast/famine regulatory protein from *Mycobacterium tuberculosis* H37Rv. *Nucleic Acids Res* 35, 7324-7335.

Singh, G., Arya, S., Narang, D., Jadeja, D., Gupta, U.D., Singh, K., Kaur, J., 2014a. Characterization of an acid inducible lipase Rv3203 from *Mycobacterium tuberculosis* H37Rv. *Mol Biol Rep* 41, 285-296.

Singh, G., Jadeja, D., Kaur, J., 2010. Lipid hydrolizing enzymes in virulence: *Mycobacterium tuberculosis* as a model system. *Crit Rev Microbiol* 36, 259-269.

Singh, P., Rao, R.N., Reddy, J.R., Prasad, R., Kotturu, S.K., Ghosh, S., Mukhopadhyay, S., 2016. PE11, a PE/PPE family protein of *Mycobacterium tuberculosis* is involved in cell wall remodeling and virulence. *Sci Rep* 6, 21624.

Singh, S., Bajpai, U., Lynn, A.M., 2014b. Structure based virtual screening to identify inhibitors against MurE Enzyme of *Mycobacterium tuberculosis* using AutoDock Vina. *Bioinformation* 10, 697-702.

Sinnott, M.L., 1990. Catalytic Mechanisms of Enzymatic Glycosyl Transfer. *Chemical Reviews* 90, 1171-1202.

Smith, I., 2003. *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. *Clin Microbiol Rev* 16, 463-496.

Smith, N.H., Hewinson, R.G., Kremer, K., Brosch, R., Gordon, S.V., 2009. Myths and misconceptions: the origin and evolution of *Mycobacterium tuberculosis*. *Nat Rev Microbiol* 7, 537-544.

Snel, B., Bork, P., Huynen, M., 2000. Genome evolution. Gene fusion versus gene fission. *Trends Genet* 16, 9-11.

Snider, D.E., Jr., La Montagne, J.R., 1994. The neglected global tuberculosis problem: a report of the 1992 World Congress on Tuberculosis. *The Journal of infectious diseases* 169, 1189-1196.

Snow, C.D., Sorin, E.J., Rhee, Y.M., Pande, V.S., 2005. How well can simulation predict protein folding kinetics and thermodynamics? *Annu Rev Biophys Biomol Struct* 34, 43-69.

Song, W.C., Funk, C.D., Brash, A.R., 1993. Molecular cloning of an allene oxide synthase: a cytochrome P450 specialized for the metabolism of fatty acid hydroperoxides. *Proceedings of the National Academy of Sciences of the United States of America* 90, 8519-8523.

Sousa da Silva, A.W., Vranken, W.F., 2012. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res Notes* 5, 367.

Speck-Planche, A., Scotti, M.T., de Paulo-Emerenciano, V., 2010. Current pharmaceutical design of antituberculosis drugs: future perspectives. *Curr Pharm Des* 16, 2656-2665.

Srivastava, S.K., Tripathi, R.P., Ramachandran, R., 2005. NAD<sup>+</sup>-dependent DNA Ligase (Rv3014c) from *Mycobacterium tuberculosis*. Crystal structure of the adenylation domain and identification of novel inhibitors. *J Biol Chem* 280, 30273-30281.

Srivastava, V., Rouanet, C., Srivastava, R., Ramalingam, B., Loch, C., Srivastava, B.S., 2007. Macrophage-specific *Mycobacterium tuberculosis* genes: identification by green fluorescent protein and kanamycin resistance selection. *Microbiology* 153, 659-666.

Stead, W.W., 1997. The origin and erratic global spread of tuberculosis. How the past explains the present and is the key to the future. *Clin Chest Med* 18, 65-77.

Sterne, J.A., Rodrigues, L.C., Guedes, I.N., 1998. Does the efficacy of BCG decline with time since vaccination? *Int J Tuberc Lung Dis* 2, 200-207.

Stinear, T.P., Seemann, T., Harrison, P.F., Jenkin, G.A., Davies, J.K., Johnson, P.D., Abdallah, Z., Arrowsmith, C., Chillingworth, T., Churcher, C., Clarke, K., Cronin, A., Davis, P., Goodhead, I., Holroyd, N., Jagels, K., Lord, A., Moule, S., Mungall, K., Norbertczak, H., Quail, M.A., Rabinowitsch, E., Walker, D., White, B., Whitehead, S., Small, P.L., Brosch, R., Ramakrishnan, L., Fischbach, M.A., Parkhill, J., Cole, S.T., 2008. Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of *Mycobacterium tuberculosis*. *Genome Res* 18, 729-741.

Strong, M., Sawaya, M.R., Wang, S., Phillips, M., Cascio, D., Eisenberg, D., 2006. Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America* 103, 8060-8065.

Sulea, T., Cui, Q.Z., Purisima, E.O., 2011. Solvated Interaction Energy (SIE) for Scoring Protein-Ligand Binding Affinities. 2. Benchmark in the CSAR-2010 Scoring Exercise. *J Chem Inf Model* 51, 2066-2081.

Sultana, R., Tanneeru, K., Guruprasad, L., 2011. The PE-PPE domain in mycobacterium reveals a serine alpha/beta hydrolase fold and function: an in-silico analysis. *PLoS One* 6, e16745.

Sutcliffe, M.J., Haneef, I., Carney, D., Blundell, T.L., 1987a. Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng* 1, 377-384.

Sutcliffe, M.J., Hayes, F.R., Blundell, T.L., 1987b. Knowledge based modelling of homologous proteins, Part II: Rules for the conformations of substituted sidechains. *Protein Eng* 1, 385-392.

Swathi Adindla, R.S., Karunakar Tanneeru, Swati Singh and Lalitha Guruprasad, 2013. To make the most of a protein sequence. *Proceedings of Andhra Pradesh of Akademik of Sciences* 15.

Talarico, S., Cave, M.D., Foxman, B., Marrs, C.F., Zhang, L., Bates, J.H., Yang, Z., 2007. Association of *Mycobacterium tuberculosis* PE\_PGRS33 polymorphism with clinical and epidemiological characteristics. *Tuberculosis (Edinb)* 87, 338-346.

Talarico, S., Zhang, L., Marrs, C.F., Foxman, B., Cave, M.D., Brennan, M.J., Yang, Z., 2008. *Mycobacterium tuberculosis* PE\_PGRS16 and PE\_PGRS26 genetic polymorphism among clinical isolates. *Tuberculosis (Edinb)* 88, 283-294.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution* 28, 2731-2739.

Tanneeru, K., Balla, A.R., Guruprasad, L., 2015. In silico 3D structure modeling and inhibitor binding studies of human male germ cell-associated kinase. *J Biomol Struct Dyn* 33, 1710-1719.

Tanneeru, K., Guruprasad, L., 2013. Ponatinib is a pan-BCR-ABL kinase inhibitor: MD simulations and SIE study. *PLoS One* 8, e78556.

Taylor, N.R., Cleasby, A., Singh, O., Skarzynski, T., Wonacott, A.J., Smith, P.W., Sollis, S.L., Howes, P.D., Cherry, P.C., Bethell, R., Colman, P., Varghese, J., 1998. Dihydropyranocarboxamides related to zanamivir: a new series of inhibitors of influenza virus sialidases. 2. Crystallographic and molecular modeling study of complexes of 4-amino-4H-pyran-6-carboxamides and sialidase from influenza virus types A and B. *J Med Chem* 41, 798-807.

Thillai, M., Pollock, K., Pareek, M., Lalvani, A., 2014. Interferon-gamma release assays for tuberculosis: current and future applications. *Expert Rev Respir Med* 8, 67-78.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL\_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25, 4876-4882.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673-4680.

Tormo, J., Lamed, R., Chirino, A.J., Morag, E., Bayer, E.A., Shoham, Y., Steitz, T.A., 1996. Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose. *The EMBO journal* 15, 5739-5751.

Torronen, A., Harkki, A., Rouvinen, J., 1994. Three-dimensional structure of endo-1,4-beta-xylanase II from *Trichoderma reesei*: two conformational states in the active site. *The EMBO journal* 13, 2493-2501.

Trott, O., Olson, A.J., 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 31, 455-461.

Tundup, S., Akhter, Y., Thiagarajan, D., Hasnain, S.E., 2006. Clusters of PE and PPE genes of *Mycobacterium tuberculosis* are organized in operons: evidence that PE\_Rv2431c is co-transcribed with PPE\_Rv2430c and their gene products interact with each other. *FEBS Lett* 580, 1285-1293.

Tundup, S., Pathak, N., Ramanadham, M., Mukhopadhyay, S., Murthy, K.J., Ehtesham, N.Z., Hasnain, S.E., 2008. The co-operonic PE25/PPE41 protein complex of *Mycobacterium tuberculosis* elicits increased humoral and cell mediated immune response. *PLoS One* 3, e3586.

Udwadia, Z.F., Amale, R.A., Ajbani, K.K., Rodrigues, C., 2012. Totally drug-resistant tuberculosis in India. *Clin Infect Dis* 54, 579-581.

Uhlmann, S., Sussmuth, R.D., Cryle, M.J., 2013. Cytochrome p450sky interacts directly with the nonribosomal peptide synthetase to generate three amino acid precursors in skylamycin biosynthesis. *ACS chemical biology* 8, 2586-2596.

Ulrich Essmann, L.P., Max L. Berkowitz, Tom Darden, Hsing Lee and Lee G. Pedersen, 1995. A smooth particle meshes Ewald method *J Chem Phys* 103, 8577–8593.

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., Berendsen, H.J., 2005. GROMACS: fast, flexible, and free. *J Comput Chem* 26, 1701-1718.

Vissa, V.D., Brennan, P.J., 2001. The genome of *Mycobacterium leprae*: a minimal mycobacterial gene set. *Genome Biol* 2, REVIEWS1023.

Voskuil, M.I., Schnappinger, D., Rutherford, R., Liu, Y., Schoolnik, G.K., 2004. Regulation of the *Mycobacterium tuberculosis* PE/PPE genes. *Tuberculosis (Edinb)* 84, 256-262.

Waight, A.B., Pedersen, B.P., Schlessinger, A., Bonomi, M., Chau, B.H., Roe-Zurz, Z., Risenmay, A.J., Sali, A., Stroud, R.M., 2013. Structural basis for alternating access of a eukaryotic calcium/proton exchanger. *Nature* 499, 107-110.

Wallace, R., 2010. History of MAC. Available: <http://www.maclungdisease.org/history-of-mac>. Accessed.

Walton, T.J., Kolattukudy, P.E., 1972. Determination of the structures of cutin monomers by a novel depolymerization procedure and combined gas chromatography and mass spectrometry. *Biochemistry* 11, 1885-1896.

Wang, J., Wang, W., Kollman, P.A., Case, D.A., 2006. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* 25, 247-260.

Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A., Case, D.A., 2004. Development and testing of a general amber force field. *J Comput Chem* 25, 1157-1174.

West, N.P., Chow, F.M., Randall, E.J., Wu, J., Chen, J., Ribeiro, J.M., Britton, W.J., 2009. Cutinase-like proteins of *Mycobacterium tuberculosis*: characterization of their variable enzymatic functions and active site identification. *Faseb J* 23, 1694-1704.

West, N.P., Wozniak, T.M., Valenzuela, J., Feng, C.G., Sher, A., Ribeiro, J.M., Britton, W.J., 2008. Immunological diversity within a family of cutinase-like proteins of *Mycobacterium tuberculosis*. *Vaccine* 26, 3853-3859.

WHO, 2006. Guidelines for the programmatic management of drug-resistant tuberculosis. .

WHO, 2014. Global tuberculosis report. World Health Organization.

Wu, H.M., Liu, S.W., Hsu, M.T., Hung, C.L., Lai, C.C., Cheng, W.C., Wang, H.J., Li, Y.K., Wang, W.C., 2009. Structure, mechanistic action, and essential residues of a GH-64 enzyme, laminaripentaose-producing beta-1,3-glucanase. *J Biol Chem* 284, 26708-26715.

Yee, D., Valiquette, C., Pelletier, M., Parisien, I., Rocher, I., Menzies, D., 2003. Incidence of serious side effects from first-line antituberculosis drugs among patients treated for active tuberculosis. *Am J Respir Crit Care Med* 167, 1472-1477.

Zaman, K., 2010. Tuberculosis: a global health problem. *Journal of health, population, and nutrition* 28, 111-113.

Zheng, Q., Jiang, D., Zhang, W., Zhang, Q., Zhao, Q., Jin, J., Li, X., Yang, H., Bartlam, M., Shaw, N., Zhou, W., Rao, Z., 2014. Mechanism of dephosphorylation of glucosyl-3-phosphoglycerate by a histidine phosphatase. *J Biol Chem* 289, 21242-21251.

Zumbo, A., Palucci, I., Cascioferro, A., Sali, M., Ventura, M., D'Alfonso, P., Iantomasi, R., Di Sante, G., Ria, F., Sanguinetti, M., Fadda, G., Manganelli, R., Delogu, G., 2013. Functional dissection of protein domains involved in the immunomodulatory properties of PE\_PGRS33 of *Mycobacterium tuberculosis*. *Pathogens and disease* 69, 232-239.

# STUDIES ON THE PE AND PPE PROTEINS OF MYCOBACTERIA

---

## ORIGINALITY REPORT

---

18%	11%	15%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

---

## PRIMARY SOURCES

---

1	Rafiya Sultana. "The PE-PPE Domain in Mycobacterium Reveals a Serine $\alpha/\beta$ Hydrolase Fold and Function: An In-Silico Analysis", PLoS ONE, 02/10/2011 Publication	1%
2	<a href="http://www.nature.com">www.nature.com</a> Internet Source	1%
3	<a href="http://onlinelibrary.wiley.com">onlinelibrary.wiley.com</a> Internet Source	1%
4	Submitted to Georgetown University Student Paper	1%
5	Mukhopadhyay, S.. "The PE and PPE proteins of Mycobacterium tuberculosis", Tuberculosis, 201109 Publication	1%
6	Submitted to University of Hyderabad, Hyderabad Student Paper	<1%

---