COMPUTATIONAL AND EXPERIMENTAL STUDIES OF PUTATIVE VIRULENCE FACTORS OF *MYCOBACTERIUM TUBERCULOSIS* H37Rv

By

Mohd. Shahbaaz

(Reg. No: 21451904)

Submitted in fulfilment of the requirements of the degree of Doctor of Philosophy in Chemistry in the Faculty of Applied Sciences at the Durban University of Technology

DECLARATION

I **Mohd. Shahbaaz** declare that the thesis submitted for the degree of Doctor of Philosophy (Ph.D.) in Chemistry at the Durban University of Technology is a result of my own investigation and has not already been accepted in substance for any degree, and that its only prior publication was in the form of conference papers and journal articles.

Student Name: Mohd. Shahbaaz

Student Signature: [redacted] **Date**: ……24/7/2017…

Supervisor Name: Prof. K. Bisetty

Signature [redacted] **Date**: ……24/7/2017…

Co-Supervisor Name: Dr. Md. Imtaiyaz Hassan, Jamia Millia Islamia (India)

[redacted]
Assistant Professor
Centre for Interdisciplinary Research in Basic Sciences
Jamia Millia Islamia, New Delhi-110025

Signature: …….. **Date:** ……24/7/2017…

ACKNOWLEDGEMENTS

I thank Almighty Allah, The most beneficent, the most merciful, for giving me the patience and perseverance throughout my life to move in the right direction.

I owe profound gratitude and sincere regards to my supervisors Prof K Bisetty, Head, Department of Chemistry, for his benevolent support and direction throughout my research. His careful corrections during regular meetings and discussions always ensured a more thoughtful approach towards problem-solving.

It is my privilege to express my gratitude and indebtedness to my supervisor Dr Md. Imtaiyaz Hassan in the Center for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, India, for his guidance, continuous encouragement, constructive suggestions and sagacious advice throughout the course of my research work. His keen sense of scientific temperament was a source of inspiration to me.

I would like to express my gratitude to the Centre for high performance computing, Cape Town for providing server and necessary software which are important for my research.

I would like to thank all my Computational Modeling and Bio-Analytical Chemistry (CMBAC) group members Bayu, Myalo, Kanchi, Nosipho and Athika for unforgettable co-operation.

I would like to thank my Lab mates of Danish, Shahzaib, and Amir in the Center for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, India with whom I share numerous memorable moments which made my research really enjoyable.

I also take this opportunity to thank the Durban University of Technology for proving the financial assistance and to all the staff members in the Department of Chemistry who have helped me in various ways during my research work.

I dedicate this thesis to my mother Shahana Ahmed and Father Naeem Ahmed as well as to my beloved younger siblings Sana, Saba, and Sufian as they provide inseparable support, love and care which help me to complete my doctoral research. I thank Allah for them, they are a great blessing.

Mohd. Shahbaaz

ABSTRACT

In drug discovery and development of anti-tubercular therapeutics, it is necessary to study the physiology and genetics of the molecular mechanisms present in the *Mycobacterium tuberculosis*. The virulence of *M. tuberculosis* is attributed to its unique genome, which contains a high frequency of glycine-rich proteins and genes involved in the metabolism of the fatty acids. Consequently, the presence of a diversity of the pathogenic pathways such as acid tolerance and drug resistance mechanisms in *M. tuberculosis* makes the treatment of Tuberculosis (TB) challenging. However, the molecular basis of the virulence factors involved in the pathogenesis is not fully understood. Accordingly, the current study focuses on better understanding of the pathogenic proteins present in this bacterium using available computational techniques.

In South Africa, there is an alarming increase in the drug-resistant TB in HIV co-infected patients, which is one of the biggest challenges to the current anti-tubercular therapies. An extensive literature search showed that the mutations in the virulent proteins of *M. tuberculosis* resulted in the development of drug tolerance in the pathogen. The molecular and genetic studies identified frequently occurring point mutations associated with the drug resistance in proteins of *M. tuberculosis*. Despite the efforts, TB infection is still increasing because different pathogenic pathways in the bacterial system are still undiscovered. Therefore, this study involves an *in silico* approach aimed at the identification of novel drug resistance implicated point mutations. The site-directed mutations leading to the development of resistance against four first-line drugs (Ethambutol, Isoniazid, Rifampicin, and Streptomycin) were studied extensively. In the primary investigation, pathogenic mutational landscapes were classified in the sequences of the studied proteins. The effects of these mutations on the stability of the proteins were studied using diverse computational techniques. The structural basis of the point mutations with the highest

destabilizing effects was analyzed using the principles of the Density Functional Theory (DFT), molecular docking and molecular dynamics (MD) simulation studies. The varied conformational behavior resulted from these predicted substitutions were compared with the experimentally derived mutations reported in the literature. The outcome of this study enabled the identification of the novel drug resistance-associated point mutations which were not previously reported.

Furthermore, a detailed understanding of the conformational behavior of diverse virulent proteins present in *M. tuberculosis* was also generated in this study. Literature study showed that inside the host's macrophage cells, the virulent proteins such as isocitrate lyase, lipase lipF, magnesium transporter MgtC, porin protein OmpATb, a protein of two component systems PhoP, Rv2136c and Rv3671c have an established role in the development of the acid tolerance. On the other hand, information regarding their role in the acid resistance is scarce. Accordingly, the structural basis of their role in acid resistance was analyzed using constant pH based MD simulations. In the studied proteins, the lipF and PhoP showed highest structural stability in highly acidic conditions throughout the course of MD simulations. Therefore, these proteins may play a primary role in the process of resistance.

In addition to these pathogenic proteins, there is a need to identify new undiscovered virulent proteins in the genome of *M. tuberculosis*, which increases the efficiency of the current therapy. The knowledge generated by the analyses of the proteins involved in resistance and pathogenic mechanisms of *M. tuberculosis* forms the basis for the identification of new virulence factors. Therefore, an *in silico* protocol was used for the functional annotations and analyses of the virulence characteristics.

M. tuberculosis contains 1000 Hypothetical Proteins (HPs), which are functionally uncharacterized proteins and their existence was not validated at the biochemical level. In this

study, the sequences of the HPs were extensively analyzed and the functions of 662 HPs were successfully predicted. Furthermore, 483 HPs were classified in the category of the enzymes, 141 HPs were predicted to be involved in the diverse cellular mechanisms and 38 HPs may function as transporters and carriers proteins. The 307 HPs among this group of proteins were less precisely predicted because of the unavailability of the reliable functional homologs. An assessment of the virulence characteristics associated with the 1000 HPs enabled the classification of 28 virulent HPs. The structure of six HPs with highest predicted virulence score was analyzed using molecular modelling techniques.

Amongst the predicted virulent HPs, the clone for Rv3906c purchased from the DNASU repository because of the ease of its availability. The gene of Rv3906c was isolated and cloned into a pET-21c expression vector. The analyses of the nucleotide sequence showed that Rv3906c gene (500 bp) encodes a 169 amino acid protein of molecular weight 17.80 kDa (~18.0 kDa). The sequence analyses of Rv3906c showed that the HPs showed high similarities with pullulanase, a thermophilic enzyme. The stability profile at different temperatures for Rv3906c generated using MD simulations showed that Rv3906c maintained its structural identity at higher temperatures. It is expected that this study will result in the design of better therapeutic against the infection of *M. tuberculosis*, as novel undiscovered virulence factors were classified and analyzed in addition to the conformational profiles of the virulent proteins involved in the resistance mechanisms.

LIST OF PUBLICATIONS

1. **Mohd. Shahbaaz**, Krishna Bisetty, Faizan Ahmad, Md. Imtaiyaz Hassan: Towards New Drug Targets? Function Prediction of Putative Proteins of *Neisseria meningitidis* MC58 and Their Virulence Characterization. OMICS A Journal of Integrative Biology 05/2015; 19(7). DOI:10.1089/omi.2015.0032
2. **Mohd. Shahbaaz**, Krishna Bisetty, Faizan Ahmad, Md. Imtaiyaz Hassan: In silico approaches for the identification of virulence candidates amongst hypothetical proteins of *Mycoplasma pneumoniae* 309. Computational Biology and Chemistry 09/2015; 59(Part A):67–80. DOI:10.1016/j.compbiolchem.2015.09.007
3. **Mohd. Shahbaaz**, Krishna Bisetty, Faizan Ahmad, Md. Imtaiyaz Hassan: Current advances in the identification and characterization of putative drug and vaccine targets in the bacterial genomes. Current topics in medicinal chemistry 08/2015; 16(9).
4. **Mohd. Shahbaaz**, Krishna Bisetty, Md. Imtaiyaz Hassan: *In silico* identification and experimental validations of putative virulence factors of *Mycobacterium tuberculosis* H37Rv. Scientific Reports (2016) (Communicated)
5. **Mohd. Shahbaaz**, Krishna Bisetty, Md. Imtaiyaz Hassan: Constant pH Molecular Dynamics Simulations based understanding of the acid resistance mechanisms of *Mycobacterium tuberculosis* H37Rv Molecular Biosystems (2016) (In preparation)
6. **Mohd. Shahbaaz**, Krishna Bisetty, Md. Imtaiyaz Hassan: Classification and analyses of the drug resistance mutational landscapes in *Mycobacterium tuberculosis* H37Rv RSC advances (2016) (In preparation)
7. Huma Naz, **Mohd. Shahbaaz**, Md. Anzarul Haque, Krishna Bisetty, Asimul Islam, Faizan Ahmad, Md. Imtaiyaz Hassan: Urea-induced denaturation of human calcium-calmodulin dependent protein kinase IV: A combined spectroscopic and MD simulation study. Journal of Biomolecular Structure & Dynamics 02/2016; DOI:10.1080/07391102.2016.1150203

8. Danish Idrees, **Mohd. Shahbaaz**, Krishna Bisetty, Asimul Islam, Faizan Ahmad, Md. Imtaiyaz Hassan: Effect of pH on structure, function and stability of mitochondrial carbonic anhydrase VA. Journal of Biomolecular Structure & Dynamics 01/2016
9. Huma Naz, **Mohd. Shahbaaz**, Krishna Bisetty, Asimul Islam, Faizan Ahmad, Md. Imtaiyaz Hassan: Effect of pH on the structure, function and stability of human calcium/calmodulin-dependent protein kinase IV: A combined spectroscopic and MD simulation studies. Biochemistry and Cell Biology 01/2016; DOI:10.1139/bcb-2015-0132
10. Shama Khan, **Mohd.Shahbaaz**, Krishna Bisetty, Asimul Islam, Faizan Ahmad, Md. Imtaiyaz Hassan: Classification and Functional Analyses of Putative Conserved Proteins from *Chlamydophila pneumoniae* CWL029. Interdisciplinary Sciences Computational Life Sciences 11/2015; DOI:10.1007/s12539-015-0134-7
11. Farha Naz, **Mohd. Shahbaaz**, Shama Khan, Krishna Bisetty, Asimul Islam, Faizan Ahmad, Md. Imtaiyaz Hassan: PKR-inhibitor binds efficiently with human microtubule affinity-regulating kinase 4. Journal of Molecular Graphics and Modelling 10/2015; 62. DOI:10.1016/j.jmgm.2015.10.009
12. Faez Iqbal Khan, **Mohd. Shahbaaz**, Krishna Bisetty, Abdul Waheed, William S. Sly, Faizan Ahmad, Md. Imtaiyaz Hassan: Large scale analysis of the mutational landscape in β-glucuronidase: A major player of mucopolysaccharidosis type VII. Gene 09/2015; 576. DOI:10.1016/j.gene.2015.09.062
13. Farha Naz, **Mohd. Shahbaaz**, Krishna Bisetty, Asimul Islam, Faizan Ahmad, Md. Imtaiyaz Hassan: Designing New Kinase Inhibitor Derivatives as Therapeutics Against Common Complex Diseases: Structural Basis of Microtubule Affinity-Regulating Kinase 4 (MARK4) Inhibition. OMICS A Journal of Integrative Biology 09/2015; 19(11). DOI:10.1089/omi.2015.0111
14. **Mohd. Shahbaaz**, Krishna Bisetty, Faizan Ahmad, Md. Imtaiyaz Hassan: Functional Insight into Putative Conserved Proteins of *Rickettsia rickettsii* and their Virulence Characterization. Current Proteomics 05/2015; DOI:10.2174/1570164612021509031138 4

LIST OF CONFERENCES PRESENTATIONS

1. ORAL PRESENTATIONS

- Presented paper titled **"Classification and analyses of putative virulence factors from *Rickettsia rickettsii*".** National Symposium on Biophysics and Golden Jubilee Meeting of Indian Biophysical Society on 14-17 February, 2015 at Jamia Millia Islamia, New Delhi.

- Presented paper titled **"Molecular dynamics simulations of putative ABC transporter from *Mycobacterium tuberculosis*".** Institutional Research Day on 26th November, 2015 at the Durban University of Technology in Durban, South Africa.

- Presented paper titled **"Understanding the Structural Diversity of Kallikreins: Targets for anticancer Therapy".** 42nd National Convention of the South African Chemical Institute (SACI) on 29th November – 4th December, 2015 at Southern Sun Elangeni Hotel in Durban, South Africa.

2. POSTER PRESENTATIONS

- **Mohd. Shahbaaz,** Faez Iqbal Khan, Md. Imtaiyaz Hassan, Faizan Ahmad and Krishna Bisetty. Functional Annotation of Conserved Hypothetical Proteins from *Mycoplasma pneumoniae* 309 using computational methods. Center for High performance computing (CHPC) conference, National Meeting 1 - 5 December 2014, Kruger National Park, South Africa.

LIST OF CONTENTS

CHAPTER 3

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS AND SYMBOLS

TB	Tuberculosis
KZN	KwaZulu-Natal
MDR	Multiple-Drug Resistant
XDR	Extensively Drug-Resistant
HBC	High Burden Countries
HIV	Human Immunodeficiency Virus
HspX	Heat-shock protein
AhpC	Alkyl hydroperoxide reductase
KatG	Catalase: peroxidase
INF-γ	Interferon gamma
DFT	Density Functional Theory
MD	Molecular Dynamics
HPs	Hypothetical Proteins
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
ATP	Adenosine Triphosphate
BLAST	Basic Local Alignment Search Tool
EMBL	European Molecular Biology Laboratory
DDBJ	DNA Data Bank of Japan
PDB	Protein Data Bank
SCOP	Structural Classification of Proteins
CATH	Class, Architecture, Topology and Homologous superfamily
BRENDA	Braunschweig Enzyme Database
HMM	Hidden Markov Models
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool
3-D	Three Dimensional
AMBER	Assisted Model Building with Energy Refinement
CHARMM	Chemistry at Harvard Molecular Mechanics
GROMACS	Groningen Machine for Chemical Simulation
NAMD	Nanoscale Molecular Dynamics
PBC	Periodic Boundary Conditions
PME	Particle-Mesh Ewald
GB	Generalized Born model
PE	Poisson Equation
SDM	Site Directed Mutator
HOMO	Highest Occupied Molecular Orbital
LUMO	Lowest Unoccupied Molecular Orbital
EMB	Ethambutol
RMSD	Root Mean Square Deviation
Rg	Radius of Gyration
RMSF	Root Mean Square Fluctuation
INH	Isoniazid
WT	Wild Type
RIF	Rifampicin
RpoB	RNA polymerase β-subunit

SM	Streptomycin
LINCS	Linear Constraint Solver
NVT	Constant Number of Particles, Volume And Temperature
NPT	Constant Number of Particles, Pressure And Temperature
SIFT	Sorting Intolerant from Tolerant

CHAPTER 1

INTRODUCTION

This chapter provides details about the global epidemiological conditions of the Tuberculosis (TB), along with the severe conditions in the province of KwaZulu-Natal (KZN), South Africa (SA). The causes of TB epidemics in KZN and the current progress of the therapy are included in this section. This chapter also provides the general understanding of the co-infections appeared during the TB infection and the resistance mechanisms present in *Mycobacterium tuberculosis* that leads to its adaptation to the host environment. Furthermore, the details about the genome of *M. tuberculosis* along with the *in silico* techniques available for the functional analysis are also included in this section. This is followed by the aims and objectives and a brief outline of the thesis.

1.1 Tuberculosis (TB)

TB is amongst the oldest known diseases infecting the human race and still is one of the major causes of mortality around the world, resulting in about two million deaths with an estimated nine million new cases per year (Smith, 2003). In most cases, about 10% of the infected individuals acquire symptomatic TB. Despite the continuous advancement in the techniques aimed at early diagnosis and treatment, the eradication of TB still remains one of the biggest challenges. In the past few decades, a small number of drugs were developed against the infection of TB (Laurenzi et al., 2007). The first line drugs include isoniazid, rifampicin, pyrazinamide, ethambutol, and streptomycin formulated nearly 40 years ago and still forms the primary source of therapy (Laurenzi et al., 2007).

The erratic, limited or ineffective therapy leads to the development of tolerance against one of the first line drugs and resulted in the Multiple-Drug Resistant (MDR) conditions in the infected patients (Laurenzi et al., 2007). The critical cases of MDR may transform into Extensively Drug-Resistant (XDR) TB, in which the pathogen can tolerate the actions of second-line drugs such as kanamycin, capreomycin, and fluoroquinolones (Laurenzi et al., 2007).

In 2014, about 20% of the reported cases developed MDR-like conditions. Approximately 300, 000 patients developed the MDR TB which lead to around 190, 000 deaths worldwide (Laurenzi et al., 2007). Therefore, there is a need to enhance the current treatment methodologies against the TB infection as well as the introduction of new drugs targets is required. This study deals with the extensive analysis of the proteins involved in the resistance mechanisms and identification of the new drug targets in the genome of *M. tuberculosis*.

1.2 Epidemiology

The global estimate of TB occurrence is about 133 cases per 100, 000 individuals, with an average rate of reduction in the TB cases was reported to be around 1.5% - 2.1% during the past four years (http://www.who.int/tb/publications/global_report). The highest number of TB infections were reported in the Asian countries (58%) followed by the African nations (28%), with a smaller proportion of cases were reported in the American and European continents along with the rest of the world (http://www.who.int/tb/publications/global_report). The World Health Organization has classified 22 nations as “High Burden Countries” (HBC) which are described in Figure 1.1. Around 83% of the TB incident cases were originated from these countries, with India, China, Indonesia, Pakistan, Nigeria, and South Africa being the top contributors (http://www.who.int/tb/publications/global_report). The South African subcontinent experiences the highest incidence rates as well as the mortality rates due to the prevalence of MDR and HIV co-infection. Therefore, understanding and identification of novel mechanisms associated with the development of MDR-like conditions are required for controlling the high mortality rates in South Africa.

1.2.1 Tuberculosis in South Africa

Regardless of the high global prevalence rates, the Southern African nations such as

Mozambique, South Africa, Namibia, Zimbabwe, and Swaziland showed the relatively higher number of TB incidences. In South Africa, approximately 450, 000 active TB cases were reported in 2014. The number of cases has increased to four folds because of the high prevalence of HIV co-infections. The TB incidence rate of South Africa is around 834 cases per 100, 000 individuals (Figure 1.1), with the provinces of Free State, Northern Cape, and KwaZulu-Natal showed highest mortality rates (Table 1.1).

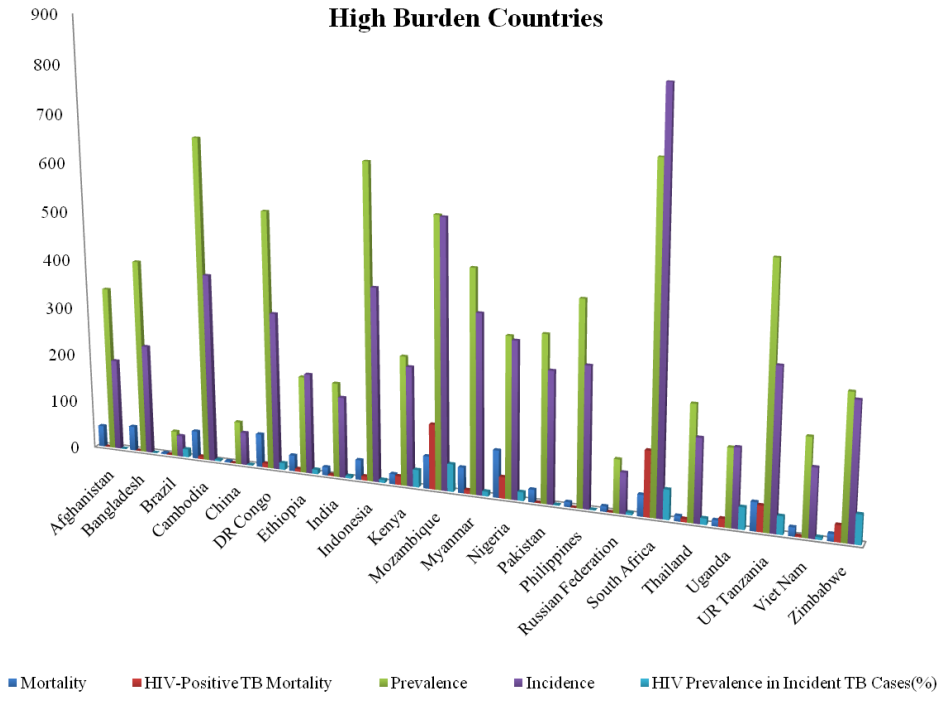


Figure 1.1: Trends in estimated rates of mortality, prevalence and incidence in HBC in 2014

The prevalence rate of the MDR in the patients of KZN and Mpumalanga were around 8.3% and 8.6% respectively (http://www.healthlink.org.za/). The major cities in KZN such as Durban and Pietermaritzburg are amongst the worst hit places, which showed a rapid increase in the MDR and XDR TB cases. The incidence rate of 843 per 100, 000 population were reported in KZN (http://www.kznhealth.gov.za/), with highest incidences rate reported in the district

eThekwini (237/100, 000 individuals), while Umzinyati district showed the highest number of MDR and XDR cases (Moodley et al., 2011). Therefore, uncovering the new pathogenic pathways followed by the identification of new virulence factors, which can act as new drug targets, is crucial for controlling the constantly increasing rates of TB infections in South Africa.

Table 1.1: Statistics of the provincial mortality rates and reported percentage of MDR cases

S. No.	Province	Mortality rate (per 100 000)	Prevalence of MDR (%)
1	Eastern Cape	87	6.0
2	Free State	99	5.6
3	Gauteng	49	5.7
4	KwaZulu-Natal	81	8.3
5	Limpopo	62	4.9
6	Mpumalanga	80	8.6
7	Northern Cape	88	5.0
8	North West	84	5.2
9	Western Cape	40	5.2

1.2.2 Co-infections

The increased severity of TB is attributed to the presence of diverse co-infections. An extensive study showed the existence of at least 24 diverse co-infections that may be present during TB prevalence (Li and Zhou, 2013). The hydatid disease, leishmaniasis and malarial parasites such as *Plasmodium falciparum, P. berghei* and *P. yoelii* form the majority of the cases and resulted in the decrease of humoral as well as the cellular immune responses (Li and Zhou, 2013). *P. yoelii* co-infection may result in the failure of the tissue growth in the certain region of the human body and increases the chances of the mortality as well as may alter the effectiveness of the TB vaccines (Li and Zhou, 2013). The triple infections of TB, leishmaniasis, and lepromatous leprosy were reported in the seasoned French patient, a rare condition in which insensitivity to IL-12 of the patient's T cells was reported (Li and Zhou, 2013). Furthermore, the

occurrence of opisthorchiasis also complicates the infection of TB by increasing the intolerance against the antibacterial therapy and may deteriorate the prognosis of both the diseases (Li and Zhou, 2013). Moreover, intestinal helminths may cause a range of immunomodulations such as elevated level of IgE, Th2-type cytokines as well as up-regulation of T-cell activities and activation of chronic immune responses.

Furthermore, among the 9.6 million reported TB incidences around the world, around 1.2 million patients are also living with HIV co-infections (http://www.who.int/tb/publications/ global_report). The percentage of TB incidences co-infected with HIV was found to be highest in African countries. In this region, around 32% of the reported cases were co-infected with HIV, which accounts for 74% of all the TB-HIV co-infections around the world. The countries located in the Southern Africa showed the presence of more than 50% HIV co-infection among the reported cases (http://www.who.int/tb/publications/global_report).

1.3 Research Problems

The manifestations resulting from TB primarily affects the pulmonary system, central nervous system, and bones (Smith, 2003). The infection is initiated by the deposition of *M. tuberculosis* coupled aerosol droplets on the alveolar surface (Smith, 2003). After colonizing the human body, a variety of virulence factors present in the *M. tuberculosis* leads to its survival inside the host. The primary factors of its pathogenesis are the proteins and lipids associated with the biosynthesis, transport and the degradation of the components that form the complex cell wall envelope (Forrellad et al., 2013). Moreover, the proteins such as Heat-shock protein (HspX), Alkyl hydroperoxide reductase (AhpC), and Catalase: peroxidase (KatG) stabilize the cellular framework of *M. tuberculosis*, necessary for its survival in the diverse stress conditions

present inside the human body (Forrellad et al., 2013). Similarly, the transporter proteins which perform the uptake of the ions such as iron and manganese are considered as essential in maintaining the cellular integrity of the *M. tuberculosis* inside the macrophages (Forrellad et al., 2013). Moreover, the higher pathogenicity of the members belonging to the genus Mycobacterium is attributed to the presence of secretory systems such as ESX-1, ESX-3, and ESX-5 which are helpful in the production of the modified proteins (Forrellad et al., 2013). The *M. tuberculosis* also synthesized and produces the toxins in the form of Phospholipase C which performs the hydrolysis of phospholipid substrates and is critical for its virulence (Forrellad et al., 2013). The diversity of the resistance mechanisms as well as biomolecules facilitating its survival against the current therapies is discussed below.

1.3.1 Resistant Mechanisms in *M. tuberculosis*

The infection of the *M. tuberculosis* occurs by its transfer from infected individual to a healthy person. Therefore, in order to survive inside the host environment, the *M. tuberculosis* modify the pattern of its gene expression (Stanley and Cox, 2013). These deviating protein expressions take place in numerous immunity providing cells such as granulomas, macrophages as well as in the lesions developed in the pulmonary region (Stanley and Cox, 2013). Within the internal environment of macrophages, the *M. tuberculosis* experiences the high acidic conditions as well as oxygen intermediates of high reactivity (Stanley and Cox, 2013). While in the granulomas, the *M. tuberculosis* encounters toxic proteases and lipases followed by low oxygen availability (Stanley and Cox, 2013). Furthermore, a high temperature is present in the TB-infected human body as *M. tuberculosis* activities may result in persistent fever conditions (Stanley and Cox, 2013). The diversified resistant mechanisms are explained below in details:

1.3.1.1 Survival in Macrophages

The pulmonary alveoli are the primary sites of the *M. tuberculosis* infection, where they encounter the macrophage cells (Stanley and Cox, 2013). As these macrophages are functionally inactive, the *M. tuberculosis* enters inside the lysosomal organelle and start replication (Stanley and Cox, 2013). After activation by Interferon gamma (INF-γ), these cells become metabolically efficient and produce the oxidative and acidic stress conditions against the *M. tuberculosis*. In order to sustain its growth, *M. tuberculosis* increases the lipid metabolism and up-regulates the activities of enzymes such as isocitrate lyase, which is the key enzyme of the glyoxylate cycle and utilizes the fatty acids instead of carbohydrate as the energy source (Stanley and Cox, 2013). Furthermore, *M. tuberculosis* also attenuates the mechanisms of phagosome–lysosome fusion, which is the key mechanism involve in the inhibition of pathogenic growth (Stanley and Cox, 2013). The *M. tuberculosis* showed the presence of anionic trehalose glycolipid (Sulfatides), which perform the characteristic anti-fusion activities. It also showed the presence of high ammonia which also affects the inhibition of the fusion process as the ammonium chloride alkylation of the intra-lysosomal space and reduces the activities of the lysosomal enzymes (Stanley and Cox, 2013).

Moreover, the process of phagocytosis lowers the phagosomal pH (Figure 1.2) which is either toxic to the pathogens or may inhibit their growth (Vandal et al., 2009). The *M. tuberculosis* showed the presence of intrinsic resistant mechanisms against the increased acidic conditions which enable its survival in the pH conditions as low as 4 (Vandal et al., 2009). A variety of characteristic genes such as isocitrate lyase, LipF, and PhoP, showed the up-regulations in the high acidic conditions (Vandal et al., 2009). The isocitrate lyase is the key enzyme which leads to the formation of glyoxylate as well as succinate by catalyzing the

cleavage of isocitrate (Vandal et al., 2009). Similarly, LipF which is characterized to be a lipase/esterase and may be involved in the modification of the bacterial cell wall (Vandal et al., 2009). It also forms the part of phoP/R regulon, a two component system, is a downstream regulator which provide response against the acidic stresses (Vandal et al., 2009). Moreover, the ompATB, a porin protein shown the activation in low pH conditions and enables the transportation of the ammonia in the phagosomal space which neutralizes the acidic environment (Vandal et al., 2009). Similarly, MgtC a transporter protein involved in the translocation of the magnesium and is required for maintaining the cell wall integrity in the acidic environment. It also acts as a cofactor for a variety of regulatory enzymes, important for the survival of the *M. tuberculosis* in the acidic stresses (Vandal et al., 2009).

Furthermore, Rv2136c and Rv3671c showed the functionalities of serine protease and undecaprenyl pyrophosphate phosphatase respectively were found to play a significant role in the acid resistance (Darby et al., 2011, Biswas et al., 2010). The Rv2136c also shown resistance to lipophilic antibiotics, sodium dodecyl sulfate, reactive intermediates and elevated temperature (Darby et al., 2011). Likewise, Rv3671c is a member of the chymotrypsin-like family and is crucial for the development of *M. tuberculosis* resistances to oxidative stresses along with the protection from the phagosomal acidification (Biswas et al., 2010). In this study, the structural basis of these virulent proteins involved in the development of acid tolerance was analyzed using constant-pH Molecular Dynamics (MD) simulations, explained in Chapter 3.

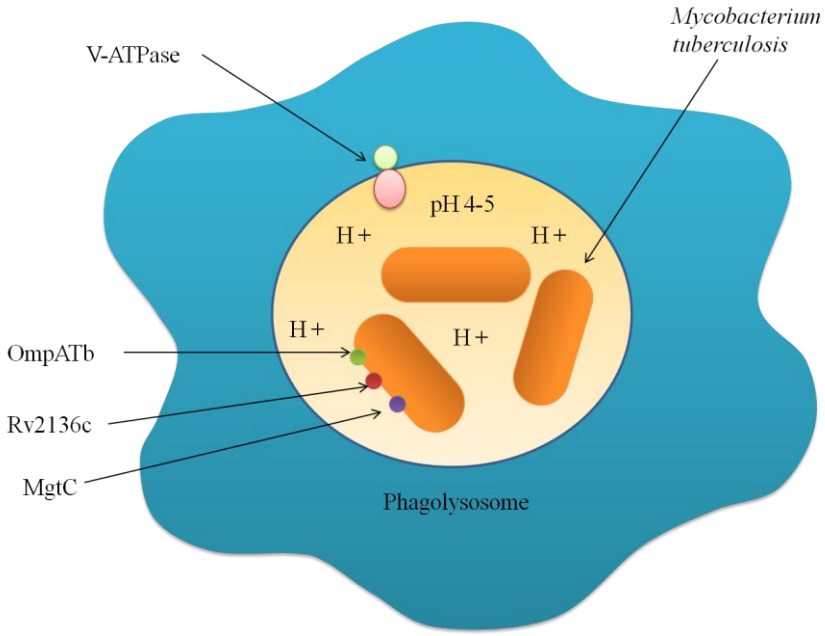


Figure 1.2: Schematic view of the phagosomal environment after fusion with lysosome with *M. tuberculosis* present in the interior of the vacuole.

1.3.1.2 Drug Resistances

In order to cure the infection of TB, the first line of drugs such isoniazid, Ethambutol, rifampicin, streptomycin, and others, form the part of intensive drug therapy (Khosravi et al., 2012). The resistance to at least two of these first line drugs is considered as the development of MDR conditions of the TB. Several isolates around the globe showed the presence of the point mutations in the primary interacting biomolecules which may lead to the development of the drug resistances in *M. tuberculosis*. The pro-drug form of isoniazid is activated by the activity of the catalase-peroxidase enzyme KatG, present in the cellular framework of the *M. tuberculosis*. The mutations observed in the codon 315 of KatG as well as in the active site and promoter region of *inhA* operon may lead to the isoniazid resistance in *M. tuberculosis* (Khosravi et al., 2012). Furthermore, the mutations in *inhA* are also associated with the development of resistance against second-line drug such as Ethionamide (Morlock et al., 2003). Likewise, the over-

expression *rspA* resulted in the development of the pyrazinamide resistance, which is attributed to the loss of Alanine 438 at the C terminus of the protein (Smith et al., 2013). Similarly, Ethambutol which is designed to be an analogue of arabinose mainly targets the cell wall of the *M. tuberculosis* through the inhibition of the arabinosyl transferases. The latter is encoded by the genes *embA*, *embB*, and *embC* which form part of *embCAB* operon. The mutation in the codon 306 of the *embB* gene may lead to the development of the Ethambutol resistances in *M. tuberculosis* (Khosravi et al., 2012).

Furthermore, DNA-directed RNA polymerase encoded by the *rpoB* gene of the *M. tuberculosis* catalyzes the process of transcription using ribonucleoside triphosphates as the substrate molecules (Khosravi et al., 2012). It was observed that mutation in the 81 bp region of the *rpoB* gene, particularly at 516, 526 and 531 positions may result in the development of the rifampicin-resistant conditions (Khosravi et al., 2012). Moreover, the ribosomal proteins play a significant role in the translational precision. It is reported that streptomycin resistance is associated with the mutations in the ribosomal proteins encoding genes such as *rrs* and *rpsL* (Sreevatsan et al., 1996). In particular, the *rrs* gene which is also involved in the formation of resistance against second-line drugs such as amikacin, kanamycin, and capreomycin (Sowajassatakul et al., 2014). In this study, the deleterious effects of the point mutations present in the literature were analyzed on the sequence and structural framework coupled with the prediction of novel disease-associated substitutions. The structural bases of the drug-resistant mutations were studied using the techniques of Density Functional Theory (DFT), molecular docking and MD simulations (described in Chapter 3).

1.3.1.3 Adaptation to High Temperature

The patients infected with TB showed many characteristic symptoms including the development of high fever as well as the night sweating (Riffo-Vasquez et al., 2004). These conditions are not suitable for its growth as high temperature may cause the degradation of the proteins by the process of unfolding (Riffo-Vasquez et al., 2004). In order to withstand this condition, the *M. tuberculosis* showed the up-regulation of the molecular chaperonins, which prevent the aggregation of the other proteins by providing the favorable conditions for their refolding (Riffo-Vasquez et al., 2004). The Acr-2 (α-crystalline protein) was shown to be activated by the heat shock and showed the presence of activities related to chaperonins (Riffo-Vasquez et al., 2004). Several other heat shock proteins such as HSP70 and 65Kd HSP take part in the refolding of the heat damaged proteins by forming complexes with them (Riffo-Vasquez et al., 2004). After the analyses of *M. tuberculosis* genome, numerous proteins were annotated as the Heat shock proteins (HSP) and may take part in the development of the resistance against high-temperature conditions (Discussed in Chapter 3). Additionally, the protein Rv3906c (Uniprot ID - O05439) showed high similarities to pullulanase, a thermophilic protein and may take part in the survival against high-temperature conditions. The stability profile of Rv3906c at different temperatures was generated by using MD simulation studies, illustrated in Chapter 4.

1.3.2 Genome of *M. tuberculosis*

The genome of *M. tuberculosis* H37Rv contains a single chromosome organized in a circular manner with around 4,411,529 bp which are structured into 4047 genes, was sequenced and published in 1998 (Figure 1.3). The GC content of the genes was assessed to be about 65.61% (Cole et al., 1998) and translated into 3906 proteins, which are categorized into 25 classes of the proteins (Figure 1.4).

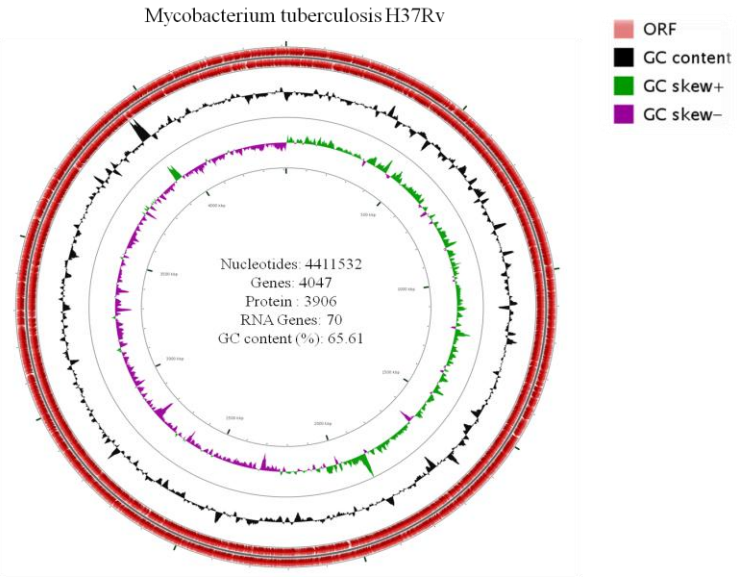


Figure 1.3: The diagrammatic view of the sequenced circular genome of *M. tuberculosis*.

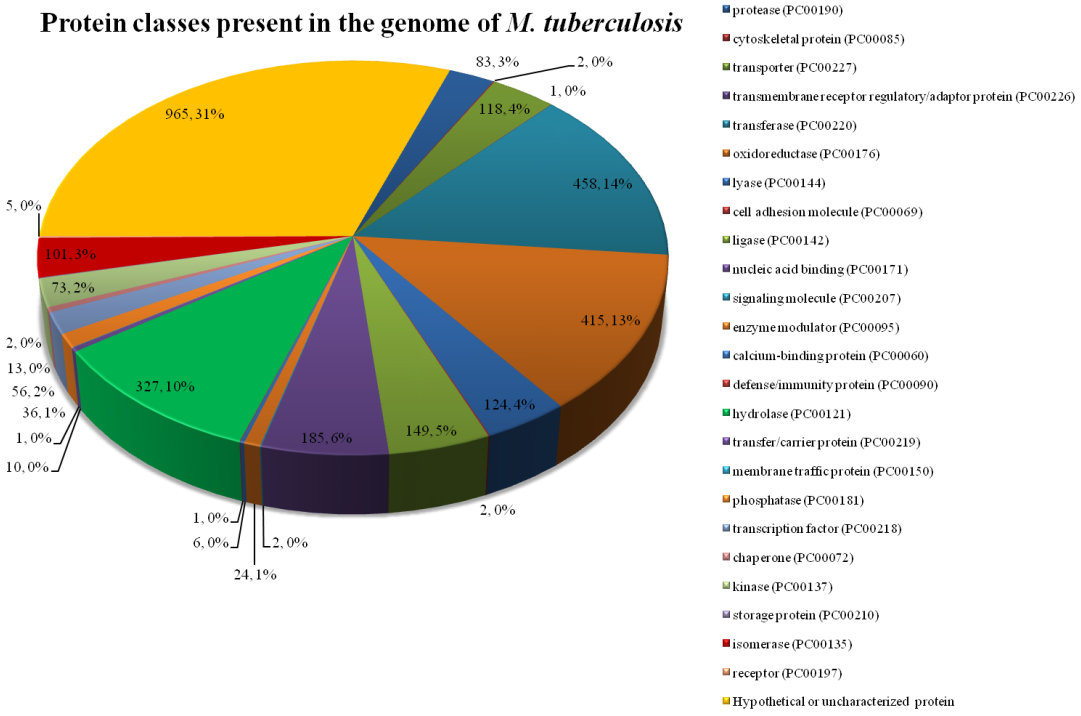


Figure 1.4: Different classes of the proteins present in the genome of *M. tuberculosis*.

1.3.3 Hypothetical Proteins (HPs)

In the group of 3,906 proteins, 1000 are classified as the "hypothetical proteins" (HPs) which form the major portion of the *M. tuberculosis* genome (Figure 1.4). The HPs are the proteins predicted from the genomic data, but their existence is not validated experimentally at the expression level (Mohd et al., 2016). The major portions of the bacterial genomes are either uncharacterized or hypothetical (Mohd et al., 2016). Information regarding this section of proteins is necessary for the completion of the knowledge obtained from the diverse genomic and proteomic studies (Mohd et al., 2016). Their annotations may lead to the detection of new structural topologies as well as new functionalities which may uncover additional metabolic pathways, crucial for the survival of the pathogenic bacteria (Mohd et al., 2016). Furthermore, these HPs can also function as the therapeutic targets in the process of drug design and discovery (Mohd et al., 2016). The functions and the virulence of 1000 HPs were predicted which discussed in depth in Chapter 3. The conformational behaviors of the virulent HPs were assessed using the diverse molecular modelling techniques.

1.4 *In silico* methods for functional annotations

There are 16,345 sequenced genomes deposited in NCBI database (http://www.ncbi.nlm.nih.gov/genome) contributing to about million of protein sequences present in the publically available protein databases (http://www.ncbi.nlm.nih.gov/proteinclusters/). The experimental determination of the functionalities of these proteins would be time-consuming and not cost-effective. The advent in the field of Bioinformatics and computational modelling techniques has been a game changer in the discovery of new genes and their hidden functionalities. The *in silico* techniques provide an alternative for predicting the functions of the uncharacterized proteins or HPs with high accuracy. Therefore, the role of computational biologists and chemists in the

discovery of new drug targets has been broadened to include the advancement of *in silico* methodologies. The term *in silico* was first used by Pedro Miramontes at a workshop in Los Alamos, New York (1989) in an article entitled: "DNA and RNA physicochemical constraints, cellular Automata and Molecular evolution" (Miramontes, 1989).

The protein functionalities are divided into various mutually dependent levels such as molecular functions, cellular compartments as well as biological processes (Lee et al., 2007). The molecular function illustrates the activity of the protein at molecular levels such as selective catalysis, while biological functions depict the broader functionalities such as metabolic pathways in which it performs its characteristic molecular activities (Lee et al., 2007). Similarly, the cellular compartments involve the portion of the cells in which the protein performs its functions. The *in silico* methods can predict the protein functionalities at all three levels along with the characteristic interactions a protein experiences during the course of biological processes (Lee et al., 2007). The sequence-based functional annotations utilize the principles of homology detection between the different sets of proteins (Lee et al., 2007).

Despite increasing accuracy and sophistication in the computational methodologies aimed at deciphering the functionalities by utilizing the sequences of the proteins, still, a very large portion of the proteins present in the genomes of the organisms remain "uncharacterized" (Lee et al., 2007). Therefore, due to high conservation of the protein structures as compared to the protein sequences, the understanding of the specific fold adopted by the given sequence can be used for predicting the function of the uncharacterized proteins (Lee et al., 2007). In this study, all the available methods were utilized for predicting the functionalities of 1000 HPs present in the genome of *M. tuberculosis* (Illustrated in Chapter 3).

1.5 Aims and Objectives

The broader goals of this study are to identify different classes of the functionalities present in the group of 1000 HPs with the view to the classification of new virulence factors which can serve as new drug targets for existing therapy against TB. The present study also focuses on the understanding of the diverse mechanisms responsible for the development of the resistances in *M. tuberculosis* against the available therapies using Computational Chemistry methods.

Objectives:

1) To study the diverse structural basis of proteins involved in the development of the resistance and survival mechanisms of *M. tuberculosis* by using MD simulations coupled with Density Functional Theory (DFT) techniques.
2) To analyze the proteome of the *M. tuberculosis* by classifying and functionally annotating the 1000 HPs by using the varieties of available *in silico* protocols. And to predict the functions of HPs by computing diverse physicochemical parameters, cellular localizations, functional predictions and domain annotations attributes as well as assessing the virulence associated with each HPs.
3) To perform the structure predictions of all the virulent proteins with highest prediction scores, using the concepts of homology modelling as well as *ab intio* methods. To also analyze the structural features responsible for their virulence nature.
4) To study the conformational behavior of all the major virulent proteins using the concepts of Molecular Dynamics (MD) simulations in explicit water conditions.
5) To validate the characterized virulence factors using the techniques of molecular cloning and analyzing its expression at the molecular level.

1.6 Thesis Outline

After giving the brief outline about the epidemiology and infection of *M. tuberculosis* along with its resistance to current therapies and its solution in this chapter, further chapters in this thesis are expanded as follows:

1. Chapter 2: This chapter deals with the review of literature primarily based on the classes of virulent proteins which are important drug targets in *M. tuberculosis* and techniques available for the functional annotation of the HPs. A discussion on the accuracy of the available methods and the generation of new drug targets is also presented. This chapter also includes the different strategies available for protein modelling and structure-based drug design.
2. Chapter 3: Deals with the computational methodologies used for the analyses of the structural basis of the drug resistance, acid resistance as well as the techniques which involve the combination of sequence and structure-based functional annotations were used for predicting the functionalities of 1000 HPs. The outcomes obtained through this extensive study are discussed and analyzed on the basis of available literature in this chapter.
3. Chapter 4: Involve the summary of the outcomes obtained regarding the experimental validation of the virulent HPs. In addition, their roles in the biochemical processes were validated by using the principles of molecular modelling.
4. Chapter 5: The concluding remarks summarize the novelty and achievements of this study.
5. Chapter 6: Include the brief outlines of the future works going to be initiated on the basis of the current study.

CHAPTER 2

LITERATURE REVIEW

This chapter highlights the historical aspects of *in silico* analyses, followed by the diversity of the algorithms and techniques available for sequence analyses and function predictions of the proteins. The protocols available for the molecular modelling of the protein structures as well as for studying their conformational and thermodynamic properties are discussed in the consecutive sections. Furthermore, the diversity of the virulence factors in *M. tuberculosis* provides an understanding of the virulent characteristics associated with the proteins of this pathogen.

2.1 Historical aspects of *in silico* analyses

In biological sciences, *in silico* analyses signifies changes from an analytical to a computational framework by the creation of computer-based artificial systems which can mimic the biological behavior. The interdisciplinary studies of biological sciences in the 20th century included a reductionist union with the concepts of physics and chemistry, which is complimented by improving the accuracy of the instrumentations (Ouzounis and Valencia, 2003). These new ideologies lay the milestone in the field of the "Bioinformatics" and lead to the formulation of new *in silico* methodologies which revolutionize the scientific research (Ouzounis and Valencia, 2003). The preliminary development in the computational studies took place in pre- or early 1970s as some of the fundamental problems in molecular biology offered some significant algorithmic problems (Ouzounis and Valencia, 2003).

These queries arose due to discoveries of the helical structure of DNA (Watson and Crick, 1953), translation of the genetic information into proteins (Gamow et al., 1956), the structural features prevailing in the protein molecules (Crick, 1953), the process of gene regulations (Britten and Davidson, 1969) and many more, are the basis for some challenges that were addressed in following decades by using the computational techniques. The essential

concepts of the computer science such as theory of information (Shannon and Weaver, 1963) as well as computation (Chaitin, 1966), randomization of the strings (Martin-Löf, 1966), context-free grammars (Chomsky, 1959), cellular automata (Neumann, 1966) and many other significant algorithms emerged in a parallel manner between 1950s-60s.

These early methodologies combined both computational and experimental knowledge to better understand the behavior of the biomolecules (Ouzounis and Valencia, 2003). Newer insights were obtained regarding the evolution of the genes as well as proteins using the concepts of the molecular homology (Florkin, 1962), identification of the evolutionary patterns (Zuckerkandl and Pauling, 1965a, Zuckerkandl and Pauling, 1965b), informational properties associated with the sequences of the DNA and proteins (Gatlin, 1966, C Nolan and Margoliash, 1968). These studies were complemented by the development of the phylogenetic trees using the properties generated from the sequence alignment of DNA and proteins (Fitch and Margoliash, 1967). This era marks the beginning of computational biology as several key concepts were developed for the first time in this period (Ouzounis and Valencia, 2003).

These innovations included the development of the sequence alignment algorithms (Gibbs and McIntyre, 1970) followed by formal studies on the preferential substitution of the residues (Clarke, 1970) and related studies on the primary structure of proteins (Krzywicki and Slonimski, 1967). Moreover, the preferences of the amino acids in the secondary structure (Ptitsyn, 1969), as well as the spatial constraints in the protein structures, were first studied in this phase (Pain and Robson, 1970). The development of the molecular graphics technologies was also initiated along with the invention of the helical wheel representations of the secondary structural elements of the proteins (Dunnill, 1968). The central dogma also postulated in the 1970s after the seminal detections of the process of transcription as well as the translation (Crick,

1970).

Consequently, the above-mentioned discoveries and inventions led to the development of the computational algorithms aimed at solving the problem in the molecular biology (Ouzounis and Valencia, 2003). The *in silico* protocols such as calculations of the solvent accessible surface areas of the protein structures (Lee and Richards, 1971), calculation of the amino acid mutational rates (Koch, 1971), addition of the parsimony algorithms regarding the determination for the topologies of the phylogenetic trees (Fitch, 1971) and prediction of the structural features associated with the RNA molecules (Tinoco et al., 1971). Furthermore, one of the most prominent innovations of this phase was the formulation of the molecular clock hypothesis which is formed by the combination of molecular evolution along with the population genetics (Kimura and Ota, 1972). The string comparison processes were also intensifying in this decade that leads to the sophistication and increased accuracy in the processes of the sequence alignment (Sankoff and Sellers, 1973).

The further development of the new sequence alignment algorithms (Ullman et al., 1976) followed by the innovation in the analyses, visualization, and prediction of the protein structure tools, were aimed at solving another "genetic code" in molecular sciences, *the problem of protein folding* (Chothia, 1975, Chothia et al., 1977). The solution was achieved by the development of the first algorithm regarding the secondary structure predictions by utilizing the sequences of the proteins (Chou and Fasman, 1974). In addition to this, the accuracy of the structure prediction methods increased by the introduction of more sophisticated tools as well as distance geometry which was aimed at the calculation of distance constraints from the protein structures (Crippen, 1977, Feldmann, 1976). Moreover, in order to increase the distribution of the generated biological data, the computational archives or databases were the compiled for the storage,

analyses, and curation of nucleic acids as well as protein sequences (Dayhoff, 1978) and their structural features (Bernstein et al., 1977). These trends of exponential increase in biological data continue till date.

The 1980s and the following decades marked the establishment of the computational biology and the related fields such as computational chemistry, an independent discipline of science (Ouzounis and Valencia, 2003). During this period, the biological and chemical data increased exponentially and more proficient algorithms were formulated in order to cope with the ever-increasing volume of the information in the publically available databases (Ouzounis and Valencia, 2003). After this period the computational biology was subdivided into various sub-disciplines, namely, *Sequence analysis*, *publically available databases*, *protein function predictions* and *molecular modelling*. The advancement took place in each sub-discipline are discussed below:

2.1.1 Sequence analysis

The development in the *in silico* process of the sequence analysis, involving calculations of the evolutionary distances between the diverse protein sequences (Sellers, 1980) and assessment of the sequence matching (Ouzounis and Valencia, 2003), were initiated by the development of the key alignment protocols such as Smith–Waterman algorithms for the sequence alignment by using the dynamic programming (Smith and Waterman, 1981a, Smith and Waterman, 1981b) and database searching algorithm of FASTA (Lipman and Pearson, 1985). Initially, the protocols based on the theoretical computer science such as analyses of the repeats (Guibas and Odlyzko, 1980) were used for the biological Sequence analysis and later modified to perform parallel analyses in diverse sequences which were later utilizing the matrix based methods (de Wachter, 1981, Fristensky, 1986). During this phase, the major development

in the automation took place which leads to the wide usage of multiple sequence alignment (Carrillo and Lipman, 1988). One of the primary applications of the sequence analysis aimed at the identification of the significant protein motifs such as the discovery of ATP binding motifs in the non-homologous proteins (Walker et al., 1982) along with the identification of the functionally conserved motifs such as zinc-finger (Klug and Rhodes, 1987) and leucine-zipper (Landschulz et al., 1988). The other applications include the identification of the homology between diverse sigma factors present in the bacterial systems (Gribskov and Burgess, 1986) along with the characteristic of the protein signal peptides (von Heijne, 1981). The performance of the protocols on the local machines showed varied efficiencies due to their computational intensive nature (Gotoh, 1987).

The other inventions that took place during this period was the development of the algorithms regarding the prediction of the RNA folding (Dumas and Ninio, 1982), identification of the translation initiation sites (Schneider et al., 1986) and open reading frames (Fickett, 1982), antigenic determinants features (Hopp and Woods, 1981) as well as the computation of the evolutionary trees (Felsenstein, 1982). Furthermore, the revolution in the analyses of the primary sequences was observed by the development of the Basic Local Alignment Search Tool (BLAST) which is a modified heuristic method, which was faster as compared to the previously developed algorithms (Altschul et al., 1990). This protocol made the analyses of the large genome sequences available in the biological databases a practical approach because of the increased automation (Ouzounis and Valencia, 2003).

2.1.2 Amplification of databases

The databases aimed at the storage and curation of the biological and chemical information were developed in order to fulfill a wide range of purposes, include varied types of

data at heterogeneous coverage and their curation was performed at several levels with distinctive methods (Kelly and Meyer, 1980). The initial stages of database development and curation for data quality restrain and compilation promptly progressed with the emergence of three major repositories for nucleotide data submission (Philipson, 1988), GenBank (Bilofsky et al., 1986), the European Molecular Biology Laboratory (EMBL) data library (Hamm and Cameron, 1986) and DNA Data Bank of Japan (DDBJ) (Tateno et al., 2002). In addition to these resources, several other initiatives such as BIONET (Kristofferson, 1987) and EMBNET (Felsenstein), were started in order to facilitate the distributions and accessibility of the biological data.

This period also marked the emergence of a variety of case-specific hardware platforms which enabled the efficient analyses of the primary structure of the proteins along with the development of the technology aimed at the construction of the relational databases (Islam and Sternberg, 1989). After this period the biological and chemical databases increased at an exponential rate (Figure 2.1). The databases were classified into various categories, namely, *protein families' database* of Pfam (Bateman et al., 2002) and Interpro (Hunter et al., 2009), *protein structure databases* such as Protein Data Bank (PDB) (Bernstein et al., 1977), Structural Classification of Proteins (SCOP) (Murzin et al., 1995), and CATH (acronym for Class, Architecture, Topology and Homologous superfamily (Orengo et al., 1997). The other categories include database aimed at the storage of the information related to the diseases, while some contain the information about the evolution such as knowledge of the taxonomic groups (Ouzounis and Valencia, 2003). There are databases that contain information derived from the scientific literature such as of derived data, collecting and systematizing the body of knowledge from the scientific literature such as ChEMBL (Gaulton et al., 2012) and BRENDA (Schomburg

et al., 2004). On the basis of 2016 report of Molecular Biology Database Collection published in the peer-reviewed journal of *Nucleic Acids Research*, the number of publically available biological and chemical databases has increased to 1685 (Rigden et al., 2016).

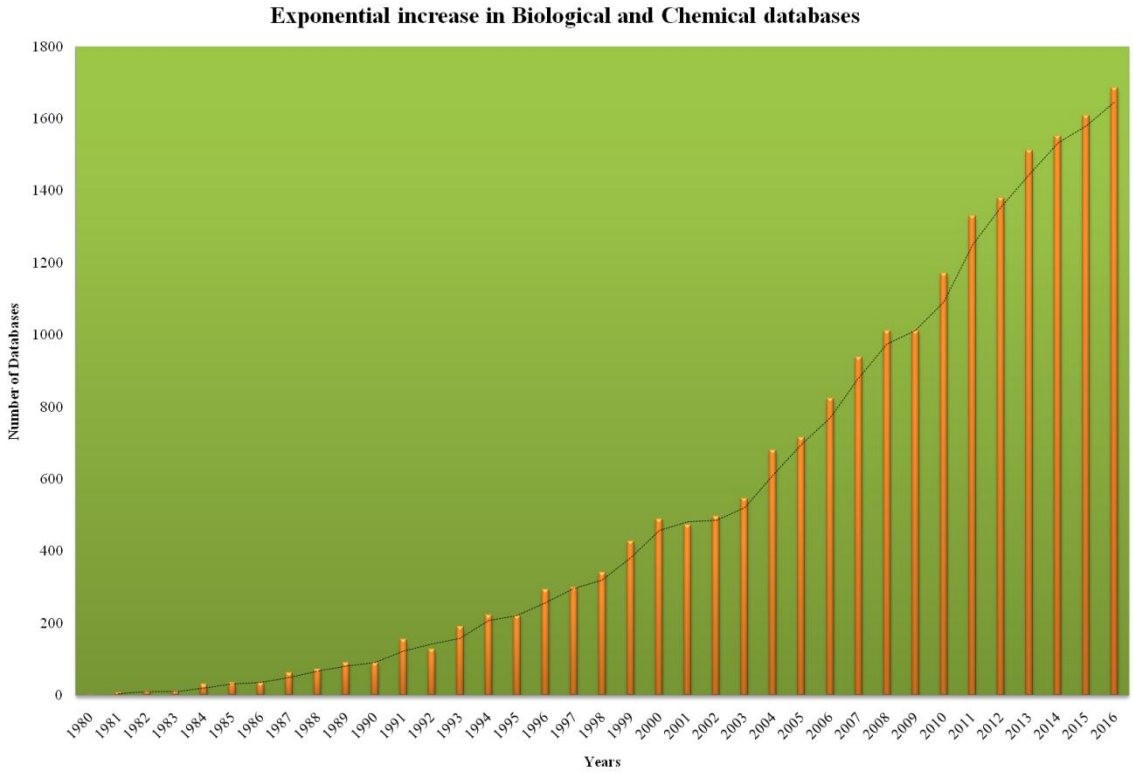


Figure 2.1: Exponential trend observed in the number of the publically available databases from 1980 to 2016.

2.1.3 Protein function predictions

The publication of the *Atlas of Protein Sequence and Structure* and its successive volumes (Dayhoff, 1978), form the basis for the functional prediction of the proteins. This book describes the principles of multiple sequence alignments as well as a graphical depiction of the protein structures (Hodgman, 2000). This study highlighted how the conserved patterns can be displayed, which arises from the functional and structural roles played by the residues at every position of the alignment (Hodgman, 2000). Short functional peptides have been identified as "*motifs*". Several motifs such as signal peptides present in the N-glycosylation site were

considered as the regular expression (Marshall, 1972). These correlations lead to the development of the databases such as 200-motif resource and later PROSITE (Bairoch, 1991), aimed at the storage and curation of the conserved motif sequences (Hodgman, 2000). The techniques for the identification of the motifs and their comparisons with the query sequences have become highly sophisticated and faster because of the development of computational infrastructures (Hodgman, 2000). The identification of the regular expression in the biological sequences leads to the development of other theories of weight matrices (Staden, 1984), neural networks (Qian and Sejnowski, 1988), perceptrons (Stormo et al., 1982), Hidden Markov Models (HMM) (Churchill, 1989), Bayesian statistics (Liu et al., 1995), sampling techniques (Lawrence et al., 1993) and several other important concepts (Hodgman, 2000). Consequently, the alignments were used for the identification of the conserved motif region in the protein sequences which were then utilized in the searching against the protein databases (Hodgman, 2000). This approach was utilized for the prediction of the functionalities of the protein belonging to the families of unknown functions. In this way, the functions of the helicase superfamily as well as the protein domains present in the RNA viruses were identified simultaneously (Hodgman, 1988). The Psi-Blast is an attempt to increasingly automate the process of the function prediction (Altschul et al., 1997).

After 2004, there was an exponential increase in the number of the sequenced genomes in the databases (Figure 2.2), which generates demand for the advancement in the high-throughput technologies along with Sequence analysis techniques of Bioinformatics (Mohd et al., 2016). As a result of these extensive analyses, the functional allocation to 40-70% of the genes was achieved, while 15-30% of the gene remains functionally "*uncharacterized*" because they have no recognizable conserved features or they belong to the families with unknown functions

(Mohd et al., 2016). These uncharacterized proteins form the basis of the current study as numerous hidden facts are present in the functionalities of these proteins. The annotation of the previously uncharacterized proteins as tRNA modification enzymes belonging to the Non-mevalonate or deoxyxylulose pathway as well as identification of the centrality of cyclic diguanylate in bacterial signaling indicated the importance of these classes of proteins in the understanding of the hidden mechanisms in the biological systems (Galperin and Koonin, 2004).

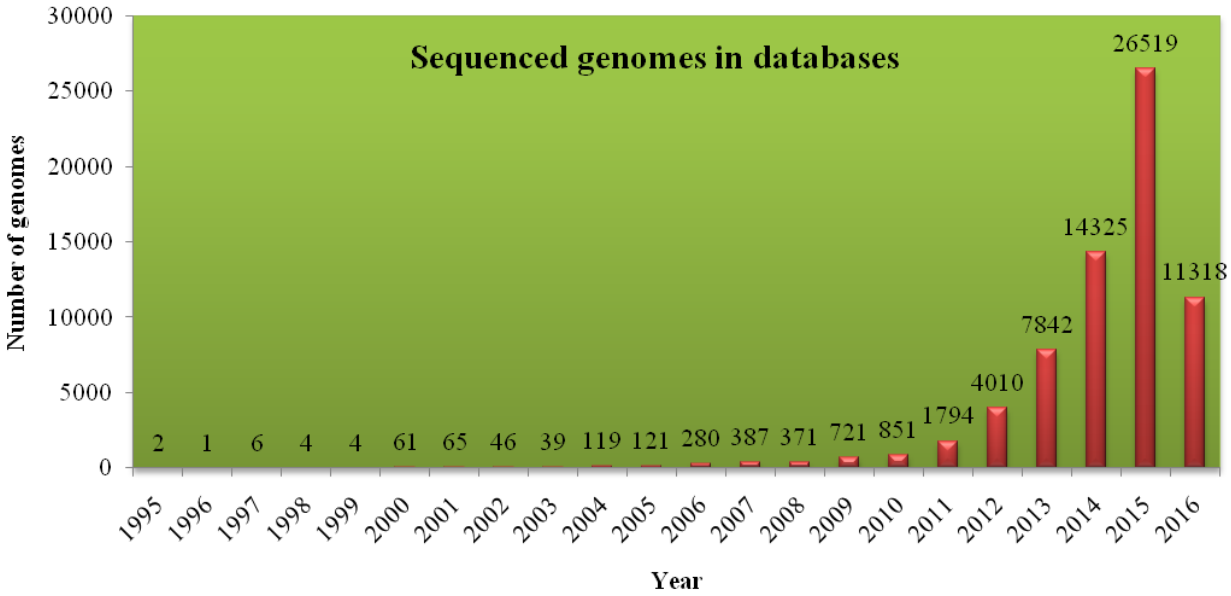


Figure 2.2 The number of the prokaryotic sequenced genomes in the biological databases increases exponentially from the time period of 1995 to 2016.

2.1.4 Molecular modelling

Molecular modelling signifies the general approach of illustrating complex biological as well as chemical structures in terms of rational atomic models, with the target of understanding and predicting the macroscopic properties associated with the systems on the basis of the knowledge obtained at the atomic scale. The molecular modelling is widely used to model novel materials by utilizing the precise prediction of physical attributes of realistic systems. In studying biological macromolecules, the computational techniques enable the analyses of structural and

conformational behavior of the biomolecules as described in the following sub-sections:

2.1.4.1 Protein structure prediction

From 1980 onwards, a major progress in the field of protein structure prediction (Ouzounis and Valencia, 2003) and analyses was reported by the introduction of improved approaches for the visualization and representations of the protein structure. These techniques include derivation of the coordinates from the stereo illustrations (Rossmann and Argos, 1980), hydrophobicity plots (Kyte and Doolittle, 1982), domain architectures (Rashin, 1981) and solvent accessible surface area (Connolly, 1983). Likewise, the significant improvements reported in the residue conservation models (Taylor, 1986), vector representation of the protein sequences as well as structures (Yamamoto and Yoshikura, 1986) along with the automatic structure sketching and identification of the conserved motifs (Rooman and Wodak, 1988) and structural building blocks (Unger et al., 1989). Furthermore, new algorithms were formulated for the structural comparisons of the proteins (Cohen and Sternberg, 1980) and new filtering steps such as class predictions were introduced in structure prediction (Klein, 1986).

The discipline of molecular modelling was developed during this period with the introduction of problems associated with the fold recognition or threading of protein sequences to the structure (Ponder and Richards, 1987). Extensive studies aimed at deriving architectural attributes of protein structures from geometrical analysis of exclusive folds and families (Brandeen, 1980). These studies included the analyses of the structural features such as disulfide bridges (Thornton, 1981), beta-sheets (Chothia and Janin, 1981), beta-sheet sandwiches (Cohen et al., 1981), beta-barrels (Lasters et al., 1988), coiled-coils (Cohen and Parry, 1986), and loops (Leszczynski and Rose, 1986). This period was also evident the development of the NMR which allowed the availability of the solution structures of the proteins (Wuthrich, 1990).

Current methodologies for the protein structure predictions are based on comparative or homology modelling, Threading or Fold recognition as well as the *ab initio* protocols (Baker and Sali, 2001). The primitive methods of the comparative modelling utilize the comparison of the target protein sequence with that of template sequences with known structures (Dorn et al., 2014, Marti-Renom et al., 2000). On the basis of the identity between the aligned sequences (>25%), the high atomic resolution structures of the target proteins are predicted by using the spatial restraints present in the structure of the templates (Dorn et al., 2014). The homology modelling techniques are generally used for the structure predictions of the proteins having evolutionary relationships with polypeptides of known structure (Dorn et al., 2014).

The amino acid residues with similar physicochemical properties simulate the equivalent positions in the 3-D structures of homologous proteins (Dorn et al., 2014). The protein structures predicted with identity (> 60%) showed the atomic resolutions comparable with the X-ray derived structures and can be utilized for mutagenesis studies, drug design and virtual screening (Dorn et al., 2014). A variety of tools have been developed by using the concepts of the comparative modelling, such as MODELLER (Eswar et al., 2006), Biskit (Grunberg et al., 2007), SWISS-MODEL (Schwede et al., 2003), SAM-T08 (Karplus, 2009), and Phyre2 (Kelley et al., 2015).

Furthermore, methods using the principles of Threading or Fold recognition for the prediction of protein structures were devised. The 3-D structures of proteins showed more conservation as compared to their primary structure during the course of evolution (Dorn et al., 2014). The process of fold recognition involves the comparison of the query sequence with the structure of the template proteins in order to identify the similar folds and then query structure is predicted by mimicking the atomic positions of the template (Dorn et al., 2014). Currently,

several threading based tools are available such as RaptorX (Kallberg et al., 2014), MUSTER (Wu and Zhang, 2008), HHpred (Soding et al., 2005), SPARKS-X (Yang et al., 2011), and pGenTHREADER (Lobley et al., 2009). Moreover, in the absence of templates or reliable fold in the structural databases, the *ab initio* protocols are used which are based on the global optimization of the energetics as well as involve the statistical propensity of conformational attributes related to the native structures (Dorn et al., 2014). The *ab initio* algorithms predict the 3-D structure by utilizing the physical theories without exploring the known structures (Dorn et al., 2014). Such algorithms are formulated in the computational techniques such as Rosetta@home (Das et al., 2007), QUARK (Xu and Zhang, 2012), Bhageerath-H (Jayaram et al., 2014), I-TASSER (Roy et al., 2010), and ROBETTA server (Kim et al., 2004).

2.1.4.2 Molecular dynamics simulations

The concept of molecular dynamics was first introduced in the late 1950s by Alder and Wainwright, in order to analyze the interactions between hard spheres (Alder and Wainwright, 1957). This study formed the basis for the current discipline as significant insights about the behavior of the simple liquid molecules resulted from this analysis. The major advancement took place in the mid-1960s, first MD simulations were performed by Rahman (known as the *father of molecular dynamics*) for the liquid argon (Rahman, 1964). Furthermore, in the later 1970s, the Rahman and Stillinger performed the MD simulation on the first realistic system of the liquid water (Stillinger and Rahman, 1974). The first biomolecule simulations were reported in 1977 in which Three Dimensional (3-D) structure of bovine pancreatic trypsin inhibitor was utilized (McCammon et al., 1977).

Nowadays, several MD simulation packages are available such as Assisted Model Building with Energy Refinement (AMBER) (Case et al., 2012), Chemistry at HARvard Molecular Mechanics (CHARMM), Groningen MAchine for Chemical Simulation (GROMACS), TINKER (http://lms.chem.tamu.edu/tinker.html), ESPResSo (http://espressomd .org/), ADUN (Johnston et al., 2005), IMD (http://imd.itap.physik.uni-stuttgart.de/), BOSS (Jorgensen and Tirado-Rives, 2005), MDynaMix (http://www.fos.su.se/~sasha/mdynamix/), CPMD (Car and Parrinello, 1985), MOLDY (http://cc-ipcp.icp.ac.ru/Moldy_2_16.html), DL_POLY (Smith and Forester, 1996), OPENMD (http://openmd.org/), LAMMPS (Grindon et al., 2004), PINY_MD (https://files.nyu.edu/mt33/public/PINY_MD/PINY.html), MOIL (Ruymgaart et al., 2011), Q (http://xray.bmc.uu.se/~aqwww/q/), NAMD (Phillips et al., 2005), SageMD (http://sagemd.com/), ORAC (Marsili et al., 2010), SIESTA (http://departments.icmab .es/leem/siesta/), VASP (Hafner, 2008), SPARTAN (http://www.wavefun.com/), and URES (Rojas et al., 2007). The AMBER (Case et al., 2012) is amongst the most widely used *ab initio* molecular modelling packages developed in the 1970s, which simulates and analyze the dynamics of biomolecules such as proteins, nucleic acids, and carbohydrates.

The AMBER package consists of two groups of algorithms, the first section contains diverse force field parameters utilized for simulations while the other parts include algorithms aimed at performing the solvation, minimization, equilibration, and production of final trajectories of molecular dynamics (Case et al., 2005). The methodology of the MD simulation is generally divided into three sections (Case et al., 2005) in which the primary phase includes the preparation of the systems by using force field parameters as well as a diverse solvent models. The generated systems of the biomolecules then simulated in order to obtain the global minima and finally, the generated trajectories were analyzed to assess their conformational behaviors

(Case et al., 2005). Similarly, the CHARMM software is another widely used package to perform MD simulations (Brooks et al., 2009). In addition to the several established protocols, some advanced features such as free energy perturbation, quasi-harmonic entropy estimation, and correlation analysis are included in the CHARMM package (Brooks et al., 2009). On the other hand, the GROMACS package (Pronk et al., 2013) is amongst the most cited MD simulation package contains high-throughput as well as highly parallel algorithms frequently utilized for running simulations for relatively larger bio-molecular systems (Pronk et al., 2013).

2.2 Virulence factors of *M. tuberculosis*

The current section highlighted the diversity of the virulence factors that may be present in the group of 1000 HPs obtained from the genome of *M. tuberculosis* and the above-explained methodologies can be used for their detailed analyses. The genomic content of the members present in the genus Mycobacterium evolved during the course of time resulting varied pathogenicity (Forrellad et al., 2013, Smith, 2003, Prozorov et al., 2014). The variation occurred because of the series of DNA deletions/insertions in the 14 conserved regions observed in the different genomes present in the genus Mycobacterium (Forrellad et al., 2013, Smith, 2003, Prozorov et al., 2014).

As a result, a diversity of virulence factors evolved in response to the reaction of the host immune system (Forrellad et al., 2013, Smith, 2003, Prozorov et al., 2014). The current advances in the biological and related sciences lead to the better understanding of the molecular basis of its virulence and tenacity inside the host systems (Forrellad et al., 2013, Smith, 2003, Prozorov et al., 2014). This leads to the identification of the virulent genes in Mycobacterium. This was mainly achieved by the combination of *in vivo* screening methods along with generation of the transposon mutant libraries (Forrellad et al., 2013, Smith, 2003, Prozorov et al., 2014). As a

result of these efforts, immense virulent genes were identified that were crucial for the bacterial survival in the hosts' systems (Forrellad et al., 2013, Smith, 2003, Prozorov et al., 2014). The ranges of virulence determinants present in the *M. tuberculosis* are depicted in Figure 2.3. On the basis of their functionalities, molecular properties as well as their localization in the cells of *M. tuberculosis*, they are grouped in different categories below.

Figure 2.3: The diversity of the virulence factors present in *M. tuberculosis*

2.2.1 Secretory Proteins

The secretion of the proteins enables *M. tuberculosis* to survive and adapt in the natural environment of the human host (Forrellad et al., 2013, Smith, 2003, Prozorov et al., 2014). The secreted proteins along with the metabolic enzymes are the major virulent determinants, as both the categories play a key role in the synthesis of the cell surface associated molecules (Forrellad

et al., 2013, Smith, 2003, Prozorov et al., 2014). The recent studies showed that the presence of specific secretory systems such as type seven and Type VII secretion systems, which enable *M. tuberculosis* to secrete the protein across the cell envelope (Abdallah et al., 2007). Around five of such Type VII secretion systems were reported in *M. tuberculosis* which is collectively known as ESX systems (Abdallah et al., 2007).

The culture studies showed that at least 200 proteins were subjected to secretion in order to interact with the environmental components (Sonnenberg and Belisle, 1997). Among different classes of the secretory proteins, the functionalities of significant proteins such as HspX, ESAT6 family proteins, Rv3763 and Glutamine synthase were discussed. HspX or Acr family proteins are the major antigenic molecules present in *M. tuberculosis*, secreted in the THP-1 macrophages of the human host under the anoxic conditions (Wayne, 1994, Yuan et al., 1998). The HspX is involved in the dormancy and persistence of *M. tuberculosis* as it's overexpression leads to the inhibition of the bacterial growth (Wayne, 1994).

Similarly, ESAT6 family proteins are the immunodominant antigens, which are crucial for the pathogenesis of virulent members of genus Mycobacterium (Skjot et al., 2000). The members of ESAT6 family such as Rv3874 as well as Rv3875 are located in the RD1 deletion region, which is the first deletion region identified after the genomic comparisons (Skjot et al., 2000, Smith, 2003). Furthermore, the Rv3763 is a glyco-lipoprotein and T cells recognized immunodominant antigens found in the majority of the TB patients (Noss et al., 2001). This surface exposed protein trigger the signaling event after interaction with TLR2 (Noss et al., 2001). Moreover, the Rv2220 which is characterized to be a glutamine synthase is involved in the synthesis of cell wall components such as a poly-L-glutamate-glutamine cell in pathogenic strains of mycobacteria (Harth and Horwitz, 1999).

2.2.2 Components of cell surface

The cell envelope of *M. tuberculosis* contains a diversity of biomolecules which are unique to this pathogenic bacteria (Smith, 2003). The Erp protein which is located on the surface of *M. tuberculosis* containing six tandem repeats may be involved in the generation of resistance against macrophages (Berthet et al., 1995). Similarly, mycocerosic acid synthase (Mas) is involved in the synthesis of cellular components by catalyzing the formation of methylated long fatty acid chains (Azad et al., 1997). The FadD26 and FadD28 are the proteins which were characterized as acyl coenzyme A synthetase and fatty acid CoA synthase respectively and may be involved in the degradation of surface exposed fatty acid (Camacho et al., 1999). These proteins are significant for the process of *M. tuberculosis* virulence (Camacho et al., 1999). Furthermore, the Fbp group of proteins presents in Mycobacteria annotated as mycolyl-transferase enzymes and catalyzes the transfer of mycolic acids long chains to the derivative molecules of trehalose (Belisle et al., 1997). These proteins are the key components involved in the binding with fibronectin protein present in the cell matrix (Belisle et al., 1997).

The proteins expressed from the genes *fbpA, fbpB,* and *fbpC* are highly antigenic in nature and depicted as antigen 85A-C (Belisle et al., 1997). Moreover, MmaA4 is a member of the group containing four closely related methyltransferases, catalyzed the formation of keto and methoxy derivatives of meromycolic acid (Dubnau et al., 1997). The formation of the derivative took place by the intimating the methylation of a double bond in meromycolic acids (Dubnau et al., 1997). These modifications may result in the variability of the cell wall permeability and make it more resistant to the oxidative stresses (Dubnau et al., 1997). The other transferase activities have been reported which are catalyzed by the protein PcaA and facilitate the alteration

of mycolic acids through the formation cyclopropane ring in its α-mycolate chains (Glickman et al., 2000).

In addition, the porin proteins such as ompA played a significant role in the stress conditions as it forms the pores in the liposomes and leads to the survival of the *M. tuberculosis* in the moderately high acidic conditions (Senaratne et al., 1998). Likewise, HbhA protein which showed the presence of heparin-binding hemagglutinin activities present on the surface of virulent strains of mycobacteria and are molecules involved in the bacterial interaction with the pneumocytes, which is significant for extra-pulmonary distribution (Pethe et al., 2001). Besides these cell adherent virulence factors, the Lipoarabinomannan (LAM) which is an immuno-modulating glycolipid causes the blocking of macrophage activation, inhibition of human protein kinase C and is a scavenger of oxygen radicals (Hunter et al., 1986).

2.2.3 Enzymes of Cellular Metabolism

The mutational analyses in the genes encoding for the specific enzymes in the metabolic pathways and acquisition systems resulted in the virulence deficit *M. tuberculosis* (Smith, 2003). Particularly, the enzymes involved in the lipid and fatty acid metabolisms such as phospholipases, isocitrate lyase, and LipF showed highest deleterious effects (Smith, 2003). The isocitrate lyase is the key component of the glyoxylate pathway, in which it catalyzes the conversion of isocitrate to succinate (Wayne and Lin, 1982). The glyoxylate shunt allows bacteria to use on the other biomolecules instead carbohydrate as the primary carbon source (Wayne and Lin, 1982). The upregulation in the activities isocitrate lyase was reported after *M. tuberculosis* infection inside the macrophages (Wayne and Lin, 1982, Camacho et al., 1999). The lipases or esterases such as LipF which may be involved in the lipid degradation may be involved in the pathogenesis of *M. tuberculosis* (Camacho et al., 1999).

Furthermore, in *M. tuberculosis* around 36 orthologs of *Escherichia coli* FadD proteins were identified (Rindi et al., 2002). The FadD protein is an acyl-CoA synthase, which is crucial for the primary step in beta-oxidation of the fatty acids and catalyzes the addition of CoA moiety to the fatty acids chains (Rindi et al., 2002). Likewise, pantothenate synthetase, as well as aspartate-1-decarboxylase, form the part of pantothenate biosynthesis in the virulent *M. tuberculosis* (Sambandamurthy et al., 2002). The pantothenate is essential intermediate in the synthesis of a variety of biomolecule involved in the fatty acid metabolism (Sambandamurthy et al., 2002). Moreover, four proteins namely, Rv2351c (*plc*A), Rv2350c (*plcB*), Rv2349c (*plcC*), and Rv1755c (*plcD*) were identified as Phospholipases C in *M. tuberculosis* (Raynaud et al., 2002). The latter three genes are closely related as compared to *plcD* and are significant for the pathogenesis of virulent Mycobacteria (Raynaud et al., 2002).

In addition to this, the enzymes such as isopropylmalate isomerase, anthranilate phosphoribosyl transferase, pyrroline-5-carboxlate reductase, and 1-phosphoribosylamino-imidazole-succinocarboxamide synthase form the part of purine and amino acid metabolic pathways, which are crucial for the pathogenesis of *M. tuberculosis* (Smith, 2003). The LeuD or isopropylmalate isomerase is essential for the leucine biosynthesis in *M. tuberculosis* (Hondalus et al., 2000). The mutation studies validated that the respective enzyme is essential for the bacteria growth in macrophages (Hondalus et al., 2000). Similarly, TrpD and ProC proteins which are anthranilate phosphoribosyl transferase and pyrroline-5-carboxlate reductase respectively and are involved in the biosynthesis of amino acids such as tryptophan and proline (Parish et al., 1999, Smith et al., 2001). These enzymes are also crucial for the survival and growth in the macrophages (Parish et al., 1999, Smith et al., 2001). Likewise, the mutation in 1-

phosphoribosylaminoimidazole-succinocarboxamide synthase or PurC resulted in the attenuation of growth in murine macrophages (Jackson et al., 1999).

2.2.4 Metal transporters

The uptake of iron and magnesium is essential for the survival and the virulence of pathogenic bacteria (Smith, 2003). The MgtC is a horizontally acquired virulence factor involved in the recovery of Mg^{2+} from the limiting environment of the macrophages (Blanc-Potard and Lafay, 2003). It is crucial for the growth and survival of the *M. tuberculosis* in the stressed conditions (Blanc-Potard and Lafay, 2003). The roles of MgtC along with other stress-related biomolecules are explained in Chapter 1. In prokaryotes, the absorption of the iron involved its solubilization and chelation by using siderophores molecules such as lactoferrin and transferrins, then its intake is carried out by using high-affinity transporters (Boradia et al., 2014). In *M. tuberculosis*, the MbtB catalyzes the formation of the mycobactin as well as carboxymycobactin, the siderophores present for the consumption of iron (Dubey et al., 2002). Likewise, the DNA binding protein such as IdeR requires binding of Fe^{2+} ions for the proper functionality (Pohl et al., 1999). This is a major regulator of iron assimilation in *M. tuberculosis* perform the repression of the iron uptake and activation of the storage genes (Pohl et al., 1999).

2.2.5 Regulatory proteins

The transcriptional regulators which control the transcription of a variety of genes are essential for the pathogenesis of *M. tuberculosis* (Smith, 2003). In order to adapt the changing environment, the prokaryotes use different strategies to drastically modify their lifestyle such formation of different RNA polymerase holoenzymes by using diverse sigma factors with varied specificities (Smith, 2003). The sigma A is the principal sigma factor in *M. tuberculosis* and necessary for the transcription of the house-keeping genes (Gomez et al., 1998). The sigma A is

considered as an essential virulence factor in *M. tuberculosis* because it interacts with the transcriptional activator which resulted in the expression of the other virulence factors (Gomez et al., 1998). Similarly, the sigma F regulates the process of bacterial sporulation and the bacterial responses to the environmental stresses (DeMaio et al., 1996). The sigma E is a member of the extracytoplasmic sigma factors that regulates the *M. tuberculosis* responses against the environmental stresses such as detergent stress along with high temperatures in macrophages (Manganelli et al., 1999). The sigma E also regulates the expression of the protein involved in the responses against the oxidative stresses, electron transport and mycolic acid biosynthesis (Manganelli et al., 1999).

Moreover, the sigma H which is also an ECF family protein which helps bacteria to withstand the oxidative stresses by regulating the expression of the structural gene such as *sigR* as well as *trx* which encodes the thioredoxin as well as the thioredoxin reductase enzyme respectively (Paget et al., 1998). The latter proteins then perform the reduction of the stress oxidized proteins (Paget et al., 1998). Additionally, the sigma H also participates in the responses against the stress of the heat shock and during the infection inside the macrophages (Paget et al., 1998).

Furthermore, another strategy utilized by *M. tuberculosis* against the changing environmental conditions involves the usage of "two-component systems", which are the signal transduction pathways which contain sensory histidine kinases which activate the associated effector proteins and response regulatory proteins (Smith, 2003). The PhoP form the part of the two-component systems that generates the response against the reduced Mg^{2+}concentrations and regulates the expression of the virulent genes (Groisman, 2001). The other two component systems present in the *M. tuberculosis* include the proteins PrrA (Ewann et al., 2002) and

Rv0981, which shows the up-regulation during the infection inside the macrophages and are important for the virulence of the bacteria (Zahrt and Deretic, 2001).

In this study, new virulence factors were classified in the genome of *M. tuberculosis*. The different methodologies adopted along with outcomes obtained regarding the functional annotation and conformational behaviors of the virulent proteins are discussed in the successive chapters.

CHAPTER 3

COMPUTATIONAL

This chapter deals with the theoretical principles of Density Functional theories (DFT) and Molecular Dynamics (MD) simulations followed by mutation-based drug resistance analysis, acid resistance analyses, sequence-based function predictions and structural analyses of virulent proteins of *M. tuberculosis*. The purpose of this study was to predict the functions and virulence characteristics of the uncharacterised proteins including the establishment of *in silico* protocols for sequence analysis. The structural basis of drug-resistant mutations were analyzed for 50 ns MD simulations, while a constant pH range from 3-6 was used to check the acid resistance for the proteins of *M. tuberculosis*. The MD simulations performed in this study provided valuable insights into the molecular mechanisms, stability profiles, and drug resistances associated with the virulent proteins. On the basis of sequence analysis, 662 of the 1000 hypothetical proteins (HPs) studied were annotated and the major functional categories are discussed in subsequent sections. Moreover, this chapter also includes the classification of novel virulence factors, in which 28 HPs were classified as virulent and the structures of six HPs with the highest predicted scores were analyzed using the principles of molecular modelling techniques.

3.1 Theoretical principles

The techniques based on Density Functional theories (DFT) and Molecular dynamics (MD) simulations enable the understanding of protein folding as well as their characteristic enzymatic catalysis mechanisms (Karplus and Kuriyan, 2005, Liao et al., 2010). In this study, these methods were used for the analyses of the structural basis of the proteins involved in the resistance mechanisms of *M. tuberculosis* as well as for deciphering the conformational behavior of the predicted virulent proteins. The proteins involved in mutation based drug resistance were studies at a time scale of 50 ns (= 600 ns) as well as virulent proteins of acid resistance mechanism present in *M. tuberculosis* were analyzed using constant pH MD simulations each at 50 ns time scale (= 1400 ns). Moreover, the structures of predicted virulent HPs were simulated for 50 ns in explicit water conditions in order to understand the dynamic of these proteins in

explicit solvent conditions. The conceptual understanding of the used DFT-based methods and MD simulations are explained in the following section:

3.1.1 Density Functional Theory

DFT is a quantum mechanical-based technique used for the investigation of the ground state electronic structures such as atoms, molecules, as well as the condensed phases in chemistry and physics. The biomolecules comprise of numerous atoms and subsequently require highly accurate DFT methods in order to perform the quantum mechanical calculations (Ban et al., 2002). In such multi-body electronic structures, the calculations were based on Born–Oppenheimer approximations in which the nuclei of the studied molecules are fixed. The energy of the system is computed on the basis of the solution of time independent Schrödinger equation:

$$\left\{\frac{-1}{2}\sum_{i}^{N}\nabla_i^2 + \hat{V}_{ext} + \sum_{i<j}^{N}\frac{1}{|r_i - r_j|}\right\}\Psi(r_1, r_2, \ldots., r_N) = E\Psi(r_1, r_2, \ldots., r_N)$$

Where Ψ is the wavefunction, E is the total energy of the system, V is the potential energy of the system and N is the number of the electron systems. The Schrödinger equation can be solved more sophisticatedly by the introduction of the term which can express the electron-electron repulsion while calculating the energy of the multi-atom system. Such calculations are presented by the Hartree-Fock methods which expresses Ψ as the anti-symmetric product of functions (ϕ_i) depending on the coordinates of the single atom and represented by the following equation:

$$\Psi_{HF} = \frac{1}{\sqrt{N!}}\det[\phi_1, \phi_2, \ldots \phi_N]$$

On the basis of Hartree-Fock theory method, the energy is calculated using the following equation:

$$\mathrm{E}_{HF} = \int \phi_i^* (\mathrm{r}) \left(\frac{-1}{2} \sum_i^N \nabla_i^2 + \hat{V}_{ext} \right) \phi_i (\mathrm{r}) d\mathrm{r} + \frac{1}{2} \sum_{i,j}^N \int \frac{\phi_i^* (\mathrm{r}_1) \phi_i (\mathrm{r}_1) \phi_j^* (\mathrm{r}_2) \phi_j (\mathrm{r}_2)}{|r_i - r_j|} d\mathrm{r}_1 d\mathrm{r}_2$$

$$- \frac{1}{2} \sum_{i,j}^N \int \frac{\phi_i^* (\mathrm{r}_1) \phi_j (\mathrm{r}_1) \phi_i (\mathrm{r}_2) \phi_j^* (\mathrm{r}_2)}{|r_i - r_j|} d\mathrm{r}_1 d\mathrm{r}_2$$

In which, the second term represents the classical Coulomb energy represents in the form of orbitals while the third term symbolizes the exchange energy.

Furthermore, in order to reduce the non-convergence of the quantum calculations resulting from the diminishing HOMO-LUMO gaps in the studied systems, the alternate DFT formulation present in the Kohn-Sham method was used (Elias, 2012). This method calculates the energy in a similar way as the Hartree-Fock but the exchange energy is calculated with higher accuracy by the following equation:

$$\left[\frac{-1}{2} \sum_i^N \nabla_i^2 + \hat{V}_{ext} (r) + \int \frac{\rho(r')}{|r - r'|} d\mathrm{r}' + \nu_{xc} (\mathrm{r}) \right] \phi_i (r) = \varepsilon_i \phi_i (r)$$

This equation illustrated the behavior of non-interacting electrons using exchange-correlation potential (Elias, 2012). These methods provided a platform for the analyses of diverse chemical properties and associated reaction mechanisms (Ban et al., 2002). Furthermore, the benchmarking of 148 molecules showed the relatively higher accuracy of the B3LYP method among the available techniques (Blomberg and Siegbahn, 2001). Therefore, in this study, the DFT methods implemented in Gaussian 9.0 (Frisch et al., 2009) were used for optimizing the geometry of the biomolecules and for analyzing the effect of the point mutations on the stability of the proteins. The DFT-based optimized structures were used as input structures in the MD simulations, as described in the subsequent sections:

3.1.2 Molecular Dynamics Simulations

Molecular dynamics (MD) simulations are one of the most important tools for the computational study of biomolecules (Lindahl, 2008). It computes the time-dependent behavior of a molecular system and therefore, provides comprehensive information about changes in protein fluctuations and conformation. These are also used in structures determination from NMR experiments and X-ray crystallography. Molecular dynamics (MD) is described as a type of computer simulation in which the atoms and the molecules are allowed to interact for a time period by approximations of known physical attributes, resulting in the simulation of the motion (Lindahl, 2008) of a system of particles (Lindahl, 2008). The analytical study of properties of such complex systems is a challenging task and therefore, an MD simulation provides a solution by means of numerical methods. It represents an interface between laboratory experiments and the theory and can be understood as an effective experiment thereby probing the link between molecular structure, function, and its movement. The laws and the theory of MD simulations are derived from interdisciplinary science like physics, chemistry, and mathematics. It utilizes algorithms from the computer science and information theory. It is applied today generally in materials science and the study of complex dynamic processes that occur in biological systems, including protein folding, protein stability, molecular recognition, conformational changes, drug design, determination of 3D structures, enzyme reactions and ion transport in biological systems. (Lindahl, 2008, Karplus and McCammon, 2002). The algorithms of MD simulations aimed at the solution the principle of Newton's Laws of motion for the system of N interacting atoms, which identifies the fluctuations in the atomic positions as well as particle velocities regarding time. This can be achieved by deriving the forces (F_i) acting on each particle with respect to time:

$$F_i = m_i \frac{\partial^2 r_i}{\partial t^2}, i = 1 \dots N$$

Whereas, the force is the negative gradient of potential energy function:

$$F_i = -\frac{\partial V}{\partial r_i}$$

"V" represents the potential energy of the particle while "r" symbolizes the position of a particle. During the MD simulations, the respective equations are solved in small time steps, simultaneously. At constant temperature and pressure values, the coordinates of the system will be written in the output files on the regular basis. These coordinates with respect to time represent the trajectory of the system. After some time the system reaches an equilibration phase and by generalizing some parameters (given in Table 3.1) from the output files, many macroscopic properties can be calculated. Similarly, the acceleration (a_i) achieved by each particle during the course of MD simulations of can be determined in terms of the forces acting on the particle and their associated masses:

$$a_i = \frac{F_i}{m_i}$$

Moreover, the total energy of the system can be represented in the form of kinetic energy as well as potential energies, while kinetic energy in terms of the velocity of the particles can be represented as:

$$K(v) = \frac{1}{2}\sum_{i=1}^{N} m_i \, {v_i}^2$$

The molecular mechanics methods such as Coulomb and Lennard-Jones PME implemented in the MD simulations were used to describe the motion of the atoms as well as for the calculation of the charge distribution.

Table 3.1: Typical vibration frequencies in molecules (wave numbers) and hydrogen-bonded liquids. (Source: GROMACS manual 5.0.7]

S. No.	Type of bonds	Type of vibrations	Wavenumber (cm^{-1})
1.	C-H, O-H, N-H	Stretch	3000 - 3500
2.	C=C. C=O	Stretch	1700 – 2000
3.	HOH	Bending	1600
4.	C-C	Stretch	1400 – 1600
5.	H_2CX	Sciss, Rock	1000 – 1500
6.	CCC	Bending	800 – 1000
7.	O-H…O	Liberation	400 – 700
8.	O-H…O	Stretch	50 – 200

Similarly, the kinetic energy can be illustrated in the form of momentum using the following equation:

$$K(p) = \frac{1}{2}\sum_{i=1}^{N}\frac{p_i^{\,2}}{m_i}$$

The Hamiltonian operator or the total energy of the system can be represented in terms of kinetic and potential energies:

$$H(q, p) = K(p) + U(q)$$

Where U(q) represents the potential energy function of the system, q is the set of Cartesian coordinates and p is the momenta of the particles.

The understanding of the functionality associated with biological macromolecules requires the knowledge of their structure as well as dynamics (Karplus and Kuriyan, 2005). Consequently, the MD simulations provide numerous methods for analyzing the conformational energy landscapes accessible to biomolecules (Karplus and Kuriyan, 2005). These methods enable the calculation of the time-dependent behavior of bio- molecular systems in the form of variations in the fluctuations of their constituents and conformational dynamics. It represents an interface between experimental as well as theoretical studies and can be effective in the

searching of a link between the molecular structure of the biomolecules, their function and dynamics (Karplus and Kuriyan, 2005). The MD simulation techniques coupled with low-resolution experimental methods can provide an alternative for conventional structure determination techniques such as NMR experiments, X-ray crystallography and others (Lindahl, 2008).

During the course of MD simulations, the constituent atoms, as well as molecules, are allowed to interact with each other in order to visualize the dynamical evolution of the system (Lindahl, 2008). As a result of its diverse application, the MD simulations can be used in designing new materials and is explored for the better understanding of the dynamics of the complex biological mechanisms including protein folding and stability, conformational changes, molecular recognition, drug design, enzyme reactions, transport mechanisms of ion in biological systems (Lindahl, 2008, Karplus and McCammon, 2002). Different methods used for obtaining the thermodynamics properties from MD simulations are discussed in the following sections:

3.1.2.1 Periodic Boundary Conditions (PBC)

In classical mechanics, the edge effect of the finite system can be minimized using the *Periodic Boundary Conditions*. During the MD simulations of biomolecules, they have put inside a space-filling box which is surrounded by its replicas in every direction in order to form a lattice. Therefore, the relatively precise approach in simulating the systems involves the usage of explicit solvent conditions, in which the system is prepared by soaking the concerned molecule in a box of solvent. The PBC conditions are frequently utilized in studying the molecules in bulk solvent conditions. Virtually in most of the MD simulations, in order to increase the computational efficiency in the determination of the potentials of the systems several cutoff

schemes are used. These cutoff schemes are based on the minimum image convention in which it was assumed that each particle interacts with the neighboring images of the n-1 particles. These cutoff techniques have been shown to simulate the artificial behavior of the system and can produce major errors in the simulations (Holden et al., 2013).

3.1.2.2 Ewald Summation Techniques

Ewald summation (ES) are used in the MD simulation studies for the calculations of the long range coulombic interactions in the periodic boundary conditions, which is a time-consuming and computationally expensive method. ES based techniques were first introduced in 1921 (Fennell and Gezelter, 2006) in order to calculate the long range interactions amongst the infinite particles systems and their numerous periodic images. Long-range interactions are calculated as the sum of a short-range (estimated in real space) as well as the long-range contributions (obtained using Fourier transformation), which do not show singularity. The principle of calculating ES involves the conversion of the outputs generated on the basis of two series potential energy as illustrated in Figure 3.1 (Fennell and Gezelter, 2006).

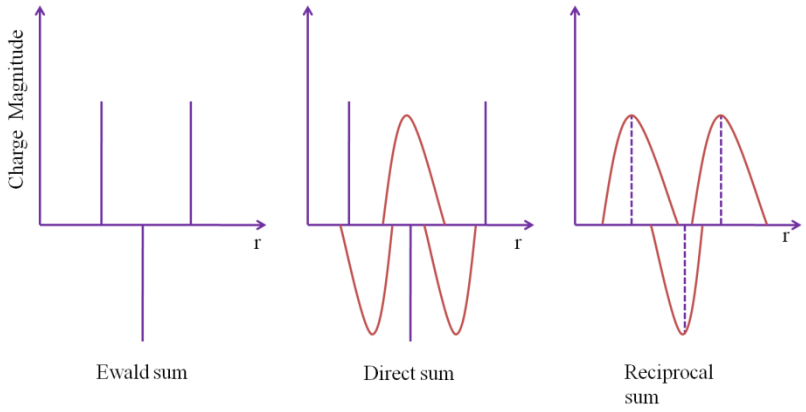


Figure 3.1: The division of the charges into discrete and smeared distributions in the real and reciprocal space.

During the calculation of the ES, the Gaussian charge distribution is frequently used. The calculation of sum regarding the point charges is modified to a summation of the interactions

among the charges and the neutralization of the distributions was done corresponding to the equation below:

$$U_{Ewald} = U^r + U^m + U^0$$

In which, the $\mathbf{U^r}$ depicts the summation of the real space whereas $\mathbf{U^m}$ is the reciprocal sum calculated on the basis of Fourier transformation, while $\mathbf{U^0}$ is a constant term (Fennell and Gezelter, 2006).

3.1.2.3 Particle Mesh Ewald (PME)

The Particle Mesh Ewald method (PME) divides the potential energy into Ewald's standard direct and reciprocal sums. Instead of directly summing wave vectors, the charges are assigned to a grid using interpolation (Norberto de Souza and Ornstein, 1999). It involves the usage of standard Gaussian charge distributions (Norberto de Souza and Ornstein, 1999). The direct summation is performed explicitly using cutoffs while the reciprocal sum is calculated through Fast Fourier Transform (FFT) with projections on a grid space followed by the interpolation of charges (Figure 3.2).

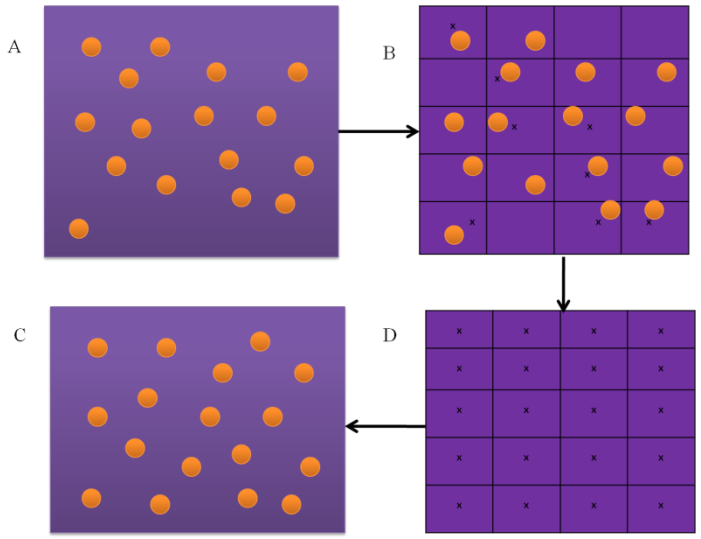


Figure 3.2: A schematic of PME technique involving **(A)** system of charged particles. **(B)** The charges are interpolated on a 2D grid. **(C)** Using FFT, the potential and forces are calculated at grid points. **(D)** Interpolate forces back to particles and update coordinates.

In addition, it calculates the forces on the basis of the analytical distribution of the energies and consequently reducing the requirements of memory significantly (Norberto de Souza and Ornstein, 1999). The MD simulations described above were used for analyzing the conformational changes of the virulent proteins of the *M. tuberculosis*, as described in the subsequent sections.

3.2 Mutation-based drug resistance analyses

The evolution of the molecular mechanisms responsible for the development of drug-resistant bacterial strains is still unclear. Several postulates have been provided that can explain the mechanisms leading to the development of drug resistance which is focused on the mutations in the bacterial proteins (Cohen et al., 2015, Eldholm et al., 2015). Therefore, *in silico* approaches were used to analyze the point mutations leading to the development of the multi-drug resistance in *M. tuberculosis*. The site-directed mutations leading to the development of resistance against four first-line drugs (Ethambutol, Isoniazid, Rifampicin and Streptomycin) were extensively analyzed in this study.

3.2.1 Material and Methods

The analyses of drug resistance associated mutations in proteins of *M. tuberculosis* were carried out in different phases. The primary analyses involved the identification and classification of the mutational landscapes associated with the disease conditions. In the following steps, the stability profiles of diverse point mutations were generated. The mutations with highest deleterious effects were selected for further studies. The extent of the changes resulted from these point mutations were analyzed by the principles of the DFT-based methods, molecular docking, and MD simulation studies. The techniques used for this study are illustrated in the consecutive sections:

3.2.1.1 Generation of mutational landscapes

The initial analyses involved the use of SIFT server, which classified the “tolerant” mutations from the positions at which the point mutation is “intolerant”. The SIFT server utilizes the sequence similarities to analyze whether an amino acid replacement will influence the protein functionalities which may result in the phenotype modifications (Ng and Henikoff, 2003).

Among the SIFT predicted intolerant mutations, the pathogenic substitutions were further classified using the neural network algorithms implemented in PMUT server. Neural networks are a computational approach modelled on the human brain and nervous system for solving problems through classification of data. Moreover, the effects of the classified pathological mutations on the stability of proteins were studied using the sequence-based as well as structure-based *in silico* methods (Ferrer-Costa et al., 2005).

The sequence-based analyses of the effects of point mutations were studied using I-Mutant, MuStab, and EASE-MM servers. The I-Mutant utilizes the knowledge generated on the basis of SVM learning algorithms to predict the sequence as well as structure based stability changes upon point mutations (Capriotti et al., 2005). The training and testing datasets for the SVM algorithms were generated using ProTherm, which is the most curated database containing the information about the mutation-driven protein stability changes (Capriotti et al., 2005). Furthermore, the outcomes of I-mutant were validated using another machine learning based method incorporated into the MuStab server. The latter server integrates 20 protein sequence features for the stability prediction upon mutation with an accuracy of 84.59% (Teng et al., 2010).

Additionally, EASE-MM server predicted the protein stability after point mutations by utilizing the five specific SVM models, while the final predictions were performed on the basis of consensus between the outcomes of models. These latter methods predicted the effect of amino acid mutations on the accessible surface area as well as secondary structures of the proteins (Folkman et al., 2016). Moreover, the effects of residue substitutions on the structure of proteins were evaluated using the SDM server, which classified the disease-associated mutations by on the basis of their effect on the stability profile of the proteins (Worth et al., 2011).

3.2.1.2 DFT based analyses

As proteins are very large biomolecules, the feasibility of the quantum calculations become limited. Therefore, in order to overcome these constraints, each protein is divided into numerous fragments and the portions containing the point mutations selected as inputs for the calculations. A total of 16 peptides each of 10 residues long were generated and subjected to the DFT based techniques present in Gaussian 9.0 (Frisch et al., 2009). In primary steps, the geometries of all the fragments were optimized using B3LYP functional theory. In Gaussian 9.0 a variety of the basis sets were implemented in this method, while on the basis of the computational efficiency the 6-31G (d,p) was selected for this study (Frisch et al., 2009). The stability energies coupled with harmonic vibrational frequencies as well as zero point vibrational energies were calculated on the same level of theory. The energy profiles generated for different peptides were compared and their structural stability was evaluated.

3.2.1.3 Molecular docking

The effect of the point mutations on binding of the drugs with the proteins was analyzed on the basis of molecular docking. The structure of the *M. tuberculosis* proteins such as Arabinosyltransferases, katG catalase - peroxidase enzyme, β-subunit of RNA polymerase and ribosomal protein rpsL which are observed to be involved in the drug resistance, were retrieved from protein data bank (PDB). The structure of all the proteins and drug molecules were optimized structure using CHARMM (Vanommeslaeghe et al., 2010) force field based refinement modules of Discovery studio 2016 (DS). The optimized structures of proteins and drug molecules were docked using rigid docking module of AutoDock 4 (Morris et al., 2009) package. The AutoDock 4 perform the prediction of the bound conformation on the basis of the free energy based empirical force field coupled with Lamarckian Genetic Algorithm (Morris et

al., 2009). The grid box of dimensions 40 x 40 x 40 Å along the XYZ directions with a grid spacing of 0.200 Å was established using the AutoGrid module and grid was centered at the site of point mutations. In order to increase the efficiency, the parameters associated with Lamarckian genetic algorithm were set to the maximum efficiency values such as a number of individuals in the population were set to 250 while the maximum number of energy evaluations was set to "longer". As a result around 100 docked conformations were obtained which were grouped according to the RMSD tolerance of 2.0 Å. The generated conformations were re-scored on the basis of the scoring function present in DrugScoreX server (Neudert and Klebe, 2011) and the conformation with the highest score was selected for the Molecular Dynamics (MD) simulations. The docking was validated by suing the CDOCKER module present which is a CHARMM (Vanommeslaeghe et al., 2010) force field based docking algorithm implemented in DS.

3.2.1.4 MD simulations of mutants

The GROMACS software package (http://www.gromacs.org/) was used for performing the MD simulations on the studied biomolecules (Van Der Spoel et al., 2005). It was developed in 1991 by the Department of Biophysical Chemistry at the University of Groningen in Netherlands, but after 2001 further development was carried out at the Uppsala University in Sweden as well as Royal Institute of Technology (Van Der Spoel et al., 2005). The GROMACS utilities and parameters used are explained the given sections:

3.2.1.4.1 Thermostats in GROMACS

In GROMACS version 5.1.2 (http://manual.gromacs.org/documentation/5.1.2/), a variety of thermostats or ensemble methodologies is available regarding the addition as well as the elimination of energy from the boundaries of a biological system on the basis of moderately

realistic approaches such as canonical or NVT ensemble (Fuzo and Degreve, 2012). In the ensemble, the number of constituent particles (N), as well as temperature (T) and the volume (V) of the system remain constant followed by the exchange of energies among endothermic along with exothermic processes (Fuzo and Degreve, 2012). In NVT ensemble, the temperature is described in terms of average kinetic energies of the constituent particles using the equation:

$$\langle \frac{1}{2} m v^2 \rangle = \frac{k_B T}{2}$$

Therefore, diverse thermostats, such as Berendsen and Langevin, Nosé-Hoover and other thermostats have been suggested to handle the particle movements during the MD simulations. The Berendsen thermostat which is used for controlling the temperature during MD simulations involves the correction of the temperature deviations (T) from the set point (T_0) by multiplication of the velocities by a factor τ (Fuzo and Degreve, 2012). Similarly, Nosé-Hoover thermostat, which is an integral thermostat involves the usage of extra degrees of freedom or momentum into the Hamiltonian of the concerned system (Fuzo and Degreve, 2012). In Langevin thermostats, Newton's equation of motion is replaced by that of Langevin equation (Fuzo and Degreve, 2012), in which velocities of the particles are proportional to the frictional forces present in addition to the conservative. In this ensemble the kinetic energy of the particles are regulated on the basis of the following equation:

$$ma = -\xi v + f(r) + f'$$

Where *m* represents is mass, *a* is acceleration, *f(r)* is a conservative force, v is the velocity, ξ is a frictional constant, and f' is a random force. The frictional force - ξv decreases with temperature as ξ is a fixed positive value.

3.2.1.4.2 Solvent models

The MD simulations of proteins with aimed at the understanding of their conformational behavior were carried out under explicit solvent conditions, where water molecules are signified by different models based on the all-atom force field. Due to the universality of water molecules, as it is found in all levels of life, it is the most explored solvent in MD simulation studies (Nguyen et al., 2014). The current version of the GROMACS contains diverse water models in its utilities such as TIP5P, SPC/E, TIP3P and others (Jorgensen and Tirado-Rives, 2005). These models were optimized on the basis of physical properties of water such as the diffusivity, radial distribution function, density anomaly, and critical parameters. In this study, SPC/E 216 water model was used for solvating the mutant proteins.

3.2.1.4.3 Energy-Minimization

On the basis of the order of derivative used by the protocol for locating the minima on the potential energy surface, the energy minimization algorithms are categorized in a variety of classes. The zero order protocols utilize energy function coupled with the grid searching algorithm in order to detect the low energy regions. Furthermore, the first-order derivatives are present in the widely used approaches such as the steepest descent method as well as the conjugate gradient algorithms, which were based on the gradient of the energy function. Moreover, the second order derivatives are included in the methods such as Newton-Raphson algorithm which utilizes the Hessian function to identify the global minima (Kini and Evans, 1991). In this study, the steepest descent and conjugate gradient algorithms present in the GROMACS 5.1.2 were used to minimize the structures of *M. tuberculosis* proteins.

The *steepest descent methods* are frequently utilized in the filtering of bad contacts and geometries in the protein structures (Kini and Evans, 1991). It shifts precisely along the steepest

slope of the potential energy surface, resulted in the partial modification in the molecular structure, and is considered as a most efficient method when the structural parameters of the molecular system extremely deviate from a minimum. Generally, the large step size was chosen for the steepest descent algorithm as the calculation regarding the minimization of the system does not readily converge and can fluctuate. Primarily, this algorithm computes the gradient at its present location and then progresses in the reverse direction of the localized gradient until a minimum is not achieved. The energy is calculated for the original geometry, followed by the movements of one of the atoms in any directions of the coordinate system. This procedure is replicated for all the constituent atoms until a new position with a lower value on the energy surface is achieved. Therefore, numerous smaller steps first proceed on the constituent atoms until the predetermined threshold condition for the whole system is fulfilled. The optimizations of the HPs were slower near the minimum and generally used for a first rough and introductory run regarding the minimization of the system. In order to achieve the minimal conformation, the conjugate gradient algorithm was further utilized for the process of minimization (Kini and Evans, 1991).

The conjugate gradient algorithm is another first-order protocol aimed at the minimization of the concerned bio-systems during MD simulations. This approach diverges from the classical steepest descent protocol, as it simultaneously performs the identification of the gradient and the back search in order to achieve the global minima (Kini and Evans, 1991). The significance of the conjugate gradient-based minimization is that it utilizes the in-house knowledge in order to compute the search direction, and achieve minima faster as compared to the other algorithms such as steepest descent. This is related to the fact that it is the first derivative of the rate of change of total energy concerning atomic arrangements with units of the

gradient are kcal mol^{-1} $Å^{-1}$. The conjugate gradient technique generates a set of directions which exists over the oscillatory effects of the steepest descents in constricted manner. Consecutive directions are not arranged in a perpendicular manner (Kini and Evans, 1991)**,** and the conjugate gradient algorithm transfers the information about the resulted functions from initial iteration to the successive one. In each step of minimization, there is the computation of the gradient and on the basis of supplementary information, the calculation regarding the vectors for further direction was performed until the global minima are achieved. Despite the computational expensive nature of the conjugate gradient method, they are generally used for minimizing the larger systems in contrast to steepest descent algorithm. The computational expensiveness is compensated by the presence of efficient convergence criteria to the system minimum (Kini and Evans, 1991).

3.2.1.4.4 Production of MD simulations

By utilizing the above methods the MD simulations were carried out using the GROMACS 5.1.2 molecular mechanics package (Van Der Spoel et al., 2005). The adopted computational workflow is illustrated in Figure 3.3. The parameterizations of proteins were performed using CHARMM 27 force field (Sapay and Tieleman, 2011). After the generation of the required inputs, the periodic conditions were generated using the "*gmx editconf*" module and the system was solvated using the SPC/E water model present in the GROMACS.

The solvated systems were neutralized by the "*gmx genion*" command, which adds the counter ions to the system using the verlet cut off scheme. Then the neutralized system undergoes a series of minimizations so as to remove the inappropriate geometry as well as structural distortions. The steepest descent algorithms were used in order to minimize the system with the maximum number of steps set to 50000 and the algorithm iteratively executed until a maximum force (<1000 kJ/mol/nm) was obtained. After the minimization of the system, it is

equilibrated using the ensemble conditions available in the GROMACS. The equilibration of the system is achieved by simulating in the NVT (constant number of particles, volume, and temperature) as well as NPT (constant number of particles, pressure, and temperature) ensembles. The final simulations were produced using the LINCS algorithm implemented in the GROMACS.

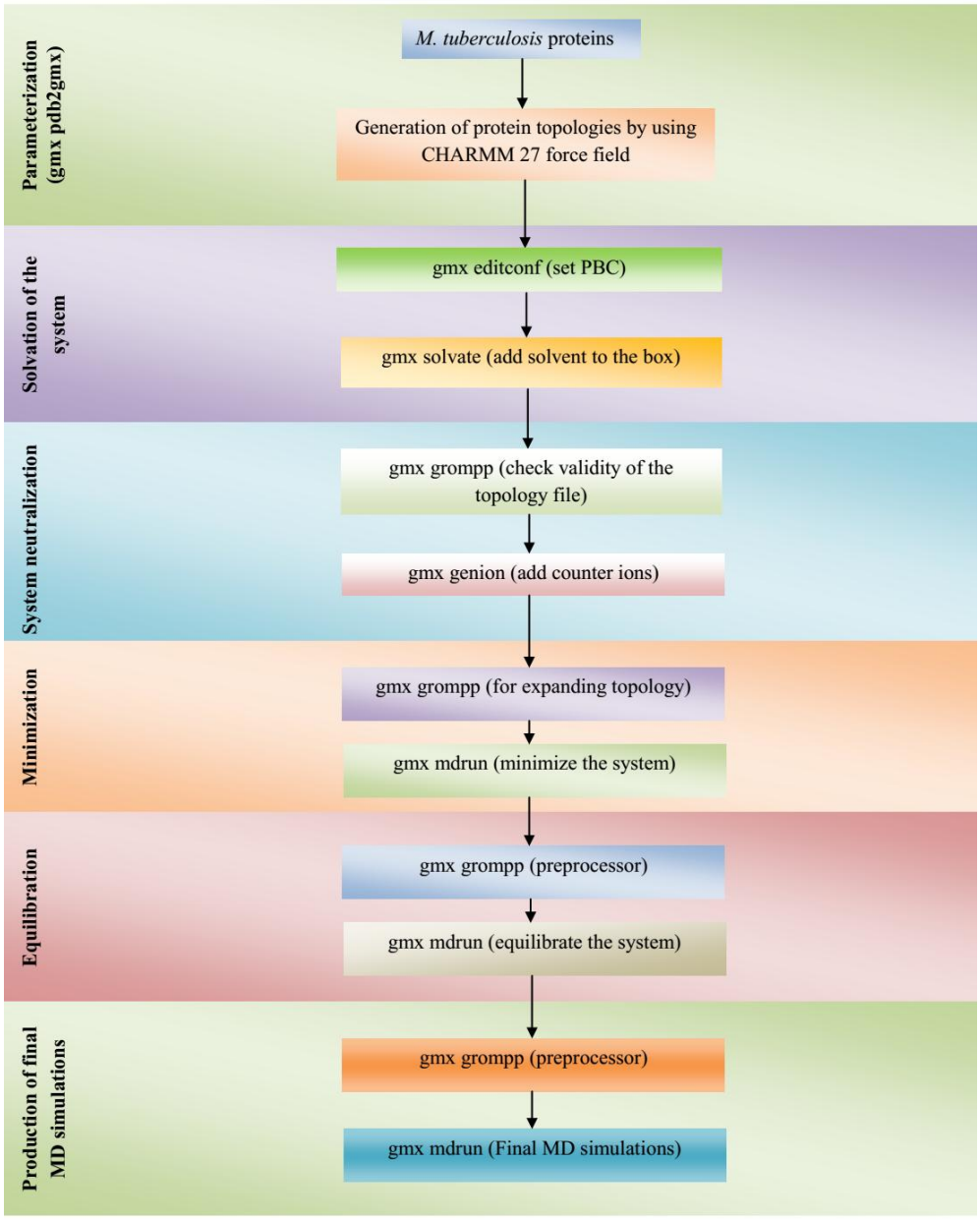


Figure 3.3: The steps involved in the MD simulations of the *M. tuberculosis* proteins using the GROMACS package.

3.2.1.4.5 Analyses of the trajectories

The utilities of the GROMACS analyzed the behavior of the trajectories obtained in the form of Root mean square deviation (RMSD), root mean square fluctuation (RMSF), the radius of gyration (R_g), Eigenvectors, free energy landscapes, and much more. The RMSD is among the most significant properties extracted from the trajectories during analysis and describe the nature of the protein in terms of stability and convenience with the experimental structure. The RMSD values of all the atoms in a protein in comparison to the reference structure were calculated on the basis of the following equation:

$$RMSD(t) = \left[\frac{1}{M}\sum_{i=1}^{N} m_i \mid r_i(t_1) - r_i(t_2)|^2\right]^{1/2}$$

Where $M=\Sigma_i m_i$, while $r_i(t)$ represents the location of atom i at time t.

Furthermore, in order to compute the average fluctuation experienced by all constituent residues, the RMSF was measured with respect to Cα atoms of protein and plotted as a function of residue number. The RMSF is calculated using "*gmx rmsf*" module of GROMACS by using:

$$\mathrm{RMSF}_i = \left[\frac{1}{N}\sum_{t_j=1}^{N}\left|r_i(t_j) - r_i(t_2)\right|^2\right]^{1/2}$$

where N is the number of polynucleotide structures under consideration, while $r_i(t_2)$ is the reference position of particle i.

Similarly, the radius of gyration of a biomolecule corresponds to the determination of its compactness. A relatively steady behavior was observed if a biomolecule assumes a stable configuration. The compactness of a structure was calculated as

$$\mathrm{R_g} = \left(\frac{\sum_i |r_i|^2 m_i}{\sum_i m_i}\right)^2$$

Where m_i is the mass of atoms and r_i is their atomic position with respect to the center of mass of the biomolecule.

Furthermore, other attributes in the form of total Mean energy of the system was calculated on the basis of the following equation:

$$\langle E \rangle = \frac{1}{N} \sum_{i=1}^{N} E_i$$

Where E_i represents the energy of atom i.

3.2.2 Results and Discussions

The information obtained from these analyses was used for the prediction of drug-resistance associated mutations. The SIFT server was used for predicting the intolerant point mutations by exhaustively mutating all the positions in the studied protein with 20 amino acids. The pathological mutations were classified using the PMut server. These classified disease-associated mutations were further analyzed using the protein stability prediction methods. They performed mutations on the sequences as well as structures of the studied proteins and provide the output in the form of free energy change (ΔΔG) and other attributes associated with protein functionalities. On the basis of consensus between the obtained outcomes, the disease-associated mutations were predicted. The conformational behaviors of the experimentally validated and computationally predicted mutations were compared and analyzed using the GROMACS package. The wild type and mutants proteins were simulated for 50 ns each (total 600 ns) time scale in explicit solvent conditions. Furthermore, the DFT-based methods provided the quantum mechanics based attributes associated with the point mutations, in which wild type and mutant type protein segment were analyzed. Moreover, the effects of point mutation on the binding of the drug molecules were studied using the molecular docking. The outcomes obtained for the

resistance studies of each of the following first line drug (Ethambutol, Isoniazid, Rifampicin, and Streptomycin) are presented below:

3.2.2.1 Ethambutol (EMB) mutations

The previous studies showed that EMB was designed to target *M. tuberculosis* arabinosyltransferases (encoded from *embABC*) which catalyze the polymerization of the arabinogalactan from arabinose (Sreevatsan et al., 1997). The arabinosyl transferase B was selected for the current study (Figure A.1). The experimentally validated drug resistant associated mutations which were collected from the literature (Baku et al., 2013) are listed in Table 3.2. Among these mutations, alteration at codon 306 of *embB* was the most frequently occurring variation and is considered as the marker for the rapid detection of the EMB mutations (Plinke et al., 2011, Baku et al., 2013).

After analyzing all these experimentally validated mutations on the basis of sequence and structure-based methods, the Leu413Pro substitution showed the highest deleterious effect on the structure of arabinosyl transferase B (Table 3.2). This mutation causes the destabilization of the arabinosyl transferase B. Furthermore, Glu504Gly also produced highly destabilizing effects, while Glu504Gln and Gly406Ala cause the stabilizing effects on the structure of arabinosyl transferase B. Therefore, the Leu413Pro was subjected to further analyses.

Despite the availability of the information regarding the mutations associated EMB resistance, still, the molecular basis of its tolerance is unknown. Therefore, there is a need to identify novel mechanisms behind the development of EMB tolerance in *M. tuberculosis*. Consequently, the diverse *in silico* methodologies were used in this study for the identification of hidden disease associated mutations.

Table 3.2: List of outcomes obtained after protein stability analyses of experimentally validated EMB resistant mutations in *M. tuberculosis*.

S. No.	Amino acid Mutation	Protein Stability							
		I-MUTANT 2.0 (Sequence based)		MuStab server		EASE-MM		SDM server	
		Predictions (Stability)	ΔΔG	Predictions (Stability)	Confidence score	Prediction (Stability)	ΔΔG	Predictions (Stability/effect)	ΔΔG
1)	**Met306Val**	(-)	-1.07	(-)	84.82%	DS	-1.04	S + N	-1.11
2)	**Met306Ile**	(-)	-0.94	(-)	84.82%	LD	-0.55	N + N	0.25
3)	**Gly406Ala**	(-)	-1.57	(-)	89.64%	Neutral	0.26	S + N	1.32
4)	**Leu413Pro**	(-)	-0.78	(-)	92.32%	DS	-2.20	HD + D	-5.96
5)	**Gln497Arg**	(-)	-0.53	(+)	26.96%	LD	-0.93	N + N	-0.47
6)	**Glu504Gln**	(-)	-0.05	(-)	79.82%	Neutral	-0.33	S + N	1.39
7)	**Arg507Gly**	(-)	-1.31	(-)	90.54%	DS	-1.11	HD + D	-2.11

LD = Likely destabilizing, DS = Destabilizing, LS= Likely stabilizing, N + N = Neutral and non-diseases associated, S + N = Stabilizing and non-disease associated, HD + D = Highly destabilizing/ cause protein malfunction and disease, DS + N = Destabilizing and non-disease associated, SD + N = Slightly destabilizing and non-disease associated, HS + D = Highly stabilizing/ cause protein malfunction and disease, SS + N = Slightly stabilizing and non-disease associated

After analyzing all the positions in the arabinosyl transferase B, 62 new drug resistance associated mutations were classified (listed in Table B.1). The Trp332Asn showed the comparable destabilizing effect to Leu413Pro. This former mutation showed the ΔΔG values to be around -2.49 kcal/mol, while the EASE-MM server showed increase ΔΔG of -3.79 kcal/mol during predictions and SDM server also indicated the destabilization of protein structure with ΔΔG values computed to be around -4.57 kcal/mol (Table B.1). Therefore, the Trp332Asn was selected for the further analyses among *in silico* predicted drug resistance associated mutations.

The DFT-based analyses of the mutations showed that both alterations Leu413Pro and Trp332Asn showed an increase in the energy. The overall energy calculated for WT Leu413 was around -3437.034 Ha (Figure A.2A), while for the mutant Pro413 the total energy was found to be about -3394.097 Ha (Figure A.2B). Similarly, the WT Trp332 and mutant Asn332 showed the calculated energy of the fragments to be around -3681.87 Ha and -3474.91 Ha respectively

(Figure A.2 C and D). Moreover, the HOMO-LUMO calculations suggested that the band gap energy after mutation Leu413Pro increase from 0.107 to 0.141 Ha, while for Trp332Asn the values decrease from 0.170 to 0.159 Ha with the distortion in the orbital space (Figure A.2). These observations indicated that both mutations causes the comparable changes in the structure of arabinosyl transferase B and may cause the EMB resistance in *M. tuberculosis*. Furthermore, in order to understand the effect of point mutations on the inhibitory action of EMB, a rigid molecular docking was performed.

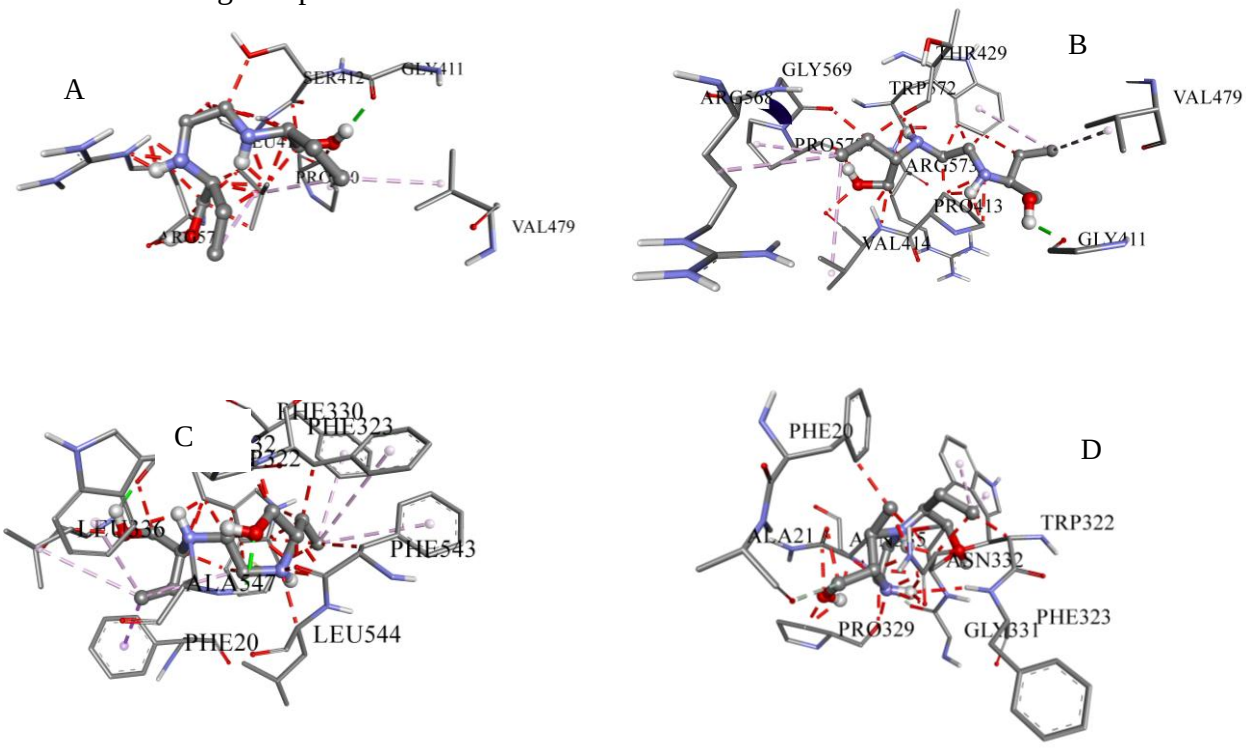


Figure 3.4: The changes in the interaction pattern due to point mutations in EMB. (A) The docked complex of EMB (Ball and stick) with the WT protein showing interaction with Leu413 (**B)** The change in the interaction pattern of EMB when the active site pocket contains mutation Pro413. **(C)** The molecular docking outcomes for EMB and WT protein showing the centered Trp332. (**D)** The resulted change in the interaction pattern when the grid is centered on mutating residue Asn332.

The interaction studies showed that the amino acid substitutions greatly affect the binding of the drug. The free energy of binding for Leu413Pro changes from -7.93 to -1.66 kcal/mol upon point mutations with the changes in the interaction pattern is depicted in Figure 3.4 A & B.

Similarly, for Trp332Asn the interaction energy changes from -5.01 to -1.99 kcal/mol (Figure 3.4 C & D)**.** These observations indicated that these point mutations greatly affect the functioning of the EMB and may cause its resistance in *M. tuberculosis*.

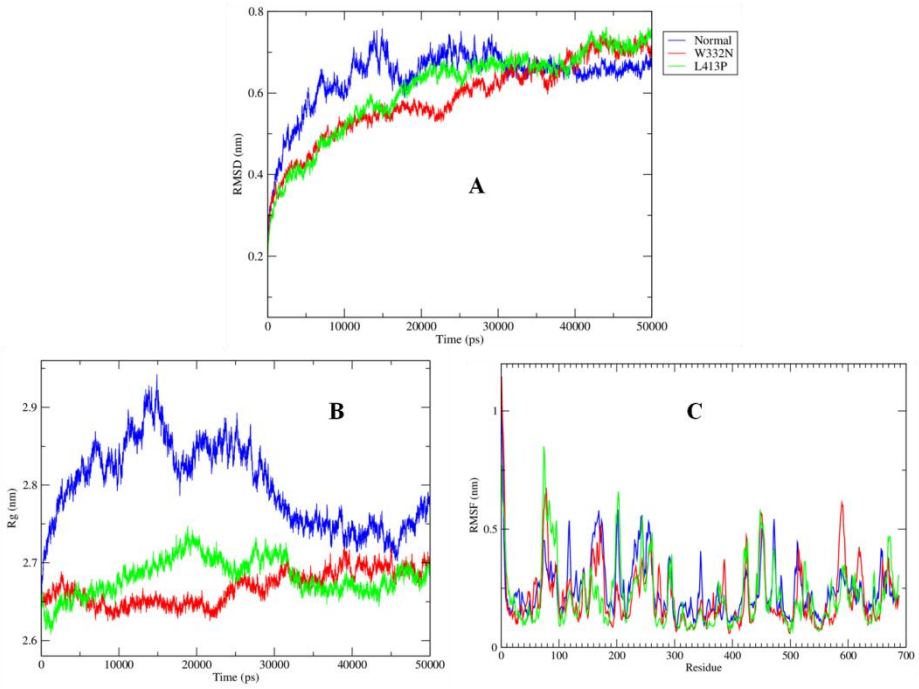


Figure 3.5: (A) The variations in the RMSD values observed after simulating each protein for 50 ns in explicit water condition. (**B)** Rg plots showing the fluctuations in the compactness of the studied structures. **(C)** The fluctuations observed in constituent residues of the WT and mutant arabinosyl transferase B.

The structural alterations were further analyzed using MD simulations in which WT and the two above studied mutations were simulated each for 50 ns time scale under explicit solvent conditions. The RMSD values showed higher fluctuations up to 30 ns and showed lower values than the rest of the mutants (Figure 3.5A). While the structure of both mutants Leu413Pro and Trp332Asn showed a steady increase in RMSD values till 50 ns. The mutant W332N showed the highest compactness in its structure, while slight changes observed for Leu413Pro (Figure 3.5B). But WT arabinosyl transferase B showed higher fluctuation in the Rg values which ranges in between 2.7 – 2.9 nm. Furthermore, the constituent residues in all the cases showed comparable

fluctuations (Figure 3.5C). These outcomes showed that higher structural compactness may lead to the EMB resistant conditions.

3.2.2.2 Isoniazid (INH) mutations

The INH resistance-conferring missense mutations in *M. tuberculosis* are generally observed in a bacterial *katG* gene which expresses to form catalase - peroxidase enzyme (Pym et al., 2002, Marttila et al., 1996). The 3-D structure used for the current study is illustrated in Figure A.3. KatG is involved in the protection of the *M. tuberculosis* against the oxidative stresses inside the macrophages (Manca et al., 1999). Through experimental studies, several INH resistance causing mutants were isolated (Pym et al., 2002). The behavior of such experimentally validated mutations was analyzed in this study (listed in Table 3.3). But still, there is a requirement for the identification of additional INH resistance-conferring mutations, which can provide conventional ways for curing the infection resulting from drug-resistant TB.

Table 3.3: List of experimentally validated mutations that leads to INH resistance in the patients infected with MDR strains of *M. tuberculosis*.

S. No.	Amino acid Mutation	Protein Stability							
		I-MUTANT 2.0 (Sequence based)		MuStab server		EASE-MM		SDM server	
		Predictions (Stability)	ΔΔG	Predictions (Stability)	confidence	Predictions (Stability class)	ΔΔG	Predictions (Stability/ effect)	ΔΔG
1)	**Ser315Thr**	Increase	0.87	Increased	24.46%	LD	-0.69	S +N	1.07
2)	**Ser315Arg**	Increase	-0.58	Increased	25.18%	LD	-0.81	S + N	1.93
3)	**Ser315Asn**	Increase	-0.18	Decreased	78.57%	DS	-1.07	SS + N	0.59
4)	**Gly299Ser**	Decrease	-0.72	Decreased	87.86%	LD	-0.88	N + N	-0.06
5)	**Gly309Val**	Decrease	-2.37	Decreased	79.29%	LD	-0.84	HS + D	4.07
6)	**Asn323Ile**	Increase	2.75	Increased	26.61%	LS	0.60	SS + N	0.58
7)	**Pro325Ala**	Decrease	2.00	Decreased	79.64%	DS	-1.44	N + N	0.30

By using the *in silico* analyses, we have classified 61 mutations in the KatG protein (Uniprot ID - P9WIE5), which are listed in Table B.2. The mutations which showed highest

deleterious effects on the structure of the KatG were selected for the further analyses. Among experimentally validated mutations the Gly309Val was selected, while among *in silico* predicted mutations the Asp419Trp showed highest deleterious effect.

In order to analyze the quantum basis of their effect, the DFT-based methods implemented in the Gaussian software was used. Both the mutations, Gly309Val and Asp419Trp showed a decrease in the total energy. The energy for Gly309 was calculated to be around -2759.66 Ha (Figure A.4A) which decreases to -2884.50 Ha upon replacement with Val309 (Figure A.4B). Similarly, the mutation Asp419Trp may lead to a decrease in the total energy from -3534.86 to -3722.18 Ha (Figure A.4 C and D). The band gap energy for both the cases showed an increase in the magnitude, with higher value was calculated for the *in silico* predicted mutation (Figure A.4). The observations suggested that these mutations cause the stabilizing effect on the structure of KatG.

The effects of these drug resistance-associated mutations were further analyzed using molecular docking which showed a decrease in the inhibition effect of INH. The experimentally validated mutation Gly309Val reduces the binding energy from -3.60 to -2.85 kcal/mol (Figure 3.6 A & B). The substitution Asp419Trp leads to the increase in the hydrophobicity by the introduction of aromatic rings of the tryptophan and may affect the orientations of amino acid in the interaction pocket of INH. These structural changes influence the INH mode of action, which can observe through the free energy of binding that changes from -3.31 to -2.12 kcal/mol. These molecular docking studies provided a better insight into the altered effect of the studied point mutations on the performance of the INH.

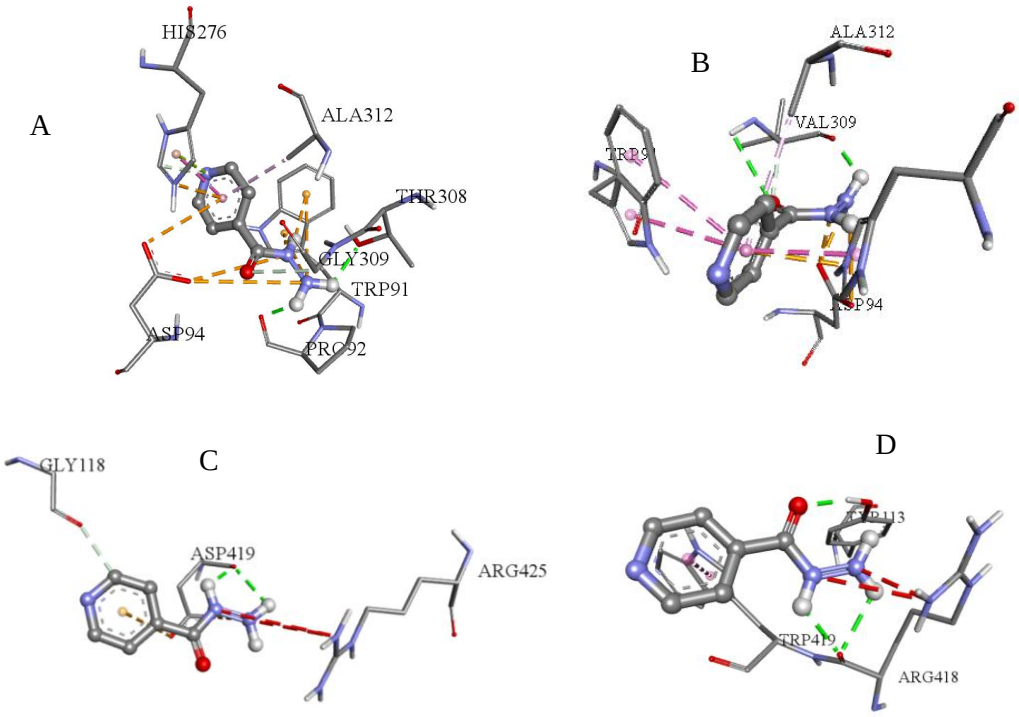


Figure 3.6: (A) The Ball and stick representation of docked complex of INH with the WT protein showing interaction with Gly309 (**B)** The observed variation in the interaction pattern of INH when the active site pocket contains Val309. **(C)** The stable docked conformation for INH and WT protein which was centered at Asp419. (**D)** The resulted alteration in the interaction pattern when the grid is center changed to Trp419.

Furthermore, the differences in the conformational behavior of the WT and mutant KatG was analyzed in the explicit solvent conditions for 50 ns time scale. The utilities of the GROMACS were used in order to analyze the trajectories of all the proteins. The RMSD plot in Figure 3.7A showed that the WT KatG showed variations between 0.4 nm – 0.5 nm after 20 ns time scale and Asp419Trp showed slightly higher values, while the Gly309Val significantly increases the structural stability as indicated by the RMSD values. Furthermore, the compactness measured in the form of Radius of gyration showed slightly lesser fluctuations for both Gly309Val and Asp419Trp as compared to the WT protein (Figure 3.7B). The majority of constituent residues showed comparable fluctuations (Figure 3.7C).

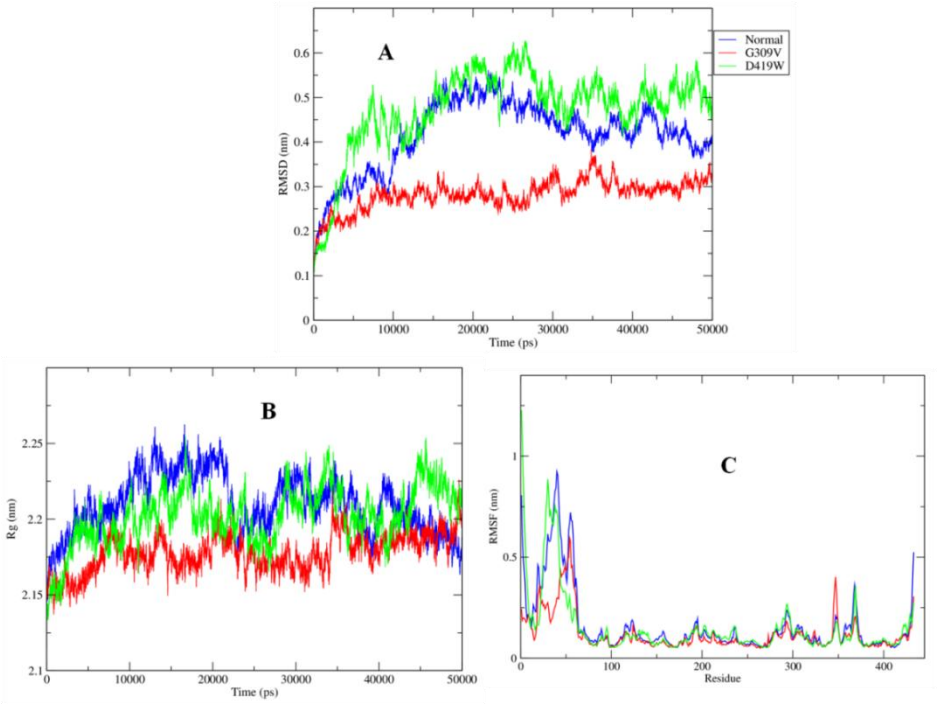


Figure 3.7: (A) The curve showing the variations in the RMSD values observed during 50 ns MD simulations in explicit water conditions. (**B)** The graph depicting the effects of point mutations on the fluctuation of the Rg values for the studied structure. **(C)** The RMSF values observed for the constituent residues of WT and mutant KatG.

3.2.2.3 Rifampicin (RIF) mutations

In addition to the above-described cases, the mutations related to the RIF associated resistance also reported in the *M. tuberculosis* (Comas et al., 2012, Boehme et al., 2010). The RIF is one of the primary first-line drugs used in treating the infection associated with TB by inhibiting the activity of β-subunit of RNA polymerase (Comas et al., 2012, Boehme et al., 2010). The RNA polymerase β-subunit (RpoB) structure is illustrated in Figure A.5. The latter protein is encoded by the *rpoB* gene (Comas et al., 2012, Boehme et al., 2010). The RIF tolerance frequently occurs by an alteration in the 81 bp segment of *rpoB* which encodes the residues ranging from position 507-533 (Sanchez-Padilla et al., 2015). Among a variety of experimentally validated mutations (Telenti et al., 1993, Valim et al., 2000, Weiss et al., 2012), we have selected the mutations listed in Table 3.4. The His451Asn showed highest deleterious

and destabilizing effect on RpoB, selected for further analyses in this study. Despite the discovery of numerous RIF resistant associated mutations in RpoB, the extensive research leads to the identification of the new alterations (Sanchez-Padilla et al., 2015). Therefore, an exhaustive search was performed for the identification of the novel RIF tolerance associated mutations. Around 104 mutations were selected that may lead to the RIF-resistant mutations in *M. tuberculosis* (Table B.3), with Leu500Lys showed the highly destabilizing effect on the structure of RpoB and selected for further study in the category of computationally predicted mutations.

Table 3.4: List of experimentally validated mutations associated with RIF's resistance in *M. tuberculosis* (http://www.uniprot.org/uniprot/P9WGY9).

S. No.	Amino acid Mutation	Protein Stability							
		I-MUTANT 2.0 (Sequence based)		MuStab server		EASE-MM		SDM server	
		Predictions (Stability)	ΔΔG	Predictions (Stability)	confidence	Predictions (Stability Class)	ΔΔG	Predictions (Stability/ effect)	ΔΔG
1)	**Val423Ala**	(-)	-1.86	(-)	92.86%	DS	-1.61	SD + N	-0.68
2)	**Leu436Pro**	(-)	-1.34	(-)	93.93%	LS	0.98	DS + N	-1.23
3)	**Ser437Thr**	(-)	-0.36	(-)	86.07%	N	-0.16	SS + N	0.61
4)	**Gln438Leu**	(+)	0.50	(-)	78.75%	N	0.13	HS + D	2.28
5)	**Phe439Val**	(-)	-2.70	(-)	88.21%	LD	-0.68	N + N	-0.09
6)	**Asp441Val**	(-)	0.32	(-)	83.39%	N	0.24	HS + D	2.58
7)	**His451Asp**	(-)	-0.15	(-)	82.86%	DS	-1.02	SD + N	-0.67
8)	**His451Leu**	(+)	0.30	(+)	27.14%	N	-0.01	SS + N	0.55
9)	**His451Asn**	(-)	-2.31	(-)	82.32%	DS	-1.08	HD + D	-3.27
10)	**His451Pro**	(+)	0.30	(-)	85.89%	LD	-0.81	HD + D	-2.03
11)	**His451Gln**	(-)	-0.19	(-)	78.93%	LD	-0.72	HD + D	-2.29
12)	**His451Arg**	(-)	-0.10	(+)	22.86%	LD	-0.58	DS + N	-1.78
13)	**Ser456Leu**	(-)	-0.48	(-)	78.75%	N	0.02	HS + D	2.64
14)	**Ser456Gln**	(+)	-0.32	(-)	84.29%	N	-0.12	N + N	-0.39
15)	**Ser456Trp**	(+)	0.16	(-)	78.75%	N	-0.34	DS + N	-1.67
16)	**Leu458Pro**	(-)	0.22	(-)	85%	DS	-1.44	DS + N	-1.43

The His451Asn leads to the loss of the imidazole side chain of the histidine and resulted in the introduction of the carboxamide side chain of asparagine may cause the reduction in the hydrophobicity. Similarly, Leu500Lys substitution causes the replacement of non-polar isobutyl

side chain of leucine with the polar lysyl side chain of lysine may also result in the decrease in the hydrophobic characteristic of RpoB. The resulted effects of both His451Asn and Leu500Lys were studied in details using the DFT-based methods which showed that increment in the total energy of the system. During His451Asn substitution, the energy increased from -3006.38 to -2946.56 Ha and for Leu500Lys the energy changes from -2869.12 to -2927.56 Ha (Figure A.6 A-D). These point mutations greatly reduce the inhibition of RpoB by RIF drug. The docking studies illustrated the effects of point mutations when RIF molecule was docked at the site of the point mutations. The free energy of binding for His451Asn showed elevation from -6.89 to -5.50 kcal/mol, while for Leu500Lys it changes from -6.41 to -5.57 kcal/mol, which indicated the high deleterious nature of these point mutations (Figure 3.8)

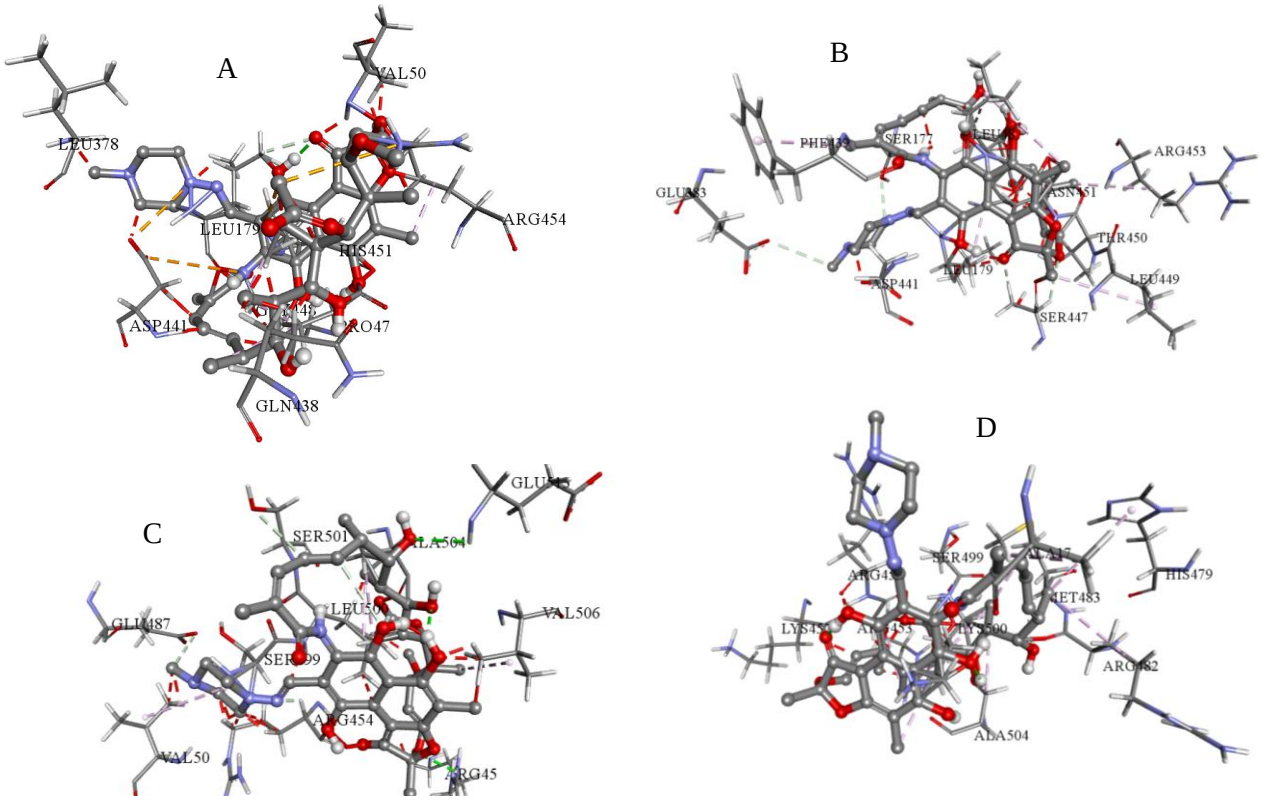


Figure 3.8: (A) The observed interacting residues when RIF docked with the WT RpoB. **(B)** The interaction pattern of the residues changes when the binding pocket contains Asn451. **(C)** The bound conformation generated after RIF docked with WT protein focusing the site of mutation Leu500. **(D)** The docked pose with centered on Lys500 showed changes in the interaction pattern.

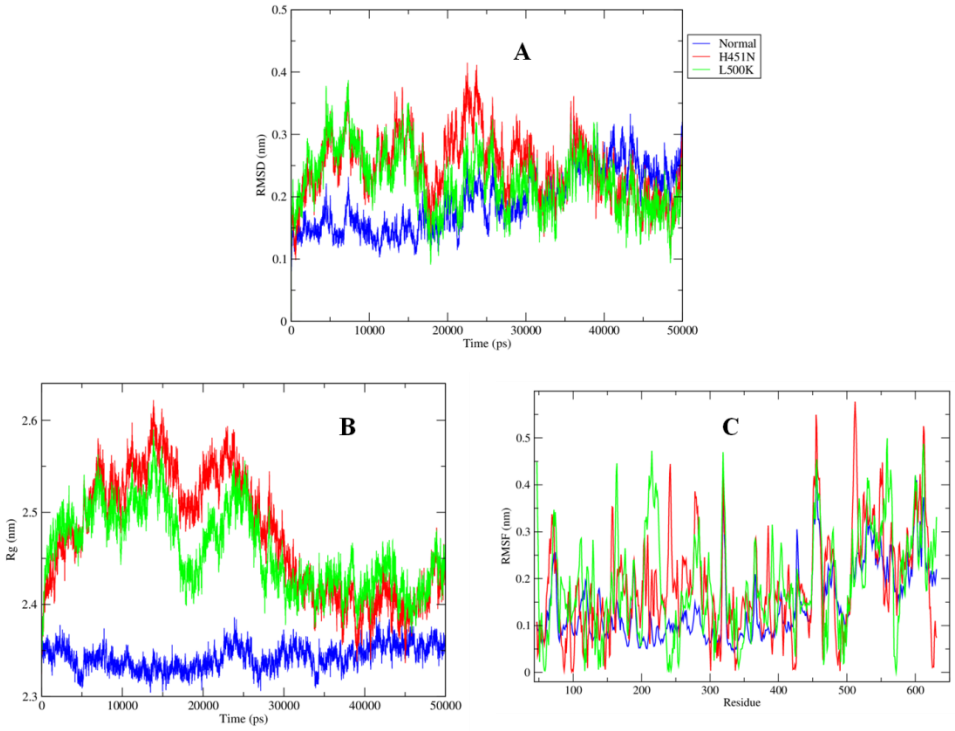


Figure 3.9: (A) The fluctuations in the RMSD values observed for the WT and mutant RpoB proteins after MD simulation studies. (**B)** The Rg curves illustrating the variations generated in the compactness of the RpoB after point mutations. **(C)** The variations in the RMSF observed for the constituent residues of the WT and mutant RpoB.

The conformational basis of the mutations was further observed using the principles of the MD simulations. Both the substitutions His451Asn and Leu500Lys exhibited a highly destabilizing effect on the structure of the RpoB, as illustrated by 50 ns MD simulations. Higher variations were observed in the RMSD values for the studied mutations in comparison to the WT RpoB (Figure 3.9A). The higher compactness was measured for the WT protein during the 50 ns MD simulations (Figure 3.9B), while constituent residues showed very high fluctuations due to the increase in the energy of the system (Figure 3.9C). These observations indicated that the RIF-resistant mutations exhibited a destabilizing effect on the structure of RpoB.

3.2.2.4 Streptomycin (SM) mutations

Streptomycin is the first drug to be formulated against the infection of the TB in 1943 and is widely used as the first-line drug (Keshavjee and Farmer 2012). The SM resistance conditions usually arise with the mutations in the ribosomal protein such as rpsL (the structure is shown in Figure A.7) in which mutation at the codon 43 is most frequently isolated from patients suffering from the drug-resistant TB (Sreevatsan et al., 1996). Among the experimentally validated substitution (listed in Table 3.5)**,** Val80Gly showed the high destabilizing effect on the rpsL (Table 3.5). While the *in silico* predicted mutation Leu102Trp (Table B.4) showed the comparable effect on the structure of rpsL by the introduction of the aromatic side which resulted in the increase in the hydrophobicity in rpsL. The DFT analyses of both the mutations showed that Val80Gly resulted in the increase in the energy from -3176.45 to -3051.13 Ha, while Leu102Trp showed a decrease in the total energy from -3313.36 to -3573.95 Ha (Figure A.8).

Table 3.5: Stability analyses of experimentally validated mutations leading to the Streptomycin resistance in *M. tuberculosis*.

S. No.	Amino acid Mutation	Protein Stability							
		I-MUTANT 2.0 (Sequence based)		MuStab server		EASE-MM		SDM server	
		Predictions (Stability)	ΔΔG	Predictions (Stability)	confidence	Predictions (Stability)	ΔΔG	Predictions (Stability/ effect)	ΔΔG
1.	**Lys43Arg**	(-)	0.06	(+)	32.14%	N	-0.43	SB + N	1.86
2.	**Ala48Pro**	(+)	-0.58	(-)	80.54%	LD	-0.72	HD + D	-3.92
3.	**Ala53Glu**	(-)	-0.42	(-)	92.86%	LD	-0.73	HD + D	-3.71
4.	**Leu74Arg**	(-)	-2.10	(-)	81.61%	DS	-1.30	HD + D	-2.29
5.	**Gln75Leu**	(+)	0.43	(-)	83.57%	N	0.09	N + N	-0.50
6.	**Val80Gly**	(-)	-4.59	(-)	89.64%	DS	-3.88	HD + D	-5.09
7.	**Arg83Gln**	(-)	-1.18	(-)	87.32%	DS	-1.29	DS + N	-1.37
8.	**Arg86Pro**	(-)	-1.07	(-)	84.11%	LD	-0.51	HD + D	-3.68
9.	**Leu90His**	(-)	-2.45	(-)	88.93%	DS	-1.63	HD + D	-2.43
10.	**Gly92Ala**	(-)	-1.56	(-)	80.71%	LD	-0.95	HD + D	-2.35
11.	**Tyr95His**	(-)	-1.01	(-)	91.79%	DS	-1.44	HD + D	-2.85
12.	**Ile98Thr**	(-)	-3.07	(+)	28.04%	DS	-2.02	HD + D	-2.29

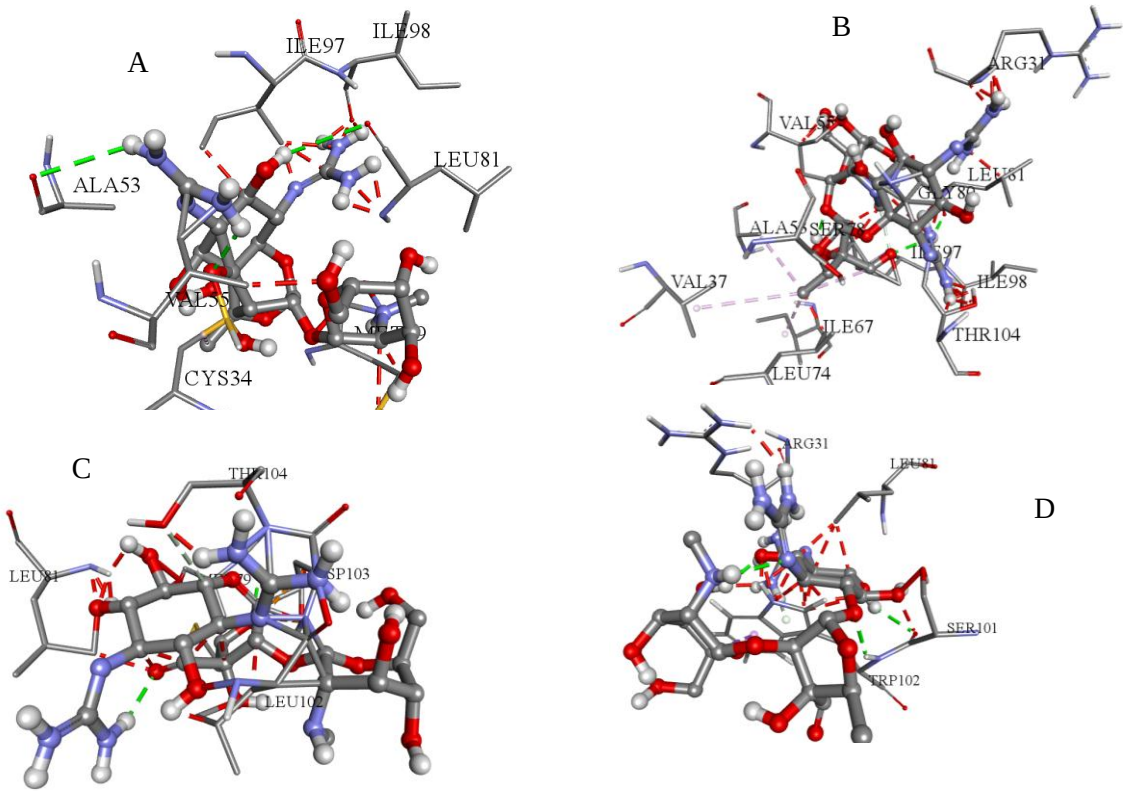


Figure 3.10: (A) The docked complex showing the WT interacting residues with SM (Ball and stick). (**B)** The change observed in the docked pattern of SM when the binding site contains Gly80. **(C)** The docked pose showing the SM and WT protein containing the mutation site of Leu102. (**D)** The illustration of mutations coupled changes in the interaction pattern.

The binding studies of SM on the site of the point mutations provide an understanding regarding the effect of amino acid alteration on the functioning of the drug (Figure 3.10). There is a difference observed in the free energy of binding between the WT and the mutant proteins. In the case of Val80Gly, the interaction energy changes from -3.31 to -2.11 kcal/mol. Similarly, for Leu102Trp the magnitude of the binding energy increases from -3.93 to -2.99 kcal/mol. These observations indicated that the predicted mutations are also causing an equivalent deleterious effect on the structure of the rpsL and can be utilized for further studies in the process of designing better drugs.

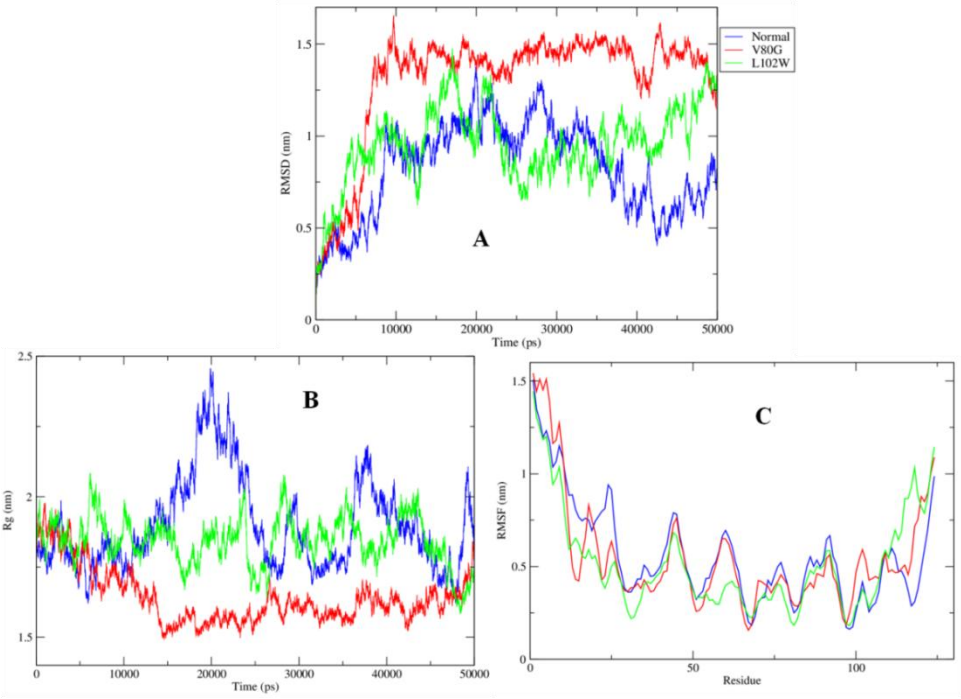


Figure 3.11: (A) The RMSD plot showing increased instability of the protein structures upon point mutations. (**B)** Rg plots illustrating that the mutations may result in higher compactness in the structure of rpsL. **(C)** The RMSF curve exhibiting the comparable fluctuation in the constituent residues of WT and mutant rpsL.

Additionally, after simulating the structure of all the proteins, both the mutants of rpsL showed comparable stability profiles in the explicit water conditions as compared to the WT rpsL. The RMSD values of rpsL mutants were fluctuating in the range of 1 nm – 1.5 nm with a magnitude higher than WT rpsL (Figure 3.11A). While Rg and RMSF plots showed higher values for WT rpsL, indicating its less stability as compared to the mutant structures studied (Figure 3.11B and C).

3.2.3 Conclusions

The development of the drug resistance in *M. tuberculosis* is the biggest challenge to the current treatment of TB infection. The isolates of patients suffering from drug-resistant TB showed the presence of various point mutations in the proteins. Therefore, extensive analyses of the mutational landscapes were performed on the proteins involved in the drug resistance. A

variety of novel point mutations that may be responsible for the development of the drug resistance were predicted using *in silico* techniques and their effects on the structure were compared with experimentally validated substitutions. The resulted structural changes were observed on the basis of available DFT, molecular docking, and MD simulations techniques. The conformational behaviors of the computationally predicted mutations were comparable with the experimentally derived mutations and therefore may be involved in the mechanism of drug resistance. These mutational studies on the *M. tuberculosis* proteins can be utilized in improving the current drug therapy. The knowledge generated through the mutational analyses enable a better understanding of the drug resistance. In addition to these drugs tolerance mechanisms, the *M. tuberculosis* possess a variety of other resistance mechanisms and the conformational analyses of the proteins involved in such virulence pathways are required, which are discussed in the subsequent sections:

3.3 Acid resistance analyses

In addition to the drug resistance, the inherent mechanisms involved in the development of the resistance in *M. tuberculosis* against the acidic conditions enable its survival in the human body. This section involves the analyses of virulence factors associated with acid resistance. A variety of acid resistance mechanisms have been discovered for enteric pathogens which can survive in extremely acidic environments in the human hosts, but little literature is present for *M. tuberculosis* (Vandal et al., 2009, Darby et al., 2011). In the immunologically inactive macrophages, the mycobacterial species inhibits the fusion of phagosomes and lysosomes and thereby exists in the mildly acidic environment of pH ~ 6.2 (Vandal et al., 2009). The pH falls to around 4.0 (Figure 1.2) after the activation of the macrophages by the action of the macromolecules such gamma interferon (Vandal et al., 2009). Several proteins such as isocitrate lyase, Lipase lipF, and others, involved in fatty acid metabolism showed over-expression during the *M. tuberculosis* infection (Vandal et al., 2009).

3.3.1 Materials and Methods

On the basis of information present in the literature, the virulence factors such as Isocitrate Lyase (ICL), Lipase (lipF), Magnesium transporter (MgtC), Porin (OmpATb), Two-component regulatory systems (PhoP), Undecaprenyl pyrophosphate phosphatase (Rv2136c), and Serine protease (Rv3671c) were analyzed to be involved in the development of the acid resistance in *M. tuberculosis* and are crucial for its survival inside the macrophages (Vandal et al., 2009). Therefore, these proteins were selected for the constant pH based MD simulation studies, in which their structural basis of the acid tolerance was observed:

3.3.1.1 Constant pH MD simulations

The primary analyses involved the calculation of the pKa values of the titratable sites present in each protein. This was achieved by the usage of "Calculate Protein Ionization and Residue pK" module of the Discovery Studio (DS) which calculate stability parameters and titratable properties for the studied proteins at the diverse pH range. These analyses indicated that all the proteins showed the minimum relative folding energy around pH 3.0 – 5.0 (Figure 3.12).

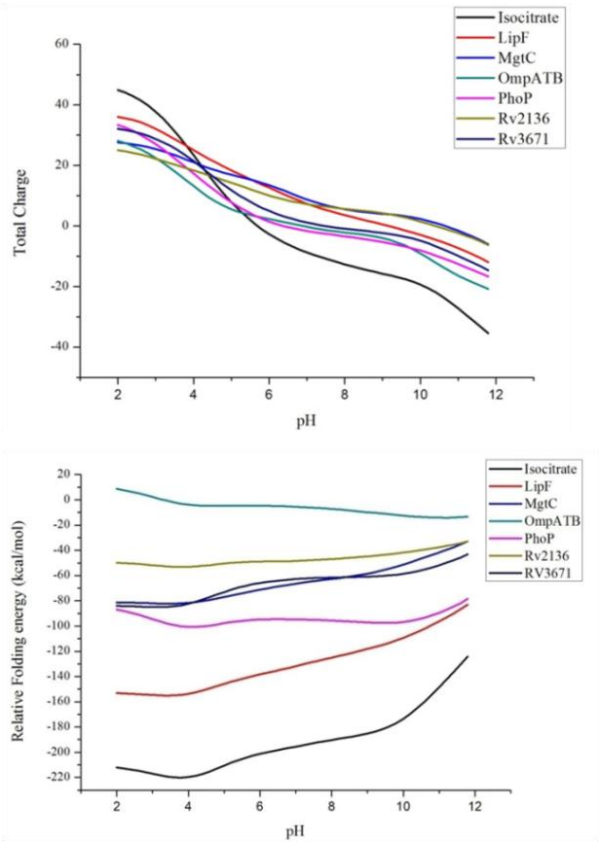


Figure 3.12: The graphical view of the variation observed in the stability parameters of the seven studied proteins at the diverse pH range.

After calculations of the pKa values of each titratable site, the desired pH conditions were created by altering the protonation state of the titratable sites (Asp, Glu, Arg, Lys and His) in the proteins using "pdb2gmx" module of the GROMACS. The acidic environment ranging from pH 3.0 - 6.0 were created and proteins were parameterized using the CHARMM 27 force field.

Then, the system was solvated and neutralized by adding the suitable counter ions. The neutralized systems were minimized for 1000 steps of steepest descent. The minimized system was equilibrated under NVT and NPT ensemble conditions each for 100 ps time period and final MD simulation was performed using LINCS algorithm for each system at 50 ns time scale (for 28 systems total time is 1400 ns). LINCS is an algorithm that resets bonds to their correct lengths after an unconstrained update.

3.3.2 Results and Discussions

The constant pH based MD simulations provide a new insight into the structural behavior of the proteins involved in the development of acid resistance. The outcomes obtained for each protein are discussed in the separate section below:

3.3.2.1 Isocitrate Lyase (ICL)

ICL is the primary enzyme of the glyoxylate cycle, an alternative pathway present in a variety of pathogenic microorganisms in the absence of tricarboxylic acid (TCA) cycle. In order to stop the infection of the bacteria in the dormant stage, this enzyme serves as the primary targets in the drug design (Lee et al., 2015). The structure of the ICL composed of eight α-helices and 17 β-strands forming an unusual α/β barrel (Figure 3.13) and was subjected to 50 ns MD simulations each at diverse pH conditions (Figure 3.14).

The utilities present in the GROMACS observed the dynamical behavior of ICL. The continuous variation was observed in the RMSD values (Figure 3.14 A), with least values between the range of 0.3 – 0.4 nm were observed at pH 4.0 followed by pH 3.0 as compared to the other conditions indicating the stability of the ICL at acidic conditions. The highest structural

compactness was observed at pH 6.0 followed by pH 4.0 indicating that ICL is able to maintain its structural integrity in acidic conditions (Figure 3.14B).

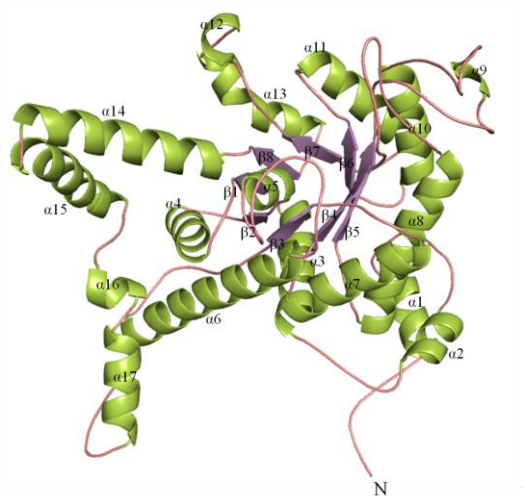


Figure 3.13: The structure of the ICL showing characteristic α/ β barrel.

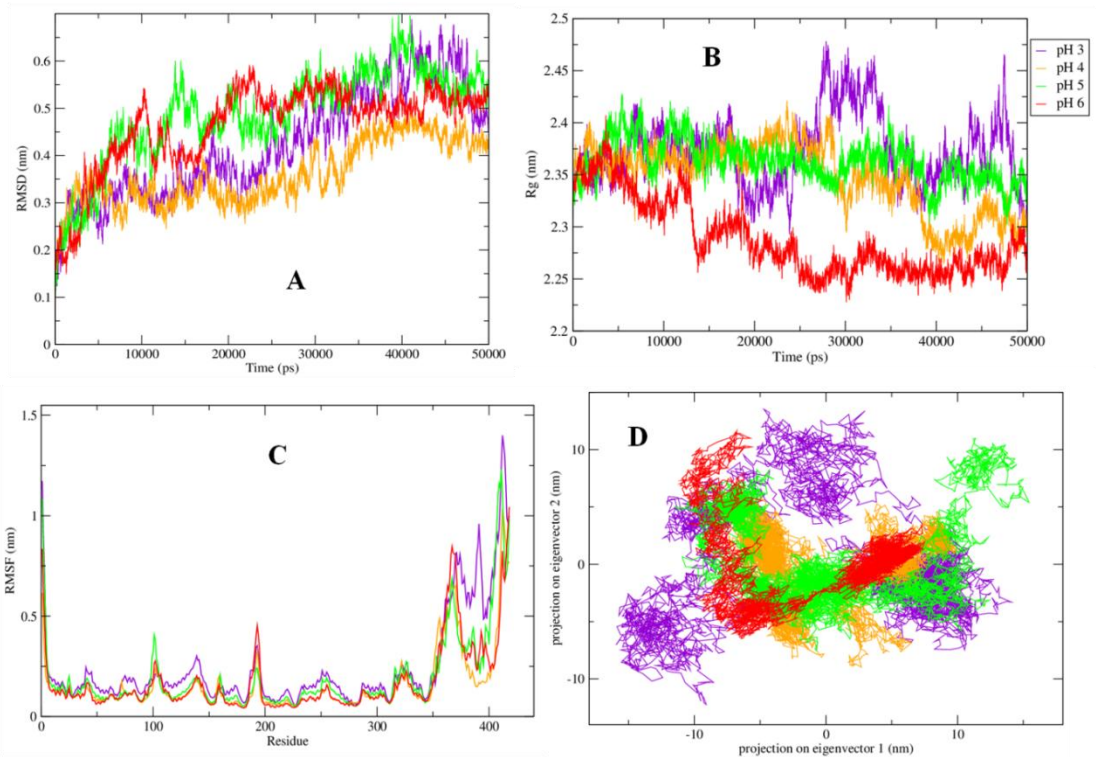


Figure 3.14: The graphs illustrating changes in **(A)** RMSD **(B)** Rg values **(C)** RMS fluctuations **(D)** the eigenvalues for ICL at a pH range of 3.0 – 6.0.

At pH 3.0, the constituent residue of the ICL showed higher fluctuation as compared to the other conditions (Figure 3.14C) and least atomic motions were experienced at pH 4.0 and 6.0 acidic

conditions (Figure 3.14D). The outcomes of these constant pH MD simulations indicated that the ICL can maintain its functional integrity at the latter pH conditions.

3.3.2.2 Lipase (lipF)

The lipF was characterized to be a lipase or esterase (Zhang et al., 2005) and is the primary virulence factor of *M. tuberculosis*, which specifically showed up-regulation in the acidic conditions (Richter and Saviola, 2009). The structure of the lipF assumes α/β topology (Figure 3.15). On simulating the structure of lipF at different acidic conditions, the least RMSD fluctuations were observed at pH 5.0 as compared to the other acidic conditions (Figure 3.16A). While highest structural compactness was observed at pH 4.0 and 6.0 after assessing the Rg plots (Figure 3.16B). The least fluctuations in the constituent residues were observed in the pH 6.0 (Figure 3.16C), while at pH 5.0 and 6.0 the minimum atomics motions were observed (Figure 3.16D). These observations indicated that lipF can maintain its functionalities at diverse pH conditions which were correlated with the information present in the literature (Richter and Saviola, 2009).

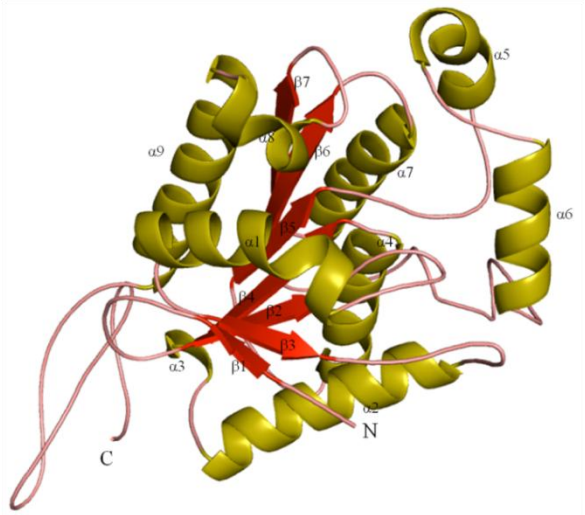


Figure 3.15: The structure of lipF showing characteristic α/β topology.

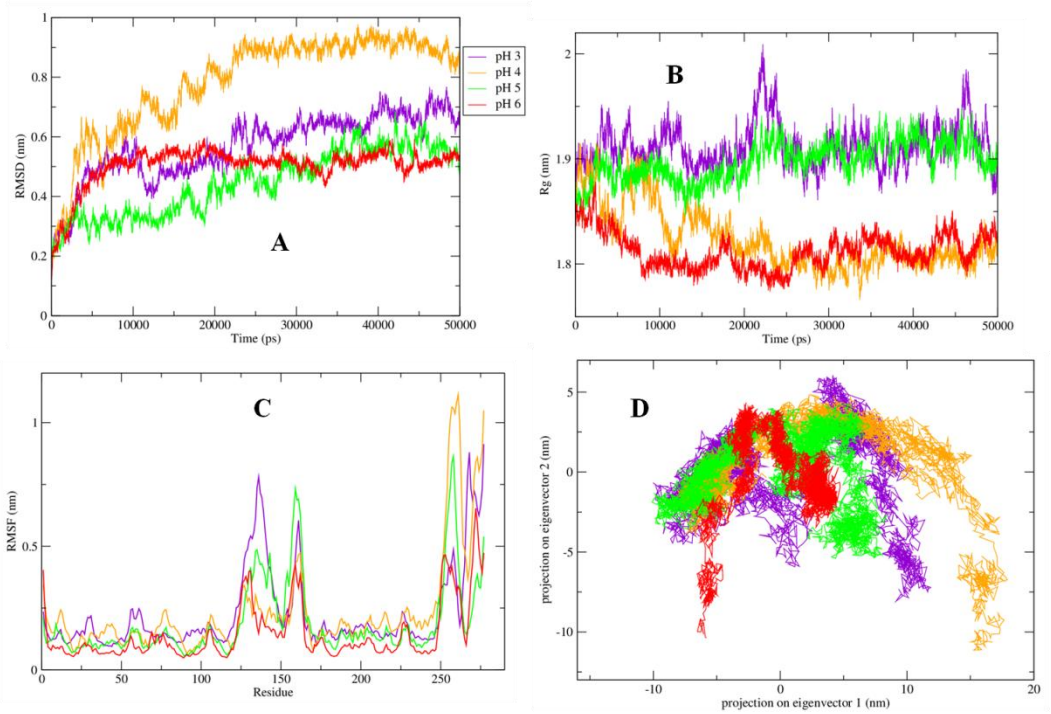


Figure 3.16: The curves highlighting the changes in **(A)** RMSD values **(B)** Rg values **(C)** RMS fluctuations **(D)** the eigenvector values for lipF at a pH range of 3.0 – 6.0.

3.3.2.3 Magnesium transporter (MgtC)

The MgtC is a transporter protein responsible for the survival of the *M. tuberculosis* in the mildly acidic environment of the macrophages as well as in the stress conditions with Mg^{2+} deprivation (Yang et al., 2012, Vandal et al., 2009). The insufficient information is available about the conformational behavior of the MgtC (Vandal et al., 2009), therefore its structure was simulated for 50 ns each at different pH conditions. The structure of MgtC showed characteristic ACT domain as its structure assumes the βαββαβ motif topology with two α-helices (α8 and α9) present on the single side of the four antiparallel β-strands forming a sheet (Figure 3.17).

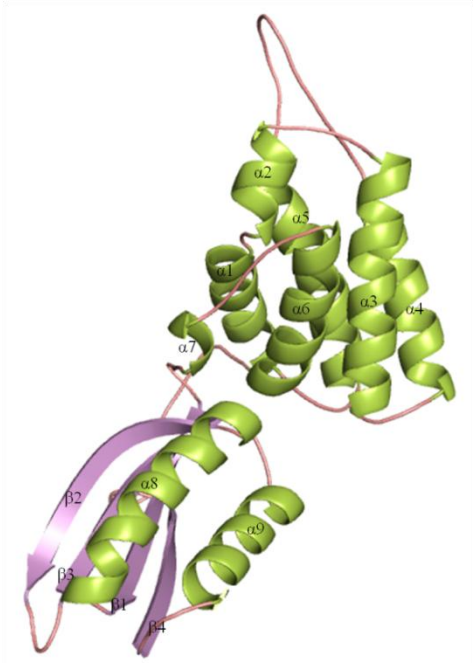


Figure 3.17: The structure of MgtC showing characteristic ACT domain

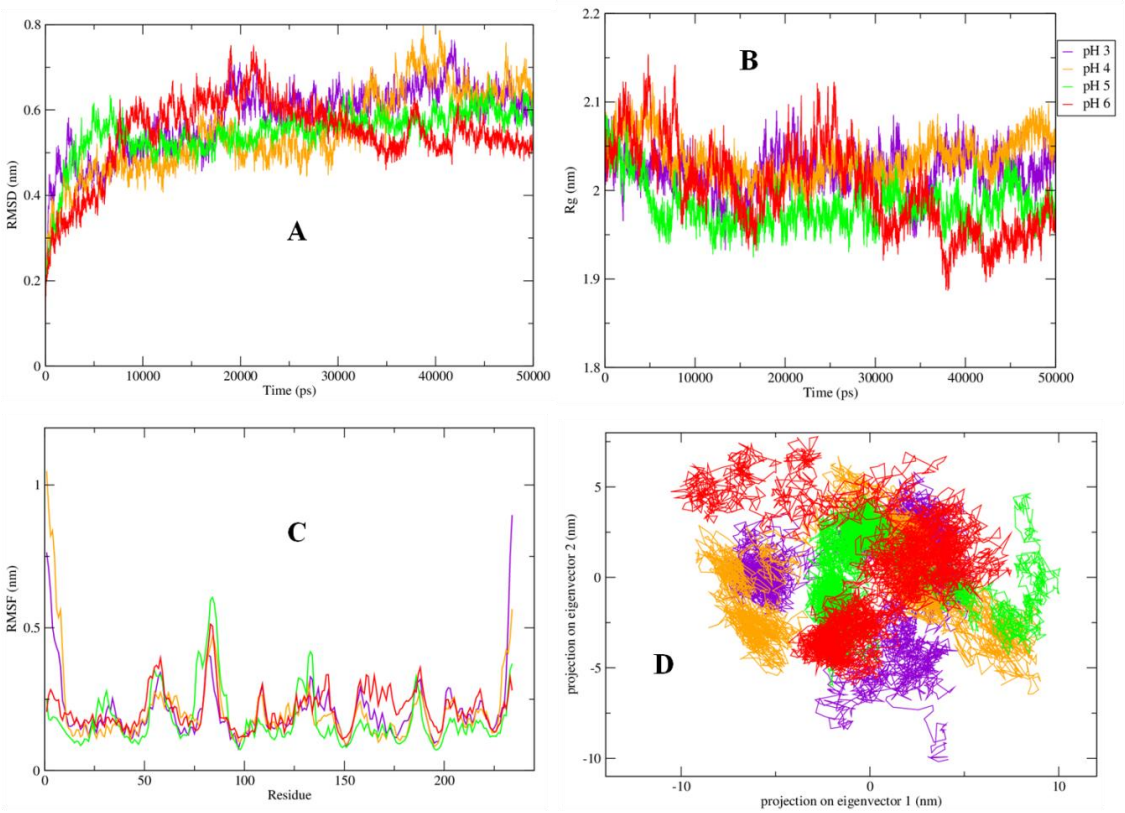


Figure 3.18: The graphical illustrations highlighting changes in **(A)** RMSD values, **(B)** Rg values, **(C)** RMS fluctuations, and **(D)** the eigenvalues, for MgtC at different acidic conditions.

The MD simulation results showed the presence of comparable conformational behavior at all the studied conditions. The RMSD values in every acidic condition were fluctuating in between 0.3 nm – 0.7 nm (Figure 3.18A) and variations in the Rg values were observed in the range of 1.9 – 2.1 nm (Figure 3.18B). This indicated the structural features of MgtC were maintained at diverse acidic conditions. Furthermore, comparable fluctuations and motions were observed in the constituent residues of the MgtC during the course of MD simulations (Figure 3.18C and D). This study indicated that the MgtC can maintain its functional integrity even in high acidic environment present in the human host.

3.3.2.4 Porin (OmpATb)

The OmpATb is characterized to be a pore forming protein and crucial for developing resistance against the acidic environment inside the host cells during the infection of the *M. tuberculosis* (Raynaud et al., 2002, Yang et al., 2011, Molle et al., 2006). The expression of OmpATb particularly induced at pH 4.5 and may facilitate the survival of *M. tuberculosis* at low pH conditions in the phagosomes by performing the acid-induced pore closing (Molle et al., 2006). Therefore, the analysis of the OmpATb structure was considered as crucial for the present study. The structure of OmpATb belongs to the class of α+β sandwich, arranged into mixed β - sheet which is formed from the parallel as well as antiparallel β strands (β1- β6) along with three α-helices (α1 - α 3) grouped on the equivalent side of the β – sheet. The OmpATb topology is divided into two β-sheet domains which were connected by α4 (Figure 3.19).

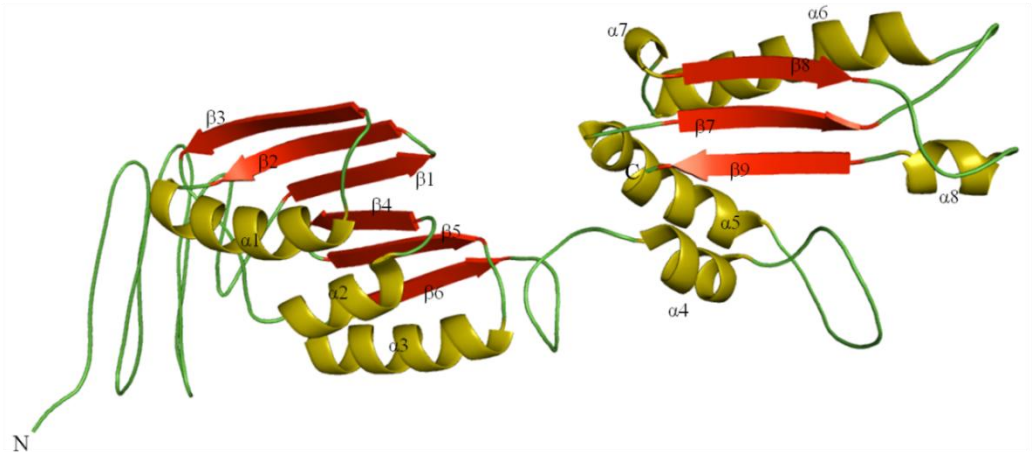


Figure 3.19: The structure of OmpATb showing characteristic bacterial OsmY and nodulation (BON) domain

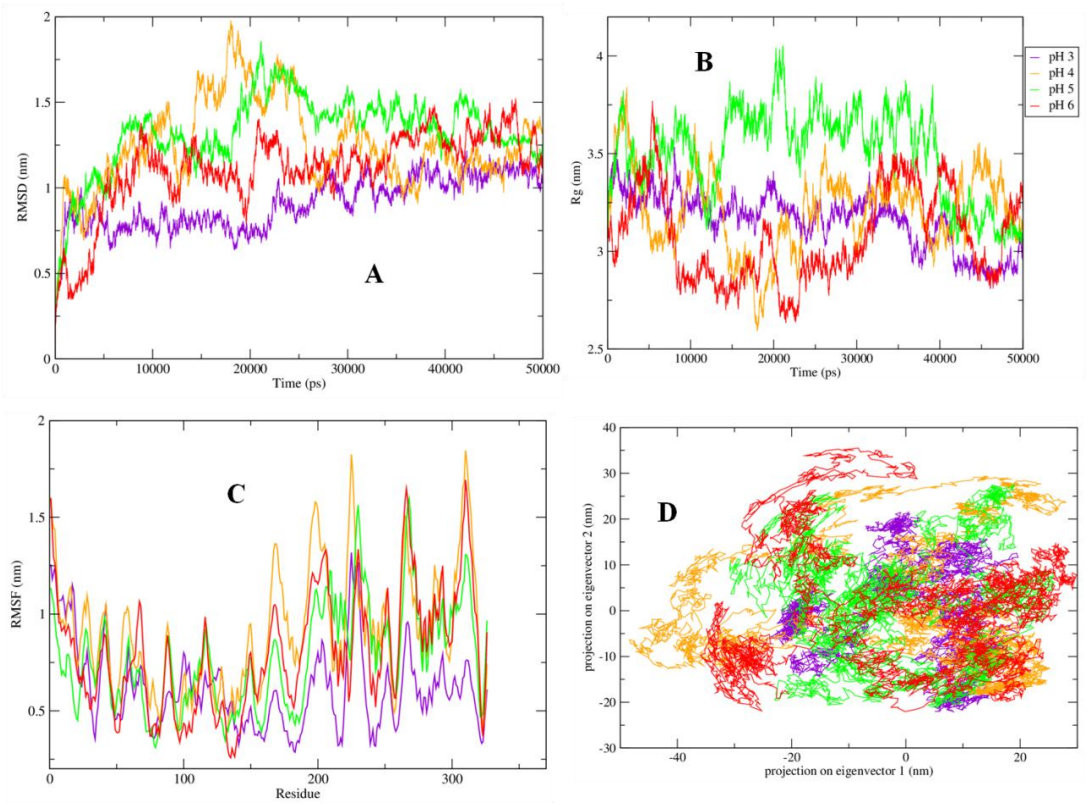


Figure 3.20: The variations in the conformational behavior of OmpATb is depicted in terms of changes in **(A)** RMSD values, **(B)** Rg values, **(C)** RMS fluctuations, and **(D)** the eigenvalues.

The study of OmpATb conformational behavior showed that the RMSD values in all the conditions were fluctuating in a comparable manner with least RMSD values were observed at pH 3.0 which is found in the range of 0.5 nm – 1 nm (Figure 3.20A). Similarly, comparatively higher fluctuations were observed in the Rg values with least fluctuations were observed at pH

3.0 with values ranging from 3 nm – 3.5 nm (Figure 3.20B). The constituent residues of OmpATb showed higher fluctuations and motions at every pH condition (Figure 3.20C and D) which are complemented by the observations obtained from free energy landscapes. This extensive analysis showed that the OmpATb can perform its characteristic functions even at the highly acidic conditions as low as pH 3.0.

3.3.2.5 Two-component regulatory systems (PhoP)

In pathogenic bacteria, the two-component regulatory systems such as PhoP/PhoQ, function as sensory as well as adaptive factors and are involved in the generation of the response against a variety of environmental changes (Perez et al., 2001). The associated PhoP protein is important for the infection of the *M. tuberculosis* (Ryndak et al., 2008). Consequently, the structure of PhoP was extensively analyzed at diverse acidic conditions. The PhoP assumes a dimerized topology in which N-terminal domain has $(\beta\alpha)_5$ fold while the C-terminal domain has winged helix-turn-helix fold. The PhoP structure dimerizes at α4-β5-α5 region (Figure 3.21).

Figure 3.21: The structure of PhoP showing characteristic dimerized topology

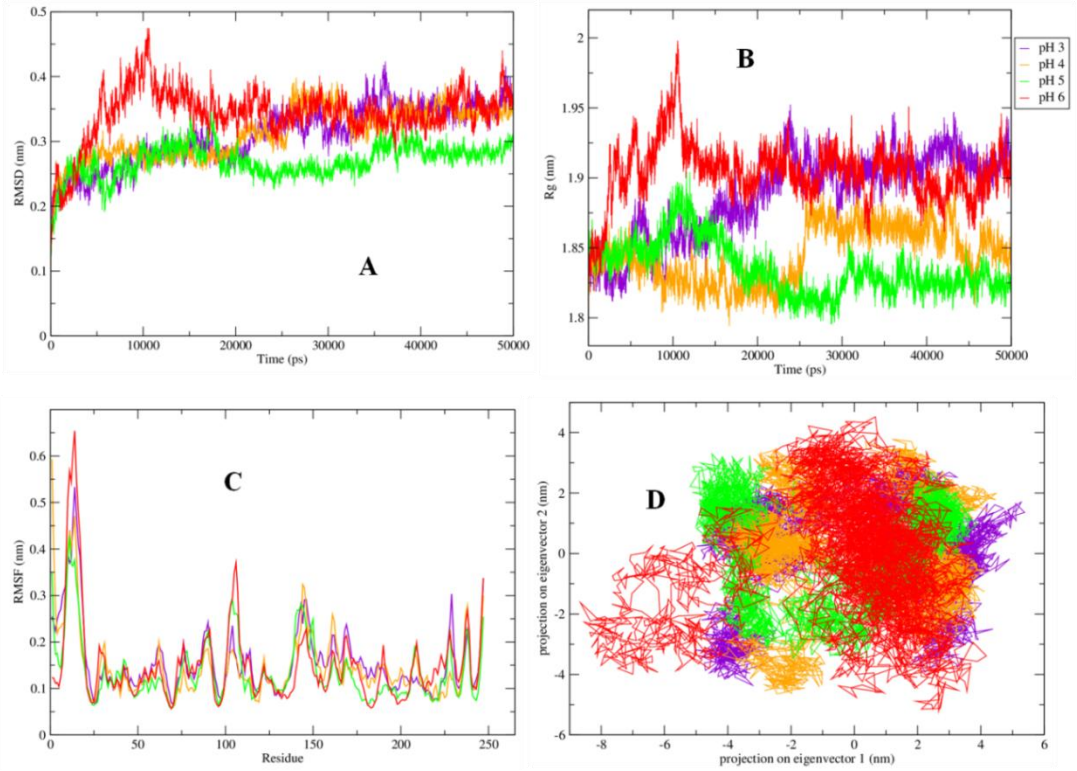


Figure 3.22: The observed structural changes in PhoP at different pH conditions are illustrated in the form of varied **(A)** RMSD values, **(B)** Rg values, **(C)** RMS fluctuations, and **(D)** the eigenvalues.

The structure of PhoP upon MD simulations showed lesser fluctuation in RMSD values at pH 5.0 with values are observed in the range of 0.2 – 0.3 nm, as compared to the other conditions in which RMSD values fall in between 0.3 – 0.4 nm (Figure 3.22A). Similarly, the Rg plots showed the presence of higher compactness in the structure of PhoP at pH 4.0 and 5.0 (Figure 3.22B), in which Rg values lie between 1.8 nm – 1.85 nm. While at other pH conditions, the Rg values fluctuating around 1.9 nm (Figure 3.22B). Furthermore, at pH 5.0 and 6.0, the comparable fluctuations were observed (Figure 3.22C) while least atomic motions were observed at pH 5.0 (Figure 3.22D). On the basis of obtained outcomes, it was inferred that PhoP maintains its structural integrity in acidic conditions below between pH 4.0.

3.3.2.6 Undecaprenyl pyrophosphate phosphatase (Rv2136c)

In *M. tuberculosis*, the Rv2136c gene was observed to be a homologue of *E. coli*'s undecaprenyl pyrophosphate phosphatase (Darby et al., 2011), which function as a lipid carrier and involved in the biosynthesis of carbohydrate intermediates which form the crucial component of bacterial cell envelope (Tatar et al., 2007). The Rv2136c is considered crucial for the development of the acid resistance in *M. tuberculosis* (Darby et al., 2011). In order to understand its molecular basis of resistance, the structure of Rv2136c was simulated for 50 ns each at different protonation states. The structure of Rv2136c assumes all α-helix topology which comprises 11 α-helices (Figure 3.23). The Rv2136c showed comparable fluctuation in the RMSD values at pH conditions ranges in between 0.3 nm – 0.5 nm, while higher fluctuation was observed after 30 ns in the case of pH 4.0 (Figure 3.24A), indicating that the protein is diverting from its native conformation and become less stable. Similar behavior was observed in the Rg plots, which showed that comparable compactness in the structure of Rv2136c was maintained at all the pH conditions with Rg values were fluctuating in between 1.85 nm – 1.95 nm, while at pH 5.0 the structure of Rv2136c is able to maintain slightly higher compactness (Figure 3.24B).

Figure 3.23: The structure of Rv2136c showing all α-helix topology

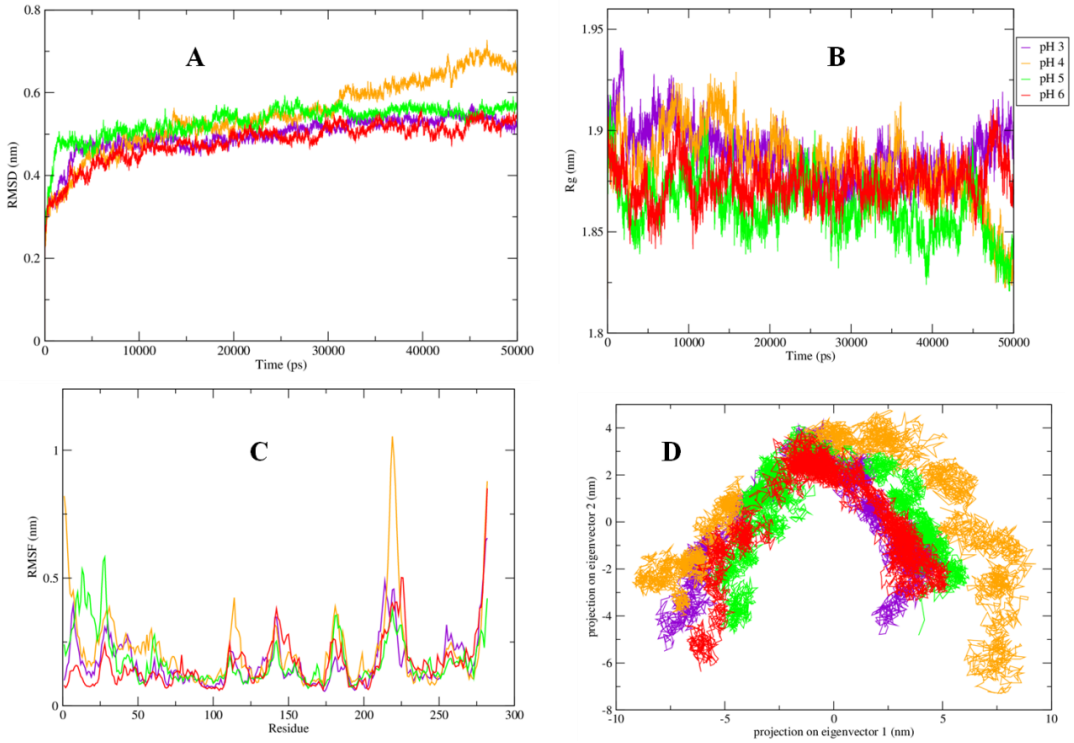


Figure 3.24: The constant pH-based MD simulations showed changes in the conformational behavior which are illustrated in the fluctuating values in **(A)** RMSD plot, **(B)** Rg plot, **(C)** RMS fluctuations plot, and **(D)** the eigenvector plots.

The highest fluctuations in the constituent residues were observed at pH 4.0 with residues 204 - 226 showed highest fluctuations as these residues forming the loop region (Figure 3.24C). These findings were complemented by the 2-D eigenvector plots which showed the presence of higher atomic motions at pH 4.0 as compared to other acidic conditions (Figure 3.24D). These findings showed that Rv2136c can function in diverse ranges of the acidic environment which is correlating with the information present in the literature (Darby et al., 2011).

3.3.2.7 Serine protease (Rv3671c)

The Rv3671c gene annotated as serine protease and is considered crucial for the survival of *M. tuberculosis* in the adverse environment of the phagosome (Biswas et al., 2010). Apart from its role in acid resistance, this protein is also involved in the development of resistance to oxidative stress (Biswas et al., 2010). The available literature on Rv3671c contains the

information about it's structural and biochemical studies which highlighted that its periplasmic domain is similar to serine protease of chymotrypsin family (Biswas et al., 2010). The structure of Rv3671c showed the presence of two β-barrel domains with surrounding 11 α-helices (Figure 3.25). After simulating the structure of Rv3671c, its dynamical behavior was analyzed in distinct explicit solvent conditions. These observations showed that the structure of Rv3671c is comparatively stable at pH 5.0 than other studied conditions. The RMSD values at pH 5.0 were fluctuating in the range of 0.4 nm – 0.6 nm (Figure 3.26A). Similar behavior was observed in pH 6.0 and pH 4.0, while the continuous increase in the RMSD values was observed at pH 3.0 (Figure 3.26A).

Furthermore, the Rg plots indicated that highest compactness in the structures was observed at pH 5.0 as compared to the other acidic environment (Figure 3.26B). Similarly, these findings were validated by RMSF and eigenvector plots which showed least fluctuations in the constituent residues and in the atomic motions at pH 5.0 (Figure 3.26C and D). This conformational analysis showed that the Rv3671c can survive in the diverse acidic environment, but structural preferences were observed at pH 4.0.

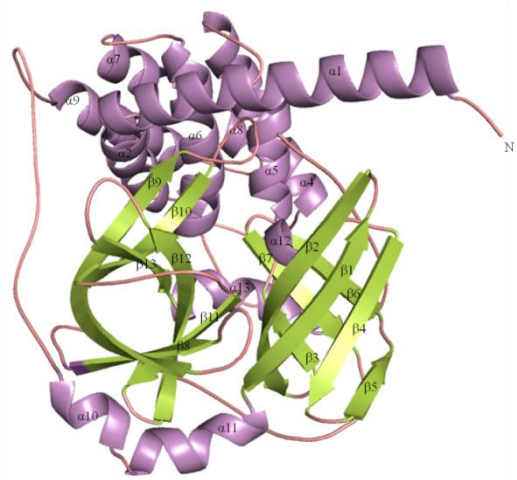


Figure 3.25: The structure of Rv3671c showing characteristic two β-barrel domains

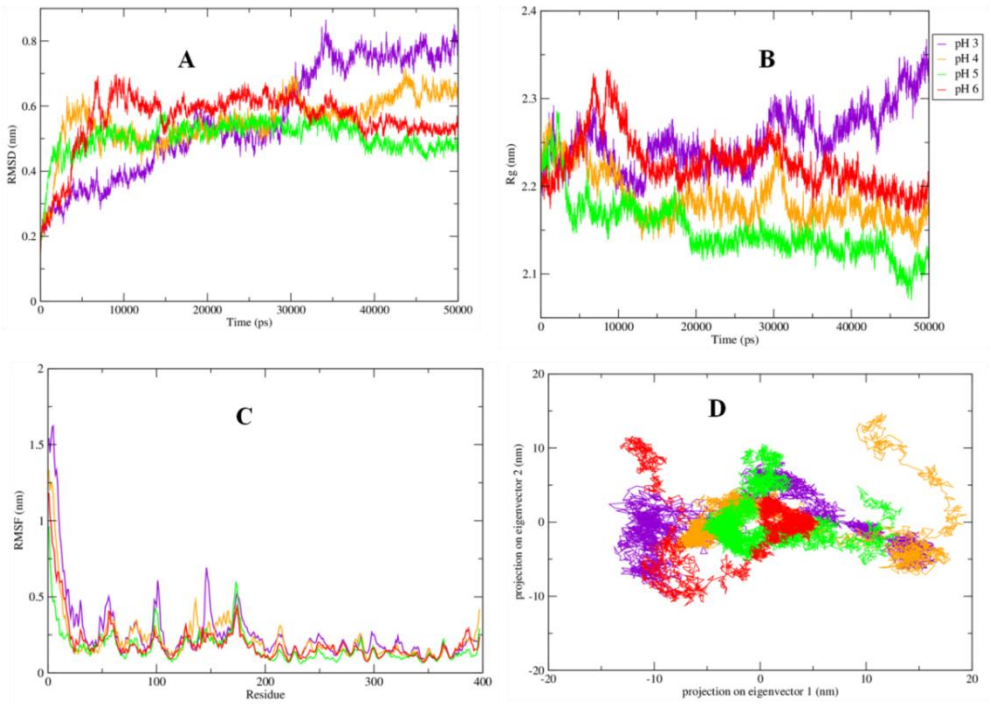


Figure 3.26: The variation observed in the structure of Rv3671c at a pH range of 3.0 – 6.0 are highlighted in the form of changes in **(A)** RMSD values, **(B)** Rg values, **(C)** RMS fluctuations, and **(D)** the eigenvalues.

3.3.3 Conclusions

The acid tolerance in *M. tuberculosis* is one of the primary causes for its survival inside the macrophage of the human host. On the basis of the literature search, seven proteins (explained in section 3.2) were found to be the primary cause of the acid tolerance. But the scarce information was available regarding their molecular mechanisms. Therefore, the constant-pH based MD simulations were performed for each protein and their behavior was observed in the mildly acidic to highly acidic environment. All the proteins were able to maintain their structure integrity in the acidic conditions, particularly lipF, PhoP, Rv2136c, and Rv3671c. The lipF and PhoP protein are showing highest compactness in the structure at pH 4.0 as compared to the other proteins and therefore, may be the primary causes of the pathogen survival and by targeting these proteins can facilitate the treatment of TB infected patients. The knowledge obtained from these conformational studies of proteins involved in resistance mechanisms were

utilized for the classification of the novel virulent protein in its genome, as explained in the following section:

3.4 Sequence-based function predictions

The extensive analyses of the virulent proteins involved in the pathogenic mechanisms of *M tuberculosis* provided an insight into their structural behaviors and this information was utilized in the classification of the additional virulent proteins. This study was aimed at the identification and analyses of the putative virulence factors in the genome of the *M tuberculosis*. Accordingly, the sequences of 1000 Hypothetical Proteins (HPs) present in *M. tuberculosis* were analyzed using diverse approaches of the similarity searches, sequence alignment, domain and functional annotations. The functions and the virulence characteristics of each HP were predicted and analyzed with these *in silico* methods to infer their characteristic molecular properties and the nature of evolution. The methodologies adopted for the functional annotations of the HPs are discussed in the consecutive sections.

3.4.1 Material and Methods

The different strategies for the sequence analysis are illustrated in Figure 3.27. The adopted methodology for the functional annotation of the HPs is divided into three phases. The primary phase involved the classification and sequence retrieval of the HPs from publically available databases (Figure 3.27), followed by the extensive analyses of the HPs' sequences using diverse *in silico* methods. On the basis of consensus of the generated output, the suitable functions were allocated for each protein. Each category of the computational technique is explained in the following sections:

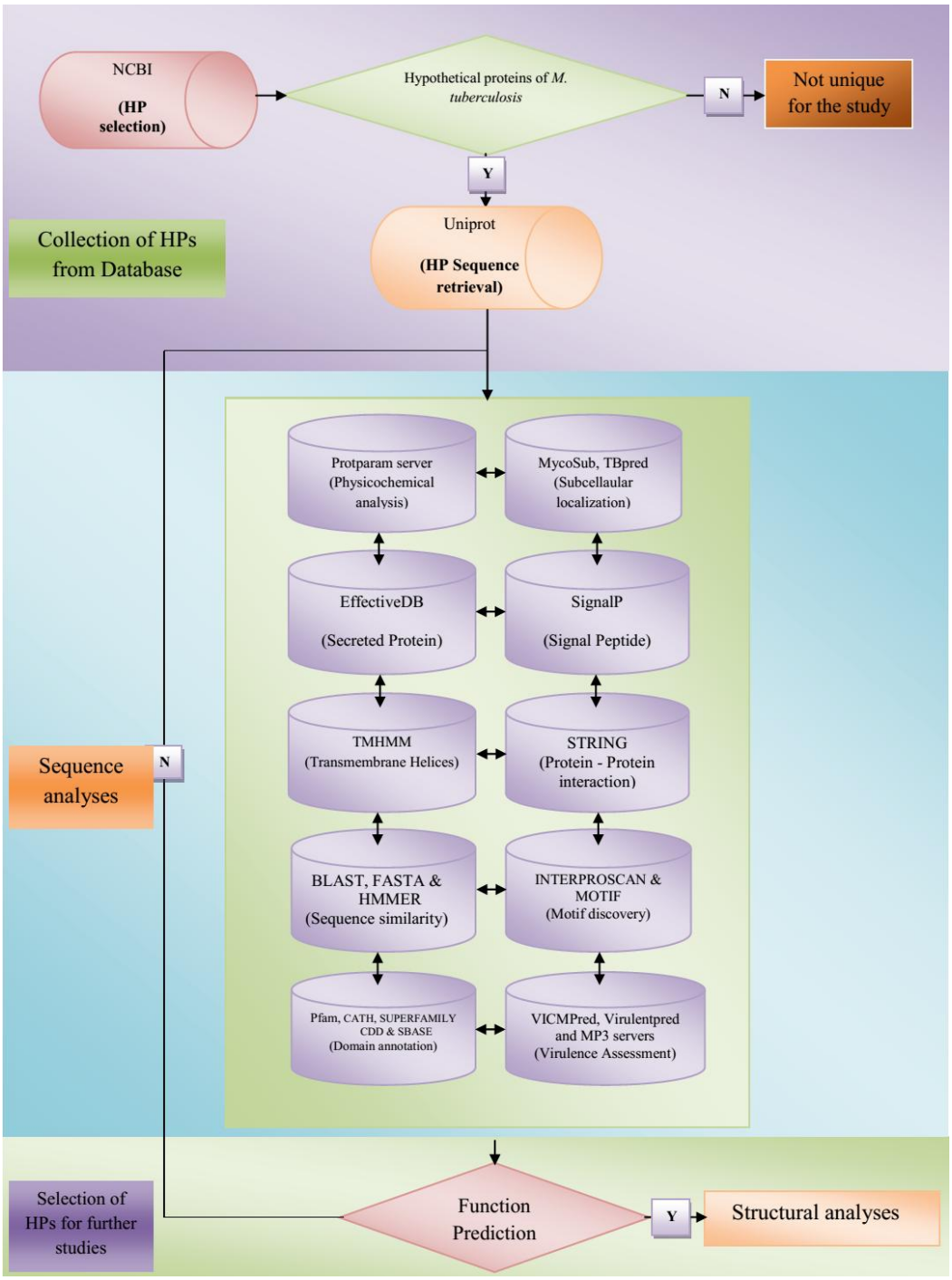


Figure 3.27: The computational workflow adopted for the function prediction of HPs

3.4.1.1 Physicochemical properties

The biophysical properties such as crystallization and thermodynamic stability of proteins can be inferred using the characterization of the physicochemical properties (Price et al., 2009).

The calculated physicochemical properties of proteins include determination of extinction coefficient, theoretical isoelectric point, molecular weight, the composition of amino acids, atomic organization, estimated half-life, grand average of hydropathicity index as well as the calculation of the indices related to instability and aliphatic characteristics (Price et al., 2009). The ExPASy ProtParam server (http://web.expasy.org/protparam/) was utilized for the calculation of the sequence based physicochemical properties.

3.4.1.2 Subcellular localization

The characteristic roles played by any protein in the biological systems are collectively known as their "functions" (Nair and Rost, 2008). As described in Chapter 1, among three aspects of protein functions, the determination of their subcellular localization has been a target for intensive research (Nair and Rost, 2008). The co-localized proteins in the cellular compartments are involved in a mutual physiological process (Nair and Rost, 2008). In addition, these analyses are crucial in the process of drug discovery, as proteins can be classified as drugs or vaccine targets on the basis of their localization in the subcellular regions (Mohd et al., 2016). Therefore, a variety of methods were utilized for the prediction of the subcellular localization of the uncharacterized HPs present in the genome of *M. tuberculosis*.

The MycoSub server predicted the localization of the HPs in the cells of *M. tuberculosis* with an accuracy of 89.71% (Zhu et al., 2015). It uses support vector machine (SVM) based algorithms, trained on 219 tri-peptide features obtained from 272 non-redundant proteins of *M. tuberculosis* (Zhu et al., 2015). Similarly, TBpred utilizes the training set of 852 proteins from the genus Mycobacterium to predict the subcellular localization of the mycobacterial proteins and achieve an accuracy of 86.8% (Rashid et al., 2007). Furthermore, in bacteria, the protein

signal targeting is generally used to transport proteins to the secretory systems as well as in the translocation through the cytoplasmic membrane (Filloux, 2010).

The transferred proteins then continue to anchor onto the bacterial cell surface before being secreted into the extracellular space (Filloux, 2010). Therefore, signal peptides in the bacterial proteins effectively control the entry of the proteins to the non-classical bacterial secretory pathways (Filloux, 2010). As a result, the SignalP a neural network based algorithm was used to predict the presence of a signal peptide in the protein sequences and their cleavage sites (Petersen et al., 2011). These predictions were complimented by the EffectiveDB, which predict the involvement of the proteins in Type III, IV, VI and VII secretory systems present in the pathogenic bacteria (Eichinger et al., 2016). Moreover, the transmembrane helices in the bacterial HPs were classified using TMHMM server (Krogh et al., 2001).

3.4.1.3 Functional annotations

The sequence-based functional analyses initially involved the similarity searches for the annotations of the uncharacterized proteins present in the genomes of pathogenic bacteria (Mohd et al., 2016). Accordingly, BLAST was used which performed the similarity searches for the sequences of HPs. BLAST is the heuristic method which utilizes multi-step algorithms for comparing the sequences (Altschul et al., 1990). In the primary or seeding stage, BLAST identifies the short matches in the protein sequences and then, identifies the strings between the concerned sequences and generates the alignments by utilizing the assembly of these strings (Altschul et al., 1990). The protein sequence alignments with brief query coverage (< 50%) and low sequence identity (< 20%) were considered as the remote homologs and were rejected. On the other hand, the protein sequences with larger query coverage and high sequence similarities (> 40%) were predicted as a reliable hit (Mohd et al., 2016). The accuracy of the similarity

searches was enhanced by utilizing position-specific score matrix generated by the PSI-BLAST for database searching (Altschul et al., 1997).

Similarly, other heuristic methods such as FASTA performed more precise K-tuple database search for the prediction of the functionalities associated with the HPs. In FASTA analysis, the speed of the database searching reduces as it compares the words throughout the protein sequences (Pearson and Lipman, 1988). Moreover, the reliability of the functional predictions was increased by the usage of HMM-based algorithms such as HMMER (Finn et al., 2011), which performed the pair-wise comparisons of the generated profile HMMs.

3.4.1.4 Domain annotations

The conserved domains in protein sequences are distinct units of molecular evolution and crucial for the characteristic molecular functions such as catalysis, transport, binding etc. (Marchler-Bauer et al., 2005). The annotations of these conserved domains in the sequences of uncharacterized HPs may provide evidence regarding their biological and molecular functionalities (Marchler-Bauer et al., 2005). The dcGO is an online database aimed at the characterization of conserved domains in the functionally uncharacterized sequences of HPs, by correlating the identified domain with the ontological terms at the level of families and superfamilies (Fang and Gough, 2013). Similarly, the Conserved Domain Database or CDD (Marchler-Bauer et al., 2015) is a repository containing information about the conserved domain signatures present in different functional classes of the proteins used for the identification of the domains in the HPs. Further analyses included the usage of Protein Analysis THrough Evolutionary Relationships (PANTHER) database, which is a manually administered database containing a significant compilation of protein families which are subdivided into functionally correlated protein subfamilies (Mi et al., 2016). These sub-families signifies the divergence that

exists within a protein family and accurately performed the function predictions by associating with the biological pathways, ontology terms and the amino acids necessary for the specificity of the protein functionalities (Mi et al., 2016). In order to classify new proteins, the sequences are compared with the HMMs generated for each family as well as sub-family (Mi et al., 2016). The version 10 of PANTHER was used that contains around 12, 000 protein families (Mi et al., 2016).

Furthermore, on the basis of 235,000 domain structures of proteins coupled with 25 million predictions of domains in the CATH database, the functional domains in the sequences of the HPs were identified (Sillitoe et al., 2015). Likewise, SUPERFAMILY database provides information about the protein structure, functions as well as about their evolutionary patterns. The SCOP domain generated HMM models at the superfamily level, form the basis for the structure-based functional annotations of the HPs (Wilson et al., 2009). Moreover, the Pfam is a collection of protein families, which includes a group of protein regions with significant sequence similarities detected by the HMMER utility present in the database (Finn et al., 2016). The current release contains around 16, 295 protein families along with the information about 559 clans (Finn et al., 2016). The Simple Modular Architecture Research Tool (SMART) is a manually curated database aimed at the identification and annotations of domain architectures (Letunic et al., 2015). The SMART version 7.0 contains around 1200 protein domain models and more than a million features which are available because of its synchronization with other specialized databases such as STRING, UniProt and Ensembl (Letunic et al., 2015). The "genomic mode" of SMART database annotates the proteins present in the genome of the organisms; currently contain the genomic information about 2031 species (Letunic et al., 2015).

The SVM-based predictions of SBASE were further utilized for the functional classification of the HPs, which performed database searching for the detection of similarities in

the domain composition (Vlahovicek et al., 2005). The current release of SBASE database contains 972,397 segments of protein sequences which were annotated on the basis of their attributes such as function, structure, cellular topology and ligand-binding and grouped into 8,547 domain clusters (Vlahovicek et al., 2005). It also contains 2,526 less annotated clusters which include 169,916 fragments of domain sequences (Vlahovicek et al., 2005).

In addition to the identification of the conserved domains in the protein sequences, the annotations of conserved motifs segments in the proteins were explored for the prediction of the protein functions. The InterProScan 5 classify the similar motifs associated with diverse protein functions in InterPro group of databases (Quevillon et al., 2005). Likewise, the MOTIF (http://www.genomC.jp/tools/motif/) identified and analyzed the functional motif signatures in the uncharacterized protein sequences of HPs.

3.4.1.5 Virulence predictions

The knowledge generated by analyzing the diversity of virulence factors present in *M. tuberculosis* form the basis for the identification of new virulence factors. Accordingly, the *in silico* methods were used for the classification of novel virulence factors in the group of 1000 HPs. The virulence factors present in the bacteria are critical for its infection (Baron and Coombes, 2007) and significant drug targets in drug discovery. The VirulentPred (Garg and Gupta, 2008), MP3 server (Gupta et al., 2014), and VICMpred (Saha and Raghava, 2006) servers were used for the classification of virulent proteins on the basis of their amino acid compositions. The VICMpred as well as VirulentPred are based on SVM-based algorithms and assessed the virulence characteristic of the proteins with a precision of 70.75% and 81.8% respectively. While the MP3 server utilizes a hybrid approach formulated by the integration of HMM as well as SVM-based methods, achieved an accuracy of 96% during the predictions (Gupta et al., 2014).

3.4.1.6 Standardization of *in silico* protocol

In order to annotate the virulent HPs in *M. tuberculosis* successfully, the utilized computational protocol for the annotation of the HPs was standardized using the genome of model pathogens such as *Neisseria meningitidis* and *Mycoplasma pneumoniae* (Shahbaaz et al., 2015b, Shahbaaz et al., 2015a). A systematic workflow was adopted regarding the functional annotation of the HPs (Shahbaaz et al., 2015b, Shahbaaz et al., 2015a). Furthermore, the accuracy of the prediction was evaluated using statistical technique of Receiver operating characteristic (ROC). The details of each model pathogens used and in the standardization of the protocol are explained below:

3.4.1.6.1 Model pathogens

N. meningitidis is a distinct Gram-negative bacteria causes bacterial meningitis (Jafri et al., 2013). The sequenced genome of *N. meningitidis* MC58 contains 2,272,351 bps and organized into 2158 coding regions which resulted in about 1953 proteins. The biological functions were allocated to 53.7 % of the coding regions while rest 35% was classified as "hypothetical" (Tettelin et al., 2000). While, *M. pneumoniae* is among the smallest self-replicating bacteria belong to the Mycoplasmataceae family and resembles Gram-positive bacteria in genomic composition (Weisburg et al., 1989). In human, it is responsible for atypical pneumonia which is an acute respiratory tract infection and associated neurological, hemolytic, cardiac and hepatic manifestations (Razin et al., 1998). The formulated protocol was able to allocate the functionalities in more than 60% of the uncharacterized HPs in the genome of each model pathogen (Shahbaaz et al., 2015b, Shahbaaz et al., 2015a).

After functional annotations of *N. meningitidis* genome, around 368 HPs showed homology with the proteins of known functions. The respective proteins were further classified

into 41 diverse functional categories of the proteins by utilizing the information present in the literature (Shahbaaz et al., 2015b). Among the characterized HPs, the 39 transferase, 17 oxidoreductase, 14 peptidase, four sulfatase, 18 hydrolase, six translocase, 11 restriction enzymes, six hydrogenase, nine synthase, eight nucleotidase, six lyase, 79 transport-associated proteins, 20 virulence-related proteins, 11 bacteriophage related proteins, 15 binding proteins, 10 immunity proteins, and many more (Shahbaaz et al., 2015b). Amongst 368 functionally characterized HPs, 18 proteins were categorized as the virulent (Shahbaaz et al., 2015b). The HP Q7DDQ9, as well as Q9JYT4, was classified with highest virulence scores and selected for further structural analyses (Shahbaaz et al., 2015b).

Similarly, in *M. pneumoniae* among 204 HPs, the functionalities of 83 HPs were allocated successfully (Shahbaaz et al., 2015a). On the basis of the obtained information, these annotated HPs were classified into various functional classes such as 23 enzymes, 27 lipoproteins, 12 transport proteins, 10 binding proteins and other proteins belonging to the cellular process such as transcription, translation, and replication (Shahbaaz et al., 2015a). On the basis of the assessments performed on the predictions, six HPs were categorized as virulence factors amongst the group of 83 functionally annotated HPs (Shahbaaz et al., 2015a). HPs H0PQA7 and H0PQI2 annotated with highest virulence scores and therefore, selected for further structural analyses (Shahbaaz et al., 2015a).

3.4.1.6.2 Accuracy assessment

The genomes used in standardization were compared and on the basis of the consensus, the protocol regarding the functional annotation of the HPs was standardized. After the functional analyses, the new functional candidates were identified along with the novel virulence factors. Furthermore, the predictive accuracy acquired during predictions was evaluated by

utilizing the statistical approach of ROC (Shahbaaz et al., 2015a). In this statistical assessment, the computational methods depicted in Figure 3.27 were utilized for the annotation of proteins with "known function" (Shahbaaz et al., 2015a). Then, the predicted functions were compared with the experimentally derived functionalities deposited in the database (Shahbaaz et al., 2015a). Afterward, the suitable scoring function was used which correspond to the confidence of the prediction (Shahbaaz et al., 2015a). And on the basis of these scores, the assessment was performed in the form of sensitivity, specificity and ROC curve area (Shahbaaz et al., 2015a). The accuracy achieved by the protocol was computed to be around 96% and average area covered by the ROC curve was calculated to be about 0.704, which designated the reliability of the adopted pipeline (Shahbaaz et al., 2015a).

3.4.2 Results and Discussions

Due to the limitations of the current therapy against the continuous evolving resistance mechanisms in *M. tuberculosis* (Tsara et al., 2009), there is a need for the discovery of new drug targets. The genome of *M. tuberculosis* contains around 1000 "Hypothetical Proteins (HPs)" (http://www.ncbi.nlm.nih.gov/genome/166). The sequences of these HPs were obtained from the Uniprot database (http://www.uniprot.org/) and extensively analyzed using a variety of Bioinformatics tools.

In the primary phase of functional annotation, the physicochemical parameters such as instability index (Guruprasad et al., 1990), isoelectric point, molecular weight, extinction coefficient (Gill and von Hippel, 1989), grand average of hydropathicity (Kyte and Doolittle, 1982) and aliphatic index (Ikai, 1980) were calculated using Expasy's protparam server. After analyzing these parameters, around 459 HPs were classified as "Unstable", while rests of the proteins were predicted to be "Stable". These calculated physicochemical properties were used to infer the characteristic of each HP.

The preliminary function predictions involved the sub-cellular localization of the HPs. Among 1000 HPs, around 637 proteins were predicted to be localized in the cytoplasm of the *M. tuberculosis*'s cells, while 273 were annotated as Integral membrane protein as well as 43 HPs may be attached to the membrane by lipid anchor and 12 proteins were classified into "Secreted" category. In order to identify the involvement of the HPs in the processes of signal transduction, the SignalP server was used which identified 39 HPs may be involved in the metabolic signaling. Moreover, around 133 HPs were found to be involved in non-classical secretory pathways observed in the pathogenic bacteria. Likewise, 115 were classified as integral membrane proteins among 1000 HPs on the basis of predicted transmembrane helices and may be involved in the transport mechanisms.

The further analyses involved identification of the functionalities associated with the 1000 HPs by performing the sequence similarities with the proteins of known functions, by identification of the functionally active motifs and domains in the sequences of the HPs and by classifying the HPs into the probable functional families using the diverse Bioinformatics tools. After performing the extensive analyses using the sequences of the HPs, the functionalities of 662 proteins were predicted among 1000 studied HPs (Table C.1 – C.3). Among functionally annotated HPs, 483 were classified as "Enzymes" (Table C.1), 141 may be involved in the diverse "Cellular Processes" (Table C.2), and 38 proteins may function as "Transporters and Carriers" (Table C.1). The rest of 307 HPs did not show predicted outcomes in less than three of the utilized functional annotation methods and categorized as "less precisely predicted" (Table C.1) due to unavailability of the suitable functional homologs in the current updated protein databases. Different classified protein functional categories are discussed below in details:

3.4.2.1 Enzymes

The enzymes are the proteins which accelerate the metabolic reactions in an organism and are crucial for the survival of the pathogenic microorganisms and drugs are mainly designed to inhibit these proteins (Nelson et al., 2013). Therefore, the knowledge about their functionalities is crucial in order to design new drugs (Nelson et al., 2013). The identification of the new enzymatic drug targets may facilitate the treatment of drug-resistant TB. Around 17 different categories of enzymes were identified among 1000 studied HPs (Figure 3.28). The major categories are discussed below in details:

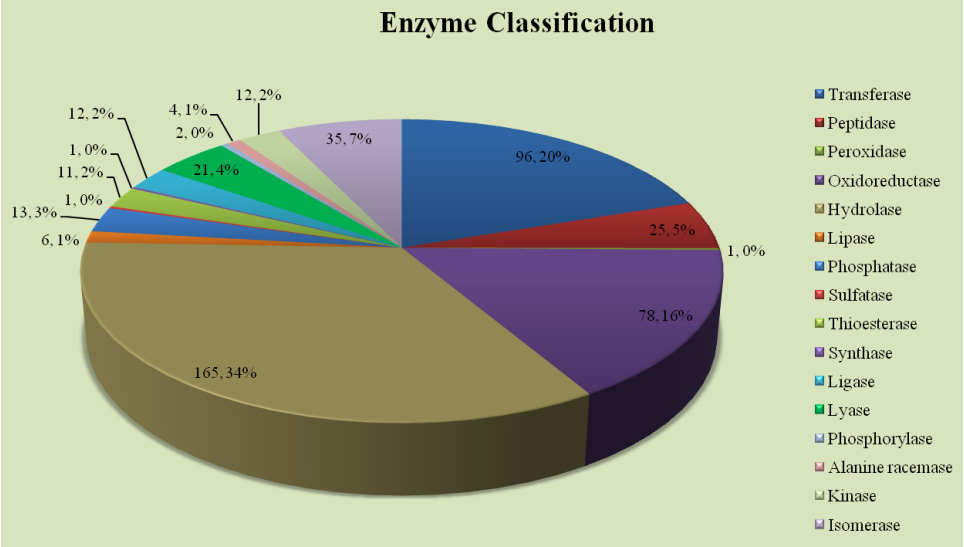


Figure 3.28: Different functional categories classified in 483 annotated enzymes

3.4.2.1.1 Transferase

This class of the enzyme plays significant roles in maintaining the cellular integrity of the *M. tuberculosis* (Leblanc et al., 2012) and involved in the bacterial pathogenesis by regulating the synthesis of crucial virulence factors (Heithoff et al., 1999). Transferase enzymes are also vital for biosynthesis of lipoprotein, which plays a significant role in the pathogenesis (Okugawa et al.). The mycobacterial 4'-Phosphopantetheinyl transferase activates the cell envelope forming lipids and considered as an essential drug target (Leblanc et al., 2012). Among 483 classified

enzymes, 96 HPs were predicted to be "Transferase" (Table C.1). The 16 HPs showed functionalities similar to SAM-dependent methyltransferases and may be involved in the synthesis of short-chain fatty acids which form the crucial part of *M. tuberculosis* cell envelope (Meena et al., 2013, Boissier et al., 2006). Likewise, five HPs were predicted to be glycosyltransferase and possibly involved in the synthesis of conjugated macromolecules such as arabinogalactan, which are essential for the synthesis of the bacterial cell wall (Berg et al., 2007).

Furthermore, the HP P9WKT1, HP O06830, HP O06625, and HP I6XFY8 were predicted to be acyltransferase and can be considered as important for the bacterial pathogenesis (Shi and Ehrt, 2006). Similarly, the HP O53699, HP P9WLG1, and HP I6YCC4 were annotated as sulfotransferase, probably involved in the biosynthesis of sulfated trehalose glycolipids, the characteristic biomolecules found in the cell wall of *M. tuberculosis*. The HP I6Y6S3, HP O53409, HP P95284, and HP I6X831 showed the presence of glutamine amidotransferase like activities and may catalyze the synthesis of NAD(+) by utilizing glutamine and ammonia (Bellinzoni et al., 2005). Moreover, the HP P9WHK1 and HP O05887 were predicted to be phosphoribosyltransferase, a protein of purine salvage pathway and can be considered as potential drug targets for acyclic nucleoside phosphonates (Eng et al., 2015). These observations indicated that these predicted transferase enzymes can be utilized in the designing novel therapeutic agents against the infection of *M. tuberculosis*.

3.4.2.1.2 Oxidoreductase

The oxidoreductases are the enzymes that catalyze varieties of the redox reactions in the biochemical pathways which are crucial for the survival of the organism and form the significant category of the drug targets (Yano et al., 2014). The enzymes such as Type II NADH-quinone oxidoreductase is important for the survival of the *M. tuberculosis* because they form the part of

bacterial oxidative phosphorylation system and catalyzes the transfer of the electrons between the NADH and quinone pool (Yano et al., 2014). In the set of characterized HPs, 78 proteins were predicted to show similarities with different classes of oxidoreductases (Table C.1). The 10 HPs in the set of predicted oxidoreductases showed high similarities to pyridoxamine 5'-phosphate oxidase. The *in vitro* studies identified that the gene encoding pyridoxamine 5'-phosphate oxidase in *M. tuberculosis* showed up-regulated expression when exposed to the stress conditions and therefore may be involved in its survival in the macrophages (Mashalidis et al., 2011). Furthermore, the HP P9WFP7 and HP P9WFP5 showed homologies with Fe-S cluster assembly proteins SufB and SufD respectively. The homologs of SufB protein may be involved in the repair of SUF machinery of *M. tuberculosis* and are vital for the survival of the pathogen (Huet et al., 2005). The HP O06216, HP P95086, HP I6XG43, and HP I6X7D4 were predicted as bacterial luciferases and can be used as marker proteins to analyze the changes in the gene expression in *M. tuberculosis* because of they're inherent destabilize nature (Roberts et al., 2005).

Furthermore, the nitro-reductases in *M. tuberculosis* are involved in the catalysis of nitroimidazoles reduction resulted in the release of the reactive nitrogen moieties and may be involved in its virulence (Cellitti et al., 2012). Therefore, the HP O06389, HP P95233, and HP P9WL07 were predicted to be nitroreductases and may be crucial for the survival of *M. tuberculosis* in the limiting environment of macrophages (Purkayastha et al., 2002). Among functionally annotated oxidoreductases, six HPs were classified in the mixed category of "Glyoxalase or Bleomycin resistance protein or Dihydroxybiphenyl dioxygenase". The glyoxalase enzyme plays a significant role in the pathogenic bacteria by catalyzing the detoxification of methylglyoxal, a α-oxoaldehyde with highly toxic byproduct which performs the modification of the biomolecules and can lead to cell death (Chakraborty et al., 2015). This

enzyme also enables the survival of the pathogenic bacteria in the cells of host immune system (Zhang et al., 2016). The glyoxalase enzyme forms the primary step of the detoxification by converting hemithioacetal to S-lactoylglutathione in the glyoxalase cycle. Therefore, these HPs with predicted glyoxalase activities can be used as drug targets for the development of the novel therapeutic agents against *M. tuberculosis* (Edagwa et al., 2013). Moreover, 10 HPs showed the presence of diverse dehydrogenase activities and may be involved in the reductive amination of glyoxylate (Giffin et al., 2012), amino acid biosynthesis (Xu and Grant, 2014) and several other significant functions in *M. tuberculosis* (Giffin et al., 2016). These observations indicated the importance of the annotated HPs in the process of drug discovery against the infection of TB.

3.4.2.1.3 Hydrolase

The hydrolases are the group of enzymes crucial for the pathological pathways present in the virulent strains of the bacteria (Li et al., 2013). Around 165 HPs were classified in the "hydrolase" category (Table C.1). A variety of hydrolases such as β-lactamase enable the pathogenic bacteria to develop resistance against the β-Lactam antibiotics (Poole, 2004). Among classified hydrolases, 14 HPs showed the similarities with diverse metallo-beta-lactamases, therefore, may be involved in the development of drug resistance in *M. tuberculosis*. Likewise, the P-loop containing nucleoside triphosphate hydrolases such as ATPase, are critical for the survival of *M. tuberculosis* in the limiting environment of macrophages (Novoa-Aponte and Soto Ospina, 2014). The membrane bound ATPase in pathogenic bacteria facilitate the catalysis of proteins folding and degradation, DNA repair, replication initiation as well as the transportation of metabolites (Chene, 2002). The heavy metal pumps such as CtpC, CtpG as well as the CtpV form the part of *M. tuberculosis* defense mechanism in the host's phagocytic cells (Botella et al., 2011). Therefore, the 17 HPs characterized as P-loop containing nucleotide triphosphate

hydrolases and 11 ATPase can be utilized as the potential drug targets in the development of anti-tubercular agents (Chene, 2002, Yatime et al., 2009).

Furthermore, 10 HNH nuclease and 18 endonucleases were characterized in the set of 1000 HPs may be involved in the pathogenesis of *M. tuberculosis* by regulating the processes of biofilm formation, nutrient uptake, and degradation of neutrophilic DNA (Dang et al., 2016). The HP I6X9N8, HP O53874, HP O50435, HP O5046, HP P95267, and I6XHV9 were annotated as "helicase", an enzyme involved in DNA repair as well as in a variety of metabolic pathways (Williams et al., 2011). Similarly, the HP I6XHX8 and HP O69700 were predicted to be nucleoside triphosphate pyrophosphohydrolase, may possibly play a role in the development of resistance against oxidative stresses in *M. tuberculosis* (Lu et al., 2010).

3.4.2.1.4 Lipase

In *M. tuberculosis*, this category of the enzyme facilitates the catalytic breakdown of triglycerides by performing the hydrolysis of the carboxyl ester bonds in unsaturated fatty acids, the primary source of energy during starvation inside the cellular granulomas (Lin et al., 2016). There were seven HPs predicted to be lipase (Table C.1). The HP O53410 and HP I6YB49 were annotated as the Patatin-like phospholipase and may act as virulence factors by inducing cytotoxicity inside the macrophages as well as regulate the cellular necrosis (Assis et al., 2014). Likewise, the HP P71725 and HP O05805 were classified to be lysophospholipase and probably involved in the breakdown of membrane lipids present in the host's cells (Cotes et al., 2007). These observations indicated that these predicted lipases can be utilized as the biomarkers in the early diagnosis of the TB and also serve as drug targets in the process of designing anti-tubercular therapeutics (Brust et al., 2011, West et al., 2011).

3.4.2.1.5 Phosphatase

There were 12 proteins in the group of functionally annotated HPs which were categorized as "phosphatase" (Table C.1) and can act as a biomarker in the differentiation of tubercular pleural effusion (Jadhav et al., 2009). Furthermore, these phosphatases facilitate the infection in *M. tuberculosis* by modulation of the metabolic signaling pathways present in the hosts' cells (Grundner et al., 2005) as well as the activities of the macrophages (Saleh and Belisle, 2000). The HP Q50699, HP P9WL81, and HP I6X827 were predicted to be purple acid phosphatase, can take part in the pathogenicity of *M. tuberculosis* by causing the modulation of the phagosomal activities (Saleh and Belisle, 2000). Likewise, the HP O06240 and HP P9WGF9 may belong to the histidine phosphatase superfamily, in which the centered histidine presents in the conserved catalytic region undergo phosphorylation during the course of the reaction (Rigden, 2008). These observations showed the functionally diverse phosphatases may be utilized for the further studies regarding the formation of novel therapeutic agents (Coker et al., 2013).

3.4.2.1.6 Thioesterase

The extensive analyses showed that thioesterases are crucial for the survival of *M. tuberculosis* inside the macrophages (Wang et al., 2007a). Therefore, the 11 predicted thioesterases among the functionally annotated HPs are also significant for the survival of the bacteria. Among this group, the HP P96817, HP O53751, HP P9WKT9, HP O53917, HP O07408, and HP P9WLN5 showed similarities to the acyl-CoA thioesterases (Table C.1). These functionally annotated HPs may be involved in the intracellular lipid metabolism by catalyzing the hydrolytic breakdown of the acyl-CoA into fatty acid and coenzyme A (Hunt and Alexson, 2002). As the lipid metabolism is crucial for the biogenesis of the bacterial membrane, therefore,

inhibition of these predicted thioesterases can be utilized in the treatment of the TB-infected patients (Parker et al., 2009).

3.4.2.1.7 Ligase

The *M. tuberculosis* proteins such as D-alanine ligase, perform the catalytic conjugation of two D-alanine to form a dipeptide, was utilized as the target for the D-cycloserine drug in order to cure the infection of the TB (Bruning et al., 2011). There were 12 HPs which may function as ligase in the pathogenic bacteria (Table C.1). The HP P96873 showed similarities to the acetyl-CoA synthetases and may be involved in a variety of catabolic and anabolic pathways as well as catalyze the conversion of acetyl-CoA from the acetate (Noy et al., 2014). Likewise, HP P9WPQ1 predicted to be an acetyl-CoA carboxylase, which is considered to be significant for the biosynthesis of the mycolic acid by catalyzing the carboxylation of the acetyl-CoA and resulted in the formation of malonyl-CoA (Reddy et al., 2014, Rainwater and Kolattukudy, 1982). The inhibition of the acetyl-CoA carboxylase present in *M. tuberculosis* may be helpful in the prevention of the severe TB infection (Reddy et al., 2014). Furthermore, HP P71733 and HP I6YCS6 were annotated as glutamate-cysteine ligase and may be essential for the homeostasis of the *M. tuberculosis* (Harth et al., 2005). Therefore, the analyses of these characterized ligases in the set of HPs are necessary for the designing of better therapeutic against the *M. tuberculosis*.

3.4.2.1.8 Lyase

The lyase group of enzymes is crucial for the virulence as well as the existence of pathogenic bacteria as they catalyze the formation of necessary nutrients and modify the host's local environment in order to facilitate the growth of bacteria (Bjornson, 1984). The cystathionine β-lyase which is involved in the catabolism and anabolism of cystathionine causes the separation of sulfur from cysteine, which alters the biosynthesis of methionine and therefore,

involved in the bacterial virulence (Ejim et al., 2004). Likewise, the isocitrate lyase which is crucial for the glyoxylate pathway, in which it catalyzes the cleavage of isocitrate in order to produce glyoxylate as well as succinate and utilized as the drug target for curing a variety of bacterial infections (Britton et al., 2001). We have successfully classified 21 lyases in the group of 483 enzymatic HPs. The carboxymuconolactone decarboxylase similar features were observed in five HPs (Table C.1) which may be involved in the anti-oxidative response in the pathogenic bacteria (Chen et al., 2015). The HP O06571 and HP O06572 were annotated as adenylate cyclase, may be involved in the virulence of the *M. tuberculosis* by producing cyclic AMP that alters the metabolic signaling pathways present in the host (Agarwal et al., 2009). These observations indicated the essentiality of these predicted lyase enzymes in bacterial virulence.

3.4.2.1.9 Isomerase

The mycothiol biosynthesized in *M. tuberculosis*, is necessary for its protection against the antibiotics and oxidative stresses (Wang et al., 2007b). It is produced through a multi-step process and requires four enzymatic mechanisms (Wang et al., 2007b). Therefore, the HP P9WM91, HP O07256, HP P9WKS3, HP P71985, and HP O53480 were annotated as mycothiol-dependent maleylpyruvate isomerase (Table C.1) may be involved in the biosynthesis of the mycothiol (Wang et al., 2007b). In addition, among 35 predicted isomerases, HP O07166 was classified to be a pseudouridine synthase and may catalyze the conversion of uridine to pseudouridine (Chaudhuri et al., 2004). These analyses highlighted the importance of predicted isomerases in the drug design.

3.4.2.2 Cellular processes and transport proteins

In the set of annotated protein, around 141 HPs may be involved in the varieties of the cellular processes, while 38 proteins may be involved in the transportation mechanisms present

in *M. tuberculosis* (Table C.2 and C.3). The proteins in the category of the cellular processes may take part in the biochemical mechanisms such as metal binding, signal transduction pathways, gene expression, DNA replication and other mechanisms (Figure 3.29).

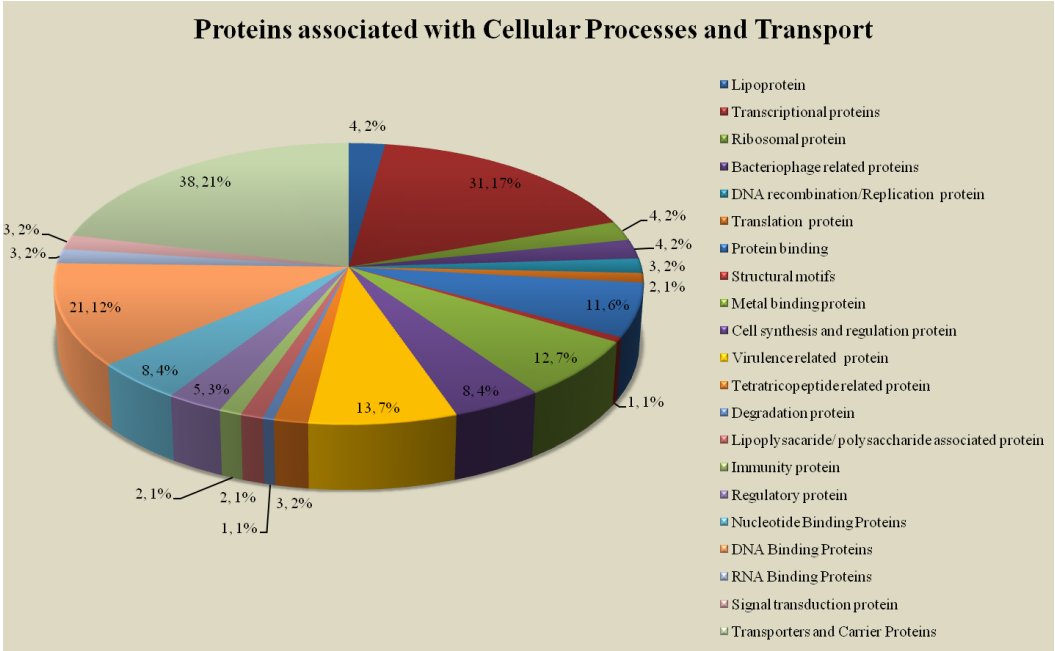


Figure 3.29: Different classified categories of HPs involved in the cellular processes and transport

The HP P9WM15, HP O07726, HP P9WLF7, HP O50383, and HP I6X7F2 showed similarities with the lipoproteins and may be involved in the virulence of the *M. tuberculosis* (Sander et al., 2004). The lipoproteins produced in this pathogenic bacteria are required for the anchoring to the host cellular surface and may involve in the processes such as inflammation, adhesion, and translocation (Kovacs-Simon et al.). While HP O05304, HP O33193, and HP I6YG27 (Table C.2) showed the presence of characteristic tetratricopeptide repeat (TPR) and may facilitate the assembly of virulent protein complexes (Kondo et al.). Furthermore, the HP P9WKU9, HP P9WKS7, and HP O05446 were predicted to be involved in the signal

transduction mechanisms and may be crucial for the persistent TB infection (Zahrt and Deretic, 2001).

Among the class of proteins involved in the cellular processes around 55 HPs were involved in diverse binding mechanisms essential for the virulence of the *M. tuberculosis* (Bortoluzzi et al., 2013). In particular, 21 HPs involved in DNA-binding proteins may play a significant role in the bacterial pathogenesis as this class of proteins are necessary for the replication of the DNA as well as in its repair and recombination (Saikrishnan et al., 2003). The sarZ protein of *Staphylococcus aureus* which showed the presence of the conserved winged-helix-turn-helix motif primarily involved in the binding processes and may take part in the bacterial virulence activating the expression of alpha hemolysin (Kaito et al., 2006).

The proteins involved in the regulation of the transcriptional mechanisms are involved in the bacterial virulence (Raman et al., 2004). The sigma factors such as SigD present in *M. tuberculosis* regulate the process of the gene expression in response to the stress conditions (Raman et al., 2004). Furthermore, the protein Srv present in the virulent strain of *Streptococcus* belongs to the CRP/FNR group of transcriptional regulators and controls the pathogenesis of the bacteria (Doern et al., 2008). Likewise, the transcriptional regulatory proteins such as HilC as well as HilD are involved in the DNA binding mechanisms of *Salmonella enteric* and facilitates its infection by initiating the invasion of the host cells (Olekhnovich and Kadner, 2002) Therefore, predicted 31 transcription associated proteins may be crucial for the virulence of *M. tuberculosis* (Table C.2).

Moreover, in order to survive inside the macrophages cells, the metal binding proteins in *M. tuberculosis* regulate and detect the level of the iron in the limiting environment of the human hosts (Banerjee et al., 2007). We have successfully characterized 12 metal binding proteins in the

group of 1000 HPs. Similarly, the HP P9WFL1, HP P96389, and HP P9WFM7 may be involved in the process of RNA binding and may play a vital role in the survival of the pathogen in the human cells by regulating a variety of the virulence factors (Ariyachet et al.). Consequently, these proteins involved in the cellular processes may play a significant role in the pathogenesis of *M. tuberculosis*.

In addition to these proteins, around 38 HPs (Table C.3) were involved in the transport associated processes such as metal uptake, translocation of the nutrient and enzymes as well as the removal of the waste product, are crucial for the survival and virulence of the bacteria (Freeman et al.). The HP I6Y870, HP O07790, HP O33182, HP O07257, and HP P9WPI7 were classified as ABC transporter proteins and may be involved in the translocation of the drugs and enable the development of the drug resistance in *M. tuberculosis* (Braibant et al., 2000). The development of MDR in the pathogenic bacteria against the available antibiotics pose challenges to the current treatment (Kumar and Varela, 2012). The primary cause for the development of the MDR-related conditions were the proteins functioned as the ABC and major facilitator superfamily (MFS) group of transporters (Kumar and Varela, 2012). Therefore, these transport associated HPs can be utilized as the drug targets in the development of anti-tubercular therapeutic agents.

3.4.2.3 Virulent HPs

Despite the availability of diverse drug targets in *M. tuberculosis*, there is a challenge to develop a reliable treatment against the MDR and XDR strains of the TB. Consequently, there is a need to identify the new potential drug targets in the genome of *M. tuberculosis*. Therefore, in this study, several *in silico* methods have been utilized for the classification of the putative virulence factors among the set of 1000 HPs. On the basis of the consensus generated between

the outcomes of VICMpred, VirulentPred, and MP3 servers, the 28 HPs were classified as the "virulence factors" (Table 3.6).

Table 3.6: List of predicted virulence factors present in the set of 1000 HPs obtained from *M. tuberculosis* (**a & b** The SVM based method for the classification of virulent factors. **c** The classification is performed using both "Genomic" and "Metagenomic" options available on the server)

S. No	Uniprot ID	VICMpred [a]	VirulentPred [b]	MP3 server [c]	Remarks
1)	P95201	Virulence factors (6.86)	Virulent	Pathogenic	HNH nucleases
2)	P9WM79*	Virulence factors (1.66)	Virulent	Pathogenic	Purple Acid Phosphatase
3)	I6WZ30	Virulence factors (1.92)	Virulent	Pathogenic	Need structure based annotations
4)	I6X9T8	Virulence factors (2.28)	Virulent	Pathogenic	Luciferase-like monooxygenase
5)	I6Y4V2	Virulence factors (1.09)	Virulent	Pathogenic	
6)	I6WZI4	Virulence factors (0.97)	Virulent	Pathogenic	
7)	P9WKQ1	Virulence factors (1.44)	Virulent	Pathogenic	
8)	P9WKP3	Virulence factors (2.12)	Virulent	Pathogenic	Metallo-beta-lactamase
9)	P9WKM3	Virulence factors (1.06)	Virulent	Pathogenic	
10)	O86370	Virulence factors (1.80)	Virulent	Pathogenic	
11)	O53410	Virulence factors (0.97)	Virulent	Pathogenic	
12)	O06570	Virulence factors (1.22)	Virulent	Pathogenic	
13)	O86348	Virulence factors (1.17)	Virulent	Pathogenic	
14)	P9WM39	Virulence factors (1.56)	Virulent	Pathogenic	
15)	P9WLX5	Virulence factors (1.33)	Virulent	Pathogenic	
16)	P9WK89	Virulence factors (2.10)	Virulent	Pathogenic	Secretory lipase
17)	P9WLQ9	Virulence factors (-1.29)	Virulent	Pathogenic	
18)	L0T9Q6	Virulence factors (1.00)	Virulent	Pathogenic	
19)	P9WLM9	Virulence factors (-0.04)	Virulent	Pathogenic	
20)	O53461	Virulence factors (1.31)	Virulent	Pathogenic	
21)	P9WLL5	Virulence factors (0.22)	Virulent	Pathogenic	
22)	Q10690	Virulence factors (-1.25)	Virulent	Pathogenic	
23)	P9WLK3	Virulence factors (1.37)	Virulent	Pathogenic	
24)	P9WFN1	Virulence factors (0.71)	Virulent	Pathogenic	
25)	P95115	Virulence factors (0.47)	Virulent	Pathogenic	
26)	O53341	Virulence factors (0.82)	Virulent	Pathogenic	
27)	O07801	Virulence factors (0.34)	Virulent	Pathogenic	
28)	O05439	Virulence factors (1.20)	Virulent	Pathogenic	

* Experimentally validated virulent hypothetical protein (Rv0574c).

3.4.3 Conclusions

The limitations of current therapy against the TB infection are the major cause of resulted in high mortality rate. Therefore, there is a requirement for exploring new drug targets in order to

increase the efficiency of the current treatment. Therefore, the 1000 HPs present in the genome of *M. tuberculosis* were analyzed in this study, in order to annotate the undiscovered drug targets in the pathogen. Several updated Bioinformatics methodologies were utilized regarding the allocation of the possible functionalities to these uncharacterized proteins. Around 662 were annotated successfully because they are showing high homology with the proteins available in the biological databases. The major functional classes of the proteins were discussed in section 5.3. Furthermore, on the basis of SVM-based classified methods, 28 HPs were predicted to be putative virulence factors. The structures of HPs with highest predicted scores were analyzed using the concepts of molecular modelling in the following section:

3.5 Structural analyses

The HP P95201, HP P9WM79, HP I6X9T8, HP P9WKP3, and HP P9WK89 were predicted to be HNH nucleases, purple acid phosphatase, luciferase-like monooxygenase, metallo-beta-lactamase, and secretory lipase respectively (Table C.1). While the function of I6WZ30 (Table D.1) was not predicted on the basis of the sequence analyses and therefore, requires further structural analyses. The structural features of functionally annotated virulence factors were also analyzed in order to understand their conformational behavior in the explicit solvent conditions using MD simulations of 50 ns each time scale (total 300 ns).

3.5.1 Materials and Methods

The 3-D structures of HPs were predicted using a variety of protocols based on "Comparative modelling", "fold recognition or threading" as well as "*ab initio*" techniques. The purpose of comparative modelling is to produce a 3-D replica for the proteins without an experimentally determined structure (i.e. the target) using the sequence-based homologies with the protein structures present in the structural databases such as PDB, SCOP and CATH (John and Sali, 2003). In order to generate a reliable model of the classified HPs, the two criteria were set in the current methodology. The primary measure involved the identification of the detectable similarities between the query sequence and the structure of the protein templates. The other criterion which is considered as significant is the accuracy of the sequence-structure alignments before the generation of the final models.

The comparative modelling based structure predictions were achieved by satisfying the spatial restraints (Chothia and Lesk, 1986). The comparative modelling is still considered as the most precise method for the prediction of 3-D structural features if a significant homology is observed between the query and target proteins (Koehl and Levitt, 1999). On the basis of the

accuracy of the alignments between targets and query proteins, diverse models were generated with resolution ranging from low to high comparable to the structures obtained through experimental techniques such as X-ray crystallography and NMR spectroscopy (Sanchez and Sali, 1997).

In some cases, the less precise models can also provide clues about functionalities of the proteins, as structure-function relationships regarding uncharacterized proteins can provide crude knowledge about their role in biological systems. These predictions are considered more reliable as the structural features of proteins are more conserved than its sequences during the course of evolutionary time (Lesk and Chothia, 1980). It was observed that a portion of sequences submitted in the biological databases shows homology with the experimentally deduced 3-D structures (Flockner et al., 1995). Therefore, the comparative or homology modelling can be explored to predict the structures of around 150,000 protein sequences amid 500,000 sequences submitted in the databases (Bairoch and Apweiler, 1999), which showed homology to around 10,000 experimentally resolved 3D structures (Berman et al., 2000). Consequently, the significance of the comparative protein modelling is progressively increasing due to a rapid increase in the experimentally derived structures and the number of characteristic structural folds present in the proteins (Holm and Sander, 1996).

In the presence of low similarities (< 25%) between the query and the template, the comparative modelling resulted in the sub-standard quality of models. In such conditions, the model needs to be predicted without using any template structure on the basis of *ab initio* protocols such as Rosetta based ROBETTA server (Kim et al., 2004), ITASSER (Roy et al., 2010) and others. The I-TASSER (Roy et al., 2010) is formulated on the basis of *ab initio* algorithms, identified the functional sections using the alignments based on multi-fold threading

algorithms and then predicted the 3D structures of the proteins using recurring simulations for structural assembly. Furthermore, the functionalities of the predicted models were identified by I-TASSER on the basis of comparisons between the generated structure and the protein with solved structures deposited in the databases. The output contains results generated on the basis of predictions regarding secondary as well as tertiary structures along with probable Enzyme Commission (EC) numbers, ligand-binding sites and other attributes (Roy et al., 2010).

Likewise, the Rosetta approach based ROBETTA server (Kim et al., 2004) utilizes a *de novo or ab initio* prediction method for the calculating the structural components of proteins with no structural analogs (Berman et al., 2000). In the preliminary steps, the ROBETTA server uses the inbuilt K*Sync module for the alignment of query protein sequence with the structure of the closely related protein. After assessing the alignment with the template structure it performed the calculations regarding the identification of flexible regions in numerous sections by forming a conformational window on the basis of present *de novo* protocols. Conversely, if no structural homologs are available then 3-D models are predicted on the basis of Rosetta based *de novo* algorithm (Misura et al., 2006), which allows the full length prediction of the domains, by determining the conformational space on the basis of fragment-inclusion protocols which results in the large number of decoy structures and the final model filtered on the basis of scoring schemes from this group.

In this study, the comparative modelling was primarily used for the prediction of 3-D structures of uncharacterized HPs. The primary steps of comparative modelling are illustrated in Figure 3.30. This technique of structure prediction involves the selection of reliable templates as well as the allocation of the structural folds, generation of the precise query-template alignments, generation of the accurate models, their evaluations along with model refinements. (Sanchez and

Sali, 1997). These steps are described below:

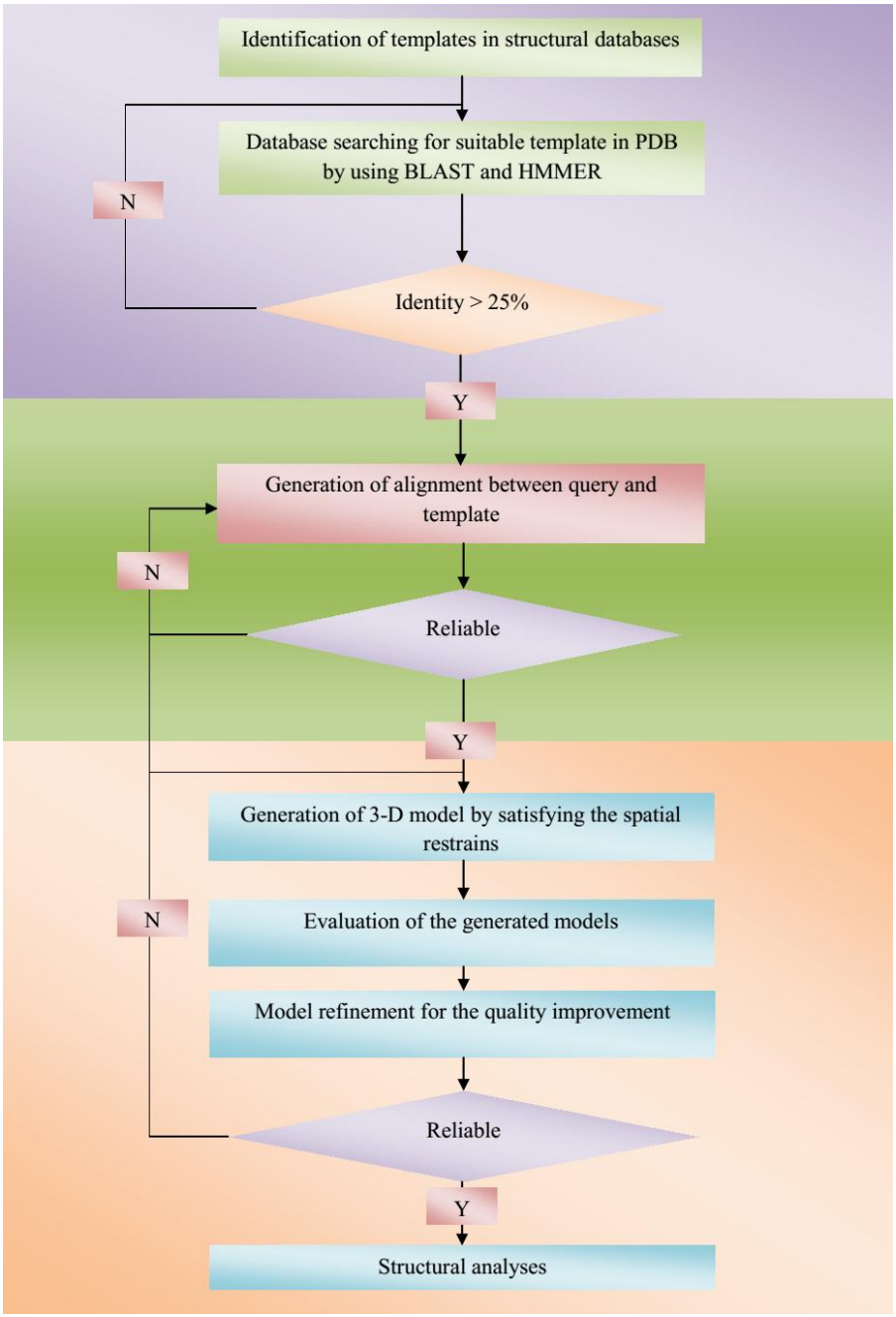


Figure 3.30: Computational framework adopted for the comparative modelling

3.5.1.1 Template search

The preliminary steps in comparative modelling aimed at the detection of homology between the query sequences and template structures deposited in the biological databases such as PDB (Berman et al., 2000), DALI (Holm and Rosenstrom, 2010), SCOP (Murzin et al., 1995),

and CATH (Orengo et al., 1997). The selection of the suitable templates was carried out on the basis of the assessments regarding the identity, query coverage, e-values and many more. On the basis of nature of genome evolution, the probability of predicting the fold using the sequence of an arbitrary protein is around 20% to 50% (Fischer and Eisenberg, 1997). There are three categories of database searching methods which are helpful in the identification of the analogous fold. The majority of the searching protocols are based on a pair-wise comparison of the query sequence with all the sequences present in the database (Apostolico and Giancarlo, 1998). Extensively, FASTA (Pearson, 1998) as well as BLAST (Altschul et al., 1990) packages are used for pair-wise comparison.

Furthermore, in order to increase the accuracy of database searching, a variety of the algorithms utilizes the comparisons of multiple protein sequences for template identification (Altschul et al., 1997). For this purpose, the PSI-BLAST (Altschul et al., 1997) is an extensively used algorithm which recursively searches suitable homologs with higher precision. The PSI-BLAST based search protocol involved grouping of the homologues proteins which were predicted as the template for a query sequence. Subsequently, weighted multiple sequence alignments were generated using the grouped sequences and then position-specific scoring matrices (PSSM) created on the basis of resulted alignments. The PSI-BLAST re-iterates the searching steps until the identification of a reliable template. PSI-BLAST searches identify double homologs for a query sequence in comparison to the searches performed by the traditional BLAST (Park et al., 1998).

Moreover, several other methods utilized the different “profiles” in the protein sequences to increase the accuracy of the database searching (Rychlewski et al., 1998). While searching a homolog, the profiles were created for all the similar protein identified in the database. Then in

order to classify the suitable templates, the profiles generated for the query sequence were compared to that constructed for the template structures. While, some methods utilizes the information about the secondary structure elements along with the multiple sequence alignments of the proteins, which was proved to be a convenient algorithm for identification of the remote structural homolog because in this case the identity between the query sequence and template structure was calculated below 25% (Fischer and Eisenberg, 1997). As discussed earlier in this chapter, the HMMER (Finn et al., 2011) an HMM-based algorithm constructs the profile-HMMs on the basis of the sequence compositions and using these profiles it performs the template searching in a range of protein databases for a query sequence.

Another class of searching algorithms used the threading or fold recognition patterns based on similar methods for the calculation of pair-wise similarities between the query sequence and template structures (Bowie et al., 1991). These protocols are based on the assumption that the probability of a query sequence to adopt a given 3D folds can be evaluated on the basis of scoring functions which derived the information from the protein alignments. This is carried out for every pair of sequence–structure alignments; as the query sequence is threaded for every protein with the possibility of having same folds. The fold recognition methods are mainly helpful when traditional searching algorithms such as BLAST are unable to identify any reliable templates for the query sequences. The HHpred server (Soding et al., 2005) was used that search the structural databases by utilizing its fold recognition algorithm that performs the pair-wise assessment of the profile HMMs between protein sequences for the detection of remote homology.

The rules for selecting the suitable template are illustrated below

- The similarities between the query and template sequences should be calculated > 25% then

the template is considered as a reliable homolog. If the template and query proteins belong to a similar subfamily of a functional class, then the selection was done on the basis of sequence alignments followed by the phylogenetic analyses.

- The template associated factors such as pH conditions, the number of ligands binding sites, solvent environment, quaternary interactions with other chains and others should be evaluated during the phase of model construction.
- The accuracy of the experimentally derived parameters such as a number of restraints per residue in NMR structure of templates, the resolution and R-factor derived from the X-ray crystallography derived structures are considered significant in the selection of a reliable template.

Among the classified templates in the databases, those satisfying the above criteria were selected for the prediction of structural features of the query sequences. In order to increase the accuracy of the model predictions, the multiple templates alignments can be utilized because some proteins showed fragmentary similarities with different sections of diverse proteins (Fiser and Sali, 2003).

3.5.1.2 Template-target alignment

A variety of structures prediction methods based on the homology assessment as well as fold assignment methods makes use of the alignment between the template structures and the query sequence. Therefore, the reliable target–template alignments are crucial for the predictions of the structural features of the proteins. As discussed in the above section that database searching algorithms identify the remote homologs in the database but the accuracy of such methods is compromised. Therefore, the alignments should be performed between the template and query proteins in order to assess the most suitable match (Baxevanis, 1998). The alignments

between the query and template proteins were considered to be reliable if the identity of 40% or more were calculated. In the case of lower similarities, regions of minimal local sequence identity were observed frequently in the outputs (Saqi et al., 1998). In the lower similarity regions (<30%) or "twilight zone" there was a difficulty in performing an accurate alignment (Rost, 1999). The lower sequence similarity resulted in the progressive alignment errors as well as the increased the number of gaps in the alignments. The consensus of some widely used methods such as CLUSTAL W (Thompson et al., 1994) as well as PRALINE (Simossis and Heringa, 2005) were used in order to increase the precision of the alignment. In particular, PRALINE, which performed the multiple sequence alignments using the information of the secondary structure elements (Simossis and Heringa, 2005).

3.5.1.3 Three-dimensional (3-D) modelling

After forming a reliable alignment between the template and query sequences, a variety of methods can be utilized for the prediction of 3-D structures of the target proteins. The structure prediction methods are widely categorized into three groups which includes structure predictions by utilizing the assembly of rigid bodies (Blundell et al., 1987) followed by the modelling using the segment matching (Claessens et al., 1989) while some methods are based on the model building by the satisfaction of the spatial restrictions (Aszodi and Taylor, 1996). In the first category, the 3-D structure was predicted by forming an assembly of the rigid bodies which resulted from the alignments of the protein structures (Blundell et al., 1987). This respective approach is based on the intrinsic conservation in the variable loops regions, protein cores and side chains of the diverse folds (Blundell et al., 1987).

Whereas, the Segment Matching or Coordinate Reconstruction techniques are based on the possibility that the major portion of the hexapeptide fragments in the structures of proteins

can be classified into around 100 classes of diverse structural features (Unger et al., 1989). Consequently, on the basis of specified atomic positions obtained from the template structures, the 3-D models can be predicted by identification and mounting of the segments present in the respective atomic positions (Unger et al., 1989). The Cα atoms are considered as the reference positions of the fragments, which are conserved in the query sequence and the template structure (Unger et al., 1989).

The methods aimed at the prediction of the 3-D structures by satisfying the spatial restraints, which are generated on the basis of information obtained from the alignments performed between the query sequence and template structures. These algorithms were based on the fact that alignments of the comparable residues in the query and the template proteins lead to the generation of restraints in the form of equivalent angles and distances along with the other significant stereochemical restraints such as bond angles, dihedral angles, non-bonded atom contacts acquired, bond lengths and other empirical force field derived restraints. On the basis of the information obtained regarding the conserved restraints, the 3-D models were constructed. The MODELLER (Fiser and Sali, 2003) module present in Discovery Studio v16 (DS) was utilized for the prediction of the 3-D structures of the HPs by satisfying the spatial restraint.

3.5.1.4 Loop Modelling

The variations among the structural features are produced on the basis of insertions, deletions and other forms of substitutions, among the members of a given protein family. These mutations generally occur in the loop regions which connects the fundamental secondary structural elements in a protein. Therefore, the conformational behaviour of the loop regions determines the functional specificity of a protein, as the binding and active sites are generally present in the loop regions of the proteins (Adhikari et al., 2012). Consequently, in order to

examine the receptor-ligand interactions, the relatively higher accuracy needs to be adopted while modelling the loop regions of the protein structure. The loop regions were re-modelled by utilizing the MODELLER (Fiser and Sali, 2003) module of the DS.

3.5.1.5 Refinement of side chain orientations

In contrast to loop modelling, the orientations in the side chains of the protein models were predicted using the structural information of the analogous structures including knowledge obtained from energetic or steric calibrations (Sali, 1995). These predictions include the calculation of the disulfide bridges which are crucial for the side chain modelling of the proteins. In the majority of the predictions, the structural information of the solved proteins in the databases was used for the side chain modelling (Jung et al., 1994), while the predictions regarding the disulfide bridges were performed utilizing the knowledge of structural properties present in the analogous proteins (Sali and Overington, 1994).

Among the diverse criteria adopted for predicting the side chain orientations, the two measures are considered as important such as coupling amongst the side as well as main chains, while the other include the constant allocation of the dihedral angles in side-chain (Vasquez, 1996). In the final predictions, the stability and packing energies of both these effects were found to be important in order to achieve the desirable accuracy in the resulted models (Lee, 1996). Therefore, in order to increase the accuracy in the side chain modelling the SCWRL 4.0 (Wang et al., 2008) was used. This is command line driven software, developed for the calculation of side-chain orientations on the basis of the fact that typically a fixed backbone resulted from the structures obtained through experimental techniques such as X-ray crystallography or NMR.

3.5.1.6 Model evaluations and optimizations

An evaluation of the accuracy of the predicted models is essential for the nature of the study used to derive structural information. The prediction accuracy can be assessed by utilizing the whole protein structure as well as the fragmented regions. In the primary steps, the adopted folds were analyzed between the query and the template using the energy-based Z-scoring. If the Z-scores between the proteins were calculated to be less than 2.5, then the proteins belong to same fold (Sanchez and Sali, 1998). A variety of methods is available for the evaluation of the predicted models such as TM-score (Zhang and Skolnick, 2005) as well as Discrete Optimized Protein Energy (DOPE) scoring methods (Shen and Sali, 2006).

Several other evaluation methods such as PROCHECK (Laskowski et al., 1996), WHAT_CHECK (Hooft et al., 1996), PROVE (Pontius et al., 1996), VERIFY3D (Eisenberg et al., 1997) and ERRAT (Colovos and Yeates, 1993) present in the Structural Analysis and Verification Server (SAVES) server (http://nihserver.mbi.ucla.edu/SAVES/) were used. The PROCHECK and WHAT_CHECK modules perform the assessment of the stereochemical quality of the predicted structure on the basis of the residue by residue comparison with the experimentally derived proteins structures of the same resolution. Furthermore, the ERRAT module computes the statistical attributes of non-bonded interactions among diverse atom types. The PROVE algorithm calculates the deviation in the Z-scores on the basis of comparisons with highly refined experimentally derived structures present in the PDB database, while VERIFY_3D perform the compatibility analyses of the 3-D model with its primary amino acid sequence.

If the low-quality 3-D models were generated, then the accuracy of the structural features were improved using the GROMOS force-field based energy minimization method present in the

DeepView (Kaplan and Littlejohn, 2001), CHARMM 36 (Vanommeslaeghe et al., 2010) based ChiRotor energy minimization algorithm in DS as well as SCWRL 4.0 (Wang et al., 2008) model refinement algorithm. The methodology adopted for MD simulations were explained in section 3.2.1.4.

3.5.2 Results and Discussions

In addition to the sequence analyses of the HPs, the structure-based analyses of the predicted virulent proteins provided an in-depth understanding of their structural elements involved in their functionalities. The outcomes of these studies are explained below in details:

3.5.2.1 HP I6WZ30

The sequence-based function annotations identified that the HP I6WZ30 showed low sequence similarities to the proteins present in the biological databases and require further structure-based analyses for the function prediction. The structure of HP I6WZ30 was predicted using the 3-D coordinates of extracellular lipase (PDB ID - 2Z8X) obtained from *Pseudomonas* sp. The structure of HP I6WZ30 assumes a β-sandwich like topology with 10 β-strands and three α-helices (Figure 3.31A). The refined structure of HP I6WZ30 showed 94.3% of the residues in the allowed region of the Ramachandran plot. The structure-based functional annotations showed the presence of lipase-like activities in HP I6WZ30.

In order to understand the conformational behavior of HP I6WZ30 in the explicit water conditions, its structure was simulated for 50 ns using the GROMACS package. The structure of HP I6WZ30 showed higher fluctuations in the Rg plot as the values showed elevation up to 15 ns, and then the magnitude decreases continuously till 45 ns (Figure 3.31B) indicating the unstable nature of HP I6WZ30. The continuous noise was observed in the RMSD plot during the course of 50 ns MD simulations (Figure 3.31B) as the values increase up to 15 ns and then start

fluctuating between 0.75 – 1 nm. The constituent residues showed the presence of higher energy with the region corresponding to the residues 55 – 78 showed relatively higher fluctuation because they expressed into the loop region (Figure 3.31C).

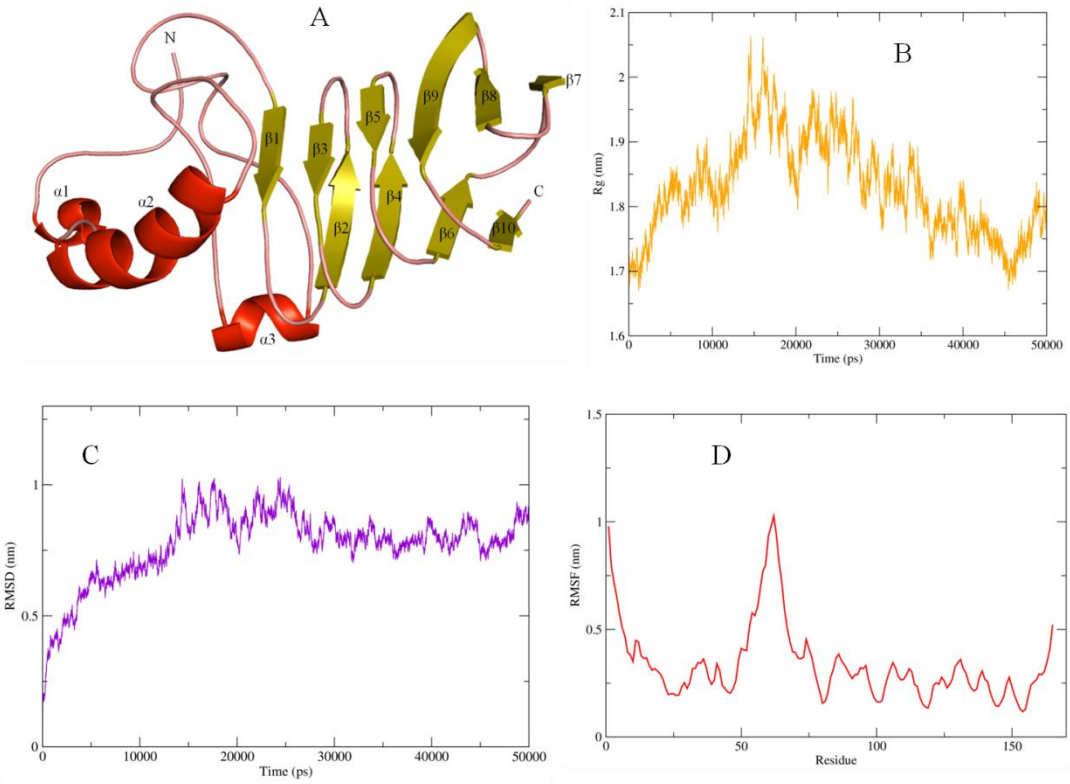


Figure 3.31: (A) The predicted structure of HP I6WZ30 **(B)** The plot showing variation in the Rg values. **(C)** The variation in the RMSD values during 50 ns MD simulations **(D)** The average fluctuations observed in the constituent residues.

3.5.2.2 HP I6X9T8

This protein was predicted to be a luciferase-like monooxygenase (Table C.1). The structure-based analyses showed that the predicted structure HP I6X9T8 showed high similarities with luciferase-like monooxygenase and adopted a TIM-barrel like topology (Figure 3.32A). After simulating the structure of HP I6X9T8, we have observed a continuous variation in the Rg values which are fluctuating between 2 nm – 2.08 nm (Figure 3.32B), while little variations were observed in the RMSD values. These observations indicated that the structural features of HP

I6X9T8 were not deviating from the native conformation during 50 ns MD simulations (Figure 3.32C). Furthermore, the regions corresponding to the residues 46 – 90, 271 - 293 and 340 – 347 showed the highest fluctuation, as these residues adopt the loop region in the three-dimensional space (Figure 3.32D).

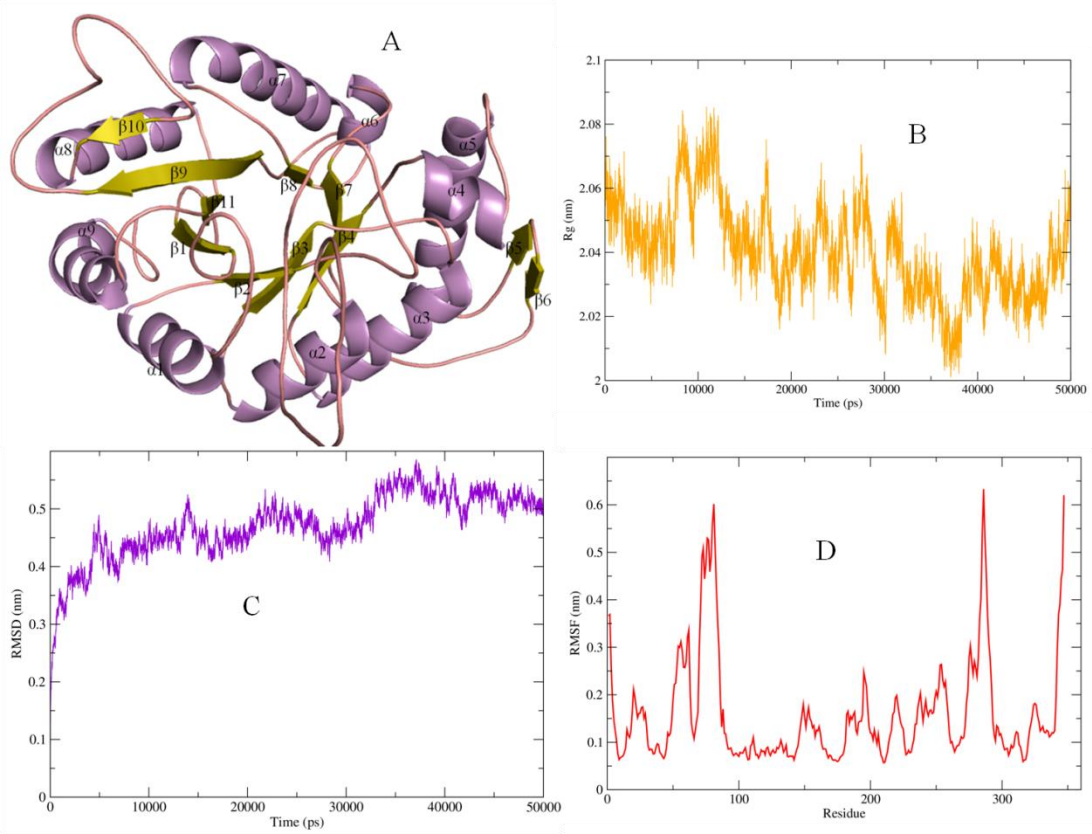


Figure 3.32: (A) The modelled TIM-barrel of HP I6X9T8 **(B)** The Rg plot showing the variations in the compactness of the predicted structure. **(C)** The variation in the RMSD values observed after MD simulations studies. **(D)** The varied fluctuations observed in the constituent residues of different structural elements.

3.5.2.3 HP P9WK89

After analyzing the sequence of HP P9WK89 by utilizing a variety of Bioinformatics tools, it was predicted to be a secretory lipase and may be involved in a variety of hydrolysis reactions. In order to understand its functionalities in details, the structure of HP P9WK89 was predicted using the MODELLER module of the DS on the basis of the 3-D coordinates present for *Candida antarctica* lipase A (PDB ID - 3GUU). The structure of HP P9WK89 showed α/β

hydrolase topology, with 21 α-helices encloses centralized nine-stranded β-sheet (Figure 3.33A). Relatively stable conformation for HP P9WK89 was observed after simulating its structure for 50 ns in explicit solvent conditions. The Rg values were showing continuous fluctuations between 2.1 nm – 2.15 nm (Figure 3.33B) and RMSD values showed increment in a continuous manner (Figure 3.33C), which indicated the unstable nature of HP P9WK89. The higher fluctuations were observed in the constituent residues corresponding to N-terminal residues (1 – 20) as compared to the other secondary structure region because these residues form the loop region (Figure 3.33D).

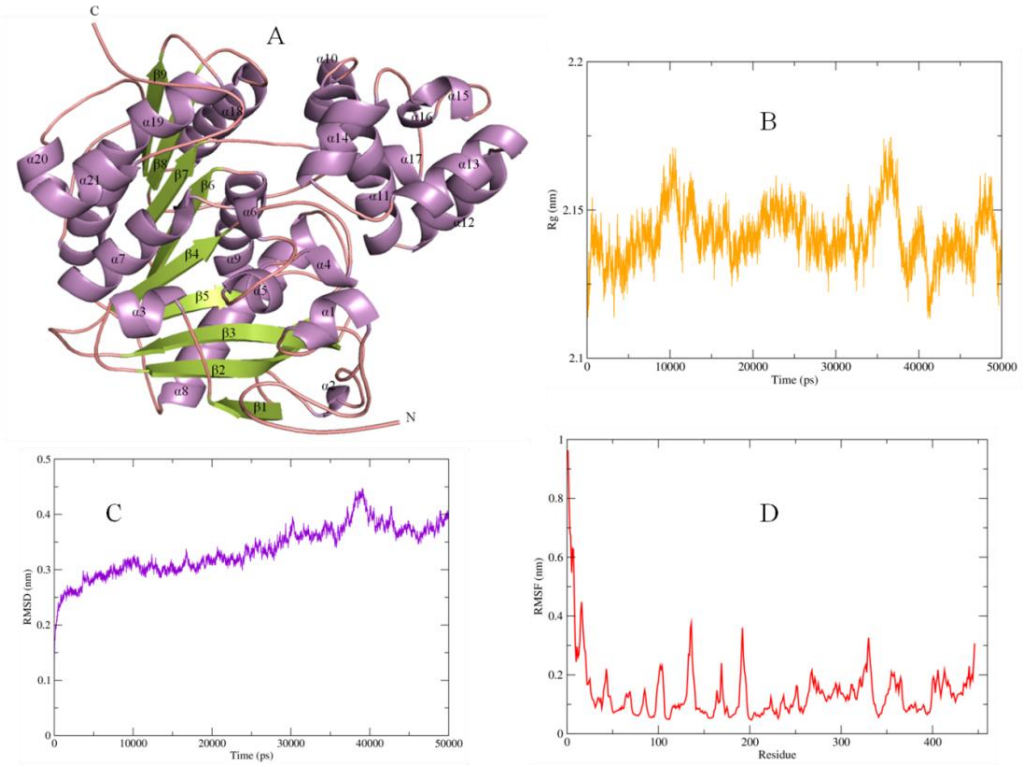


Figure 3.33: (A) The predicted α/β hydrolase topology of HP P9WK89 **(B)** The Rg curve showing variation in the parameters generated for inferring the compactness of the structural element in HP P9WK89. **(C)** The variation in the RMSD values observed during 50 ns MD simulations **(D)** The average fluctuations observed in the constituent residues.

3.5.2.4 HP P9WKP3

The HP P9WKP3 was predicted to be a metallo-β-lactamase, a crucial enzyme that leads to the development of the drug resistance in pathogenic bacteria. The predicted structure of HP

P9WKP3 assumes a sandwich-like topology with 13 β-strands aligned in the form of a central sheet which is surrounded by 13 α-helices which are characteristic to the metallo-β-lactamase (αββα) fold (Figure 3.34A). After minimization and equilibration, the structure of HP P9WKP3 showed higher fluctuations in the compactness of the structure up to 20 ns but the structure becomes relatively stable after 20 ns (Figure 3.34B). These observations were complemented by the inferences obtained from the RMSD plots, which showed a steep increase in the values till 10 ns, while the variations were reduced after this time interval and this stable pattern was continued till 50 ns (Figure 3.34C). The higher fluctuations were observed in the constituent residues which is indicative of the high flexible region in the structure of HP P9WKP3.

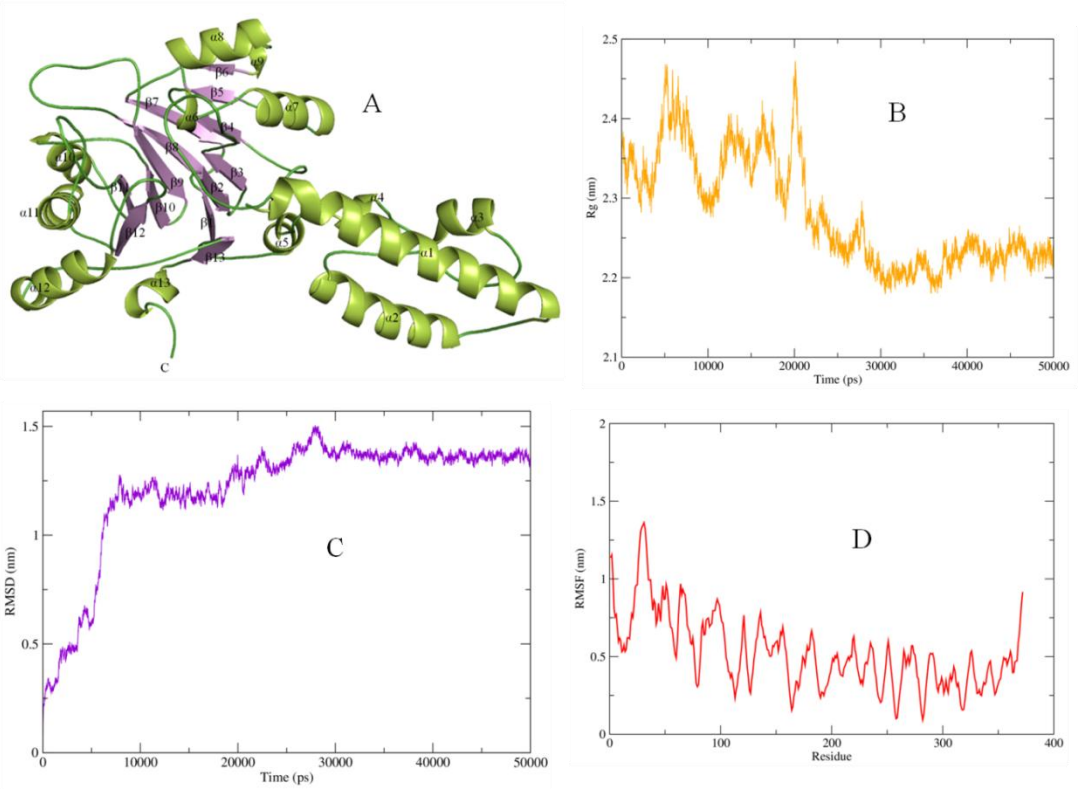


Figure 3.34: (A) The predicted sandwich topology of HP P9WKP3 **(B)** The structural compactness is illustrated in the form of variation in the Rg values. **(C)** The variation in the RMSD values highlighting the stability profile for HP P9WKP3. **(D)** The average fluctuations observed in the constituent residues.

3.5.2.5 HP P9WM79

The sequence-based functional analyses identified the presence of purple acid phosphatase-like characteristics in this protein. The HP P9WM79 (Rv0574c) was experimentally validated regarding its involvement in the maintaining cellular integrity, enable the pathogen survival in the stress conditions as well as crucial for the virulence of *M. tuberculosis* (Garg et al., 2014). The structure of HP P9WM79 assumes α/β fold topology with seven α-helices surrounding a set of 10 antiparallel β-strands (Figure 3.35A). The HP P9WM79 showed comparatively stable nature in the explicit solvent conditions, there were lesser variations were observed in the studied parameters (Figure 3.35B - D).

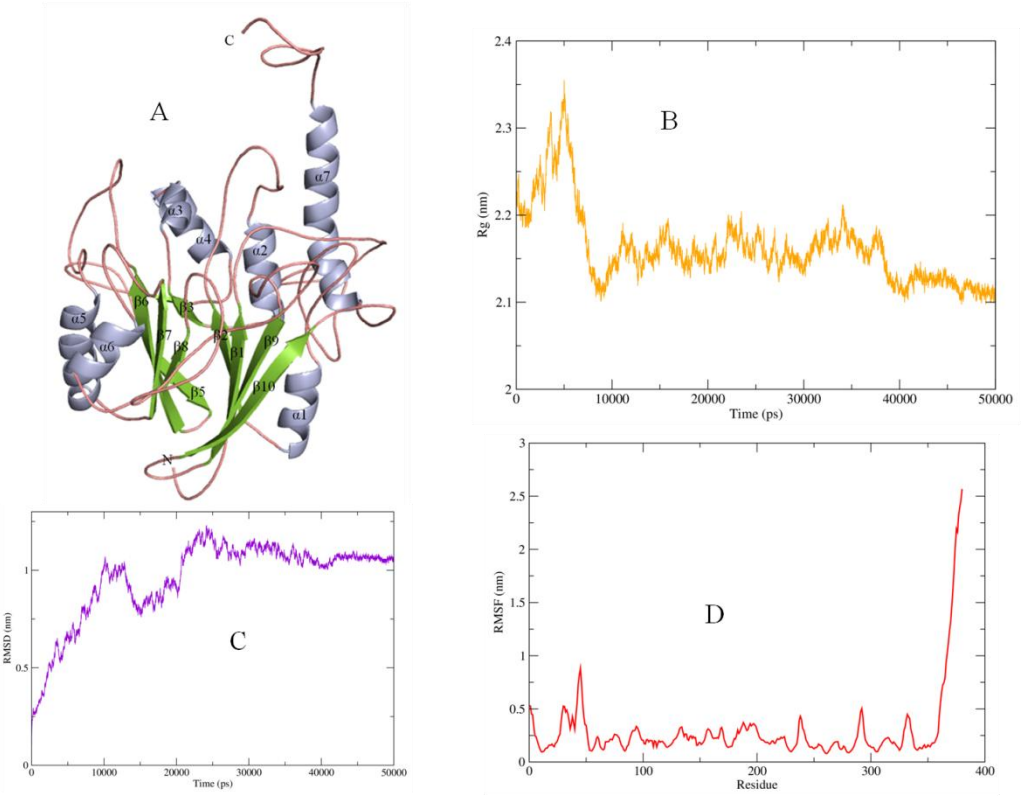


Figure 3.35: (A) The predicted structure showing α/β topology of HP P9WM79 **(B)** The plot showing variation in the Rg values. **(C)** The variation in the RMSD values during 50 ns MD simulations **(D)** The average fluctuations observed in the constituent residues.

3.5.2.6 HP P95201

The HP P95201 was predicted to be a member of HNH nucleases family (Table C.1) showed unique fold in the predicted structure which is showing less structural similarities with the crystal structures deposited in the publically available protein databases. The structure of HP P95201 showed the presence of 17 α-helices and with protruding two β-strands in the form of the hairpin (Figure 3.36A). The dynamics studies of HP P95201structure in explicit solvent conditions showed that a relatively stable behavior as the Rg values corresponding to the compactness of the protein structure showed relatively lower variations (Figure 3.36B).

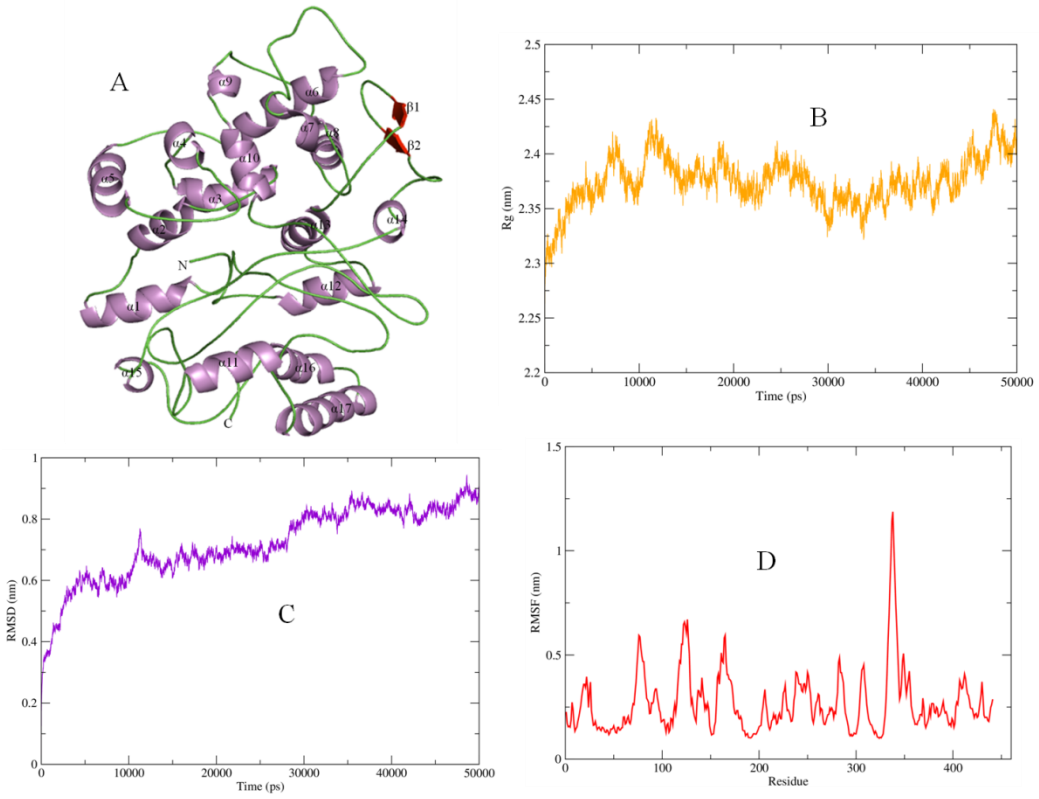


Figure 3.36: (A) The mixed topology of the modelled structure generated for HP P95201 **(B)** The Rg plot illustrating the compactness of the 3-D model. **(C)** The resulted stability parameters are illustrated in the form of varied RMSD values. **(D)** The average fluctuations observed in the constituent residues.

Furthermore, the RMSD plots showing a steady increase in the values throughout the 50 ns MD simulations and fluctuating in the range of 0.4 nm – 0.8 nm (Figure 3.36C). While the

RMSF plot showed higher fluctuations in the constituent residues especially in the region of corresponding to the residues from 325 – 350, this fragment of HP P95201 structure express to form the protruding β-hairpin region.

3.5.3 Conclusions

The structural analyses of these six classified virulence factors provide us an insight into their conformational behavior, which showed that these proteins are quite unstable in nature. As stability is related to the activity of the proteins (Shoichet et al., 1995), therefore, these predicted virulent proteins actively participate in the metabolic pathways of the pathogen and can be utilized as the possible drug targets in the other biochemical studies.

3.6 Summary

The current study focuses on the computational analyses of the virulent proteins present in *M. tuberculosis*. The computational methodology deals with the conformational analyses of the virulent proteins using MD simulation techniques. The proteins associated with the multi-drug resistance (MDR) and acid resistance were extensively analyzed. Novel MDR associated mutations were predicted by analyzing the experimentally validated mutational landscapes available in the literature. The structural changes upon mutations were observed using molecular modelling techniques. A variety of new point mutations associated with the development of the resistance against each first line drugs were predicted. Around 62 mutations in arabinosyltransferase were predicted to be involved in ethambutol resistance followed by the analyses of catalase-peroxidase protein which showed the presence of 61 isoniazid tolerant substitutions. Similarly, the classification for rifampicin and streptomycin identified around 104 and 64 drug resistant mutations respectively. The conformational behaviors of the computationally predicted mutations were comparable with the experimentally derived mutations in the explicit solvent conditions.

Furthermore, a constant pH based MD simulations in acidic conditions were performed on the virulent proteins showed that the 3-D structural features of lipF and PhoP were conserved in the solvent conditions of pH 4. These observations indicated that the respective proteins can be considered as the primary factors in the development of acid tolerance in *M. tuberculosis* and can be targeted for the development of better therapeutics.

The knowledge developed by the analyses of these virulent proteins was used for the classification of the novel virulence factors in the set of 1000 HPs present in *M. tuberculosis*. The sequence-based function prediction tools successfully annotated 662 HPs, which were

further classified into 483 enzymes, 141 HPs of cellular processes and 38 proteins were found to be involved in the transportation associated mechanisms. The 28 HPs were classified as putative virulence factors and six proteins with the highest predicted scores selected for the structural analyses. The MD simulations highlighted their instability in explicit water conditions. Due to the availability of the suitable clones in the repositories, HP O05439 was selected for the experimental validation. The details of the experimental procedures and outcomes obtained are explained in next chapter.

CHAPTER 4

EXPERIMENTAL

In the previous chapter, the virulence characteristics of the HPs were analyzed and predicted. The generated outcomes were utilized in the experimental validation of the Rv3906c (HP O05439) in this section. The gene of Rv3906c was cloned and expressed in the bacterial systems. The pathological significance of Rv3906c is highlighted in this chapter through the principles of Molecular modelling and MD simulations.

After the classification of virulent Hypothetical Proteins (HPs), the experimental validation of Rv3906c (HP O05439) was performed using recombination DNA technology, which was aimed to observe the expression of cDNA. The cDNA of *Rv3906c* gene was successfully cloned. The cloning vectors are the bacterial plasmids containing necessary genetic signals and segments for the replication, foreign DNA insertion, recognition of cells along with recombinant DNA, and expression the foreign DNA. Furthermore, the restriction enzymes, as well as the DNA ligase, are main tools in the production of recombinant DNA.

The primary steps of molecular cloning involve PCR-based amplification of the DNA fragments (gene of interest) to be cloned followed by the restriction digestion of the resulted product and vector plasmid. These generated fragments of concerned gene and vector plasmid were ligated in the presence of T4 DNA ligase. Finally, the recombinant DNA was transformed into the host cells DH5α strain of *E. coli*, where they replicate and produce clones. The replicated clones were isolated, purified and again transformed into BL21(DE3) strain of *E. coli* for the further analyses of the protein expression and purifications.

4.1 Materials and methods

4.1.1 Reagents

Protease inhibitor cocktail was procured from Roche (Life Science Custom Biotechnology USA). NaCl, EDTA, and other reagents were purchased from Merck (India). DNA preparation kits, ampicillin, kanamycin, monoclonal anti-His antibody and mouse secondary antibodies, dialysis tubing, Luria broth, Luria agar were from the Sigma Chemical Co. (St. Louis, USA). Ni-HF column was obtained from GE healthcare (GE Healthcare Life Sciences, Uppsala, Sweden). All chemicals and reagents were of analytical grade and highly pure. Luria-Bertani broth was purchased from Merck, Darmstadt, Germany. Ampicillin was purchased from Sigma, Saint Louis, MO, USA while plasmid pET-21c (Novagen, Wisconsin, USA) was used as an expression vector.

4.1.2 Strains and plasmid

The clones of Rv3906c were purchased from the DNASU plasmid repository (https://dnasu.org/DNASU/), which is a repository for the collection and distribution of plasmid clones. The DNASU currently has over 200,000 plasmids in addition to about 75,000 clones for human as well as mouse. Furthermore, DH5α and BL21(DE3) strains of *Escherichia coli* were used for cloning and expression, respectively.

4.1.3 Primer design

The sequences of the forward and reverse primers were estimated by using Oligo version 6.0 software (Rychlik, 2007), based on the *Rv3906c* gene present in the genome of *M. tuberculosis* H37Rv (Figure 4.1). The primary step involved includes the usage of the *Rv3906c* gene as input and investigate the restriction sites for a suitable enzyme which can be used for the

cloning. The *NheI* and *XhoI* restriction sites were selected as the respective enzymes were found; do not cleave within the coding region of the *Rv3906c* gene. Moreover, the selected enzymes also had unique restriction sites in the MCS (Multiple Coding Sequence) of the pET21c cloning vector. The primers were synthesized, DST purified and supplied in the lyophilized form by Sigma Aldrich (USA). These primers were dissolved in suitable volumes of the sterilized water as instructed in the company technical datasheet in order to a definitive concentration of 100µM and then they are stored at -20°C until further use.

>Rv3906c
ATGGAGTACTGCATAGCCGGCGACGACGGCAGCGCCGGGATCTGGAACCGCCCGTTCGACGTCGAC
CTCGACGGTGACGGCCGGCTGGACGCGATTGGCCTGGATCTCGACGGCGACGGTCTGCGCGATGACGC
GCTGGCCGACTTCGACGGCGACGACGTTGCCGACCACGCCGTATTCGACGTCGACAACGACGGCACCC
CGGAAAGCTACTTCATCGACGACGGATCGGGGACCTGGGCGGTCGCCGTCGACCGCGGCGGACAACT
GCGCTGGTATGGGCTCGACGGCGTCGAGCACACCGGTGGTCCACTGGTTGACTTCGACGGGTTCGGTG
GTTTAGACGACCGGCTACTCGATACCGACGGTGACGGGCTGGCCGATCGGGTGCTGTGTGCTGGTGAG
CAGCGTGTGACCGGATACGTCGACACCGATGGCGACGGCCGCTGGGACGTACGGCTGACCGACACCG
ACGGCGACGGCACC**GCCGACGGCGCCAGCAGCCTT**

Figure 4.1: The *Rv3906c* gene sequence with primer sites (highlighted)

Sequences of primers used in this study, with relevant restriction sites, are as follows

Fp_Rv3906c: 5’ GCTAGCATGGAGTACTGCATAGCCGGCG 3’

Rp_Rv3906c: 5’ CTCGAGAAGGCTGCTGGCGCCGTCGGC 3’

4.1.4 Polymerase chain reaction

The Polymerase chain reaction (PCR) is an essential method in the molecular biology, which is used to amplify the DNA contents of the genes and produces thousands to millions copies of the concerned sequence. The PCR techniques involve the usage of a thermostable enzyme, the *Taq* DNA polymerase, which could be added at the start of the reaction. The primary step of three staged PCR involves denaturation of DNA into two strands by the use of high temperature, followed by the primers annealing to the complementary sequences within the target gene and finally the *Taq* DNA polymerase adds dNTPs complementary to the parent DNA

strand sequence in 5'-> 3' direction to extend the new daughter strand. The *Rv3906c* gene was amplified by using the above primers and following optimized conditions in a thermocycler;

Initial denaturation at 98 °C for 8 min

Cyclic denaturation at 98 °C for 30 s
Primer annealing at 55 °C for 30 s
} 25 cycles

Extension at 72 °C for 45 s

Final extension at 72 °C for 10 min.

4.1.5 Agarose gel electrophoresis

The PCR product was validated by using the 1% agarose gel electrophoresis. The required quantity of agarose was placed in an Erlenmeyer flask along with the suitable amount of 1 X TAE buffer, which was diluted from a 50 X TAE stock (242 g Tris, 57.1 ml acetic acid, 100 ml 0.5 M EDTA). This mixture of agarose was heated in a microwave for 1-2 min in order to obtain a clear solution; ethidium bromide (0.05 mg/ml) was added before pouring into a casting tray containing well combs and allowed to solidify. Then the PCR products were mixed with gel loading buffer in a ratio of 1:5, which was then loaded into the gel and run at 60 V for approximately 1 h. The size of the PCR product was confirmed by running a DNA Ladder alongside the samples. Gels were viewed on a UV trans-illuminator. The EtBr or ethidium bromide intercalates between the bases in DNA double helix and results in a strong, UV-excitable orange fluorescence that makes the position of DNA bands visible on an agarose gel. The desired bands were excised from the gel and purified by using the Gel Filtration Cal Kit (GE Healthcare) according to manufacturer's protocol.

4.1.6 Restriction digestion and ligation

The molecular scissors or restriction endonucleases are the enzymes utilized for precise digestion of both the insert DNA and the pET21c vector. These cleavages generated fitting sticky

ends which for the process of ligation catalyzed by T4 DNA ligase. The restriction digestion was carried out on the basis of the standard protocols (Sambrook and Russell, 2001). For digestion, 0.1 volume of the equivalent restriction buffer (10X) was added to the DNA solution. The incubation of the respective restriction enzymes along with the produced mixtures was carried out at 37 °C overnight. After around 4 hours, the restricted DNA fragments of *Rv3906c* were analyzed on 1 % agarose gels. The desired bands were excised and purified by using Gel Filtration Cal Kit (GE Healthcare). In order to increase the efficiency ligation reaction between of the restricted DNA insert into the vector pET21c, the insert ratio of 1:3 was used. For the ligation reaction, the reaction mixture was prepared to contain the vector, insert, ligation buffer (has ATP) as well as T4 DNA ligase. The respective ligation mixture was incubated overnight at 4 ^{o}C. Furthermore, the ligated plasmids were transformed into competent host cells DH5α strain of *E. coli*.

4.1.7 DNA quantification

DNA quantity was verified by using the NanoDrop Spectrophotometer ND-1000 (Thermo Scientific). A one microliter aliquot of the sample was loaded onto the stage, and its absorbance was measured at 260 nm. This was done in triplicate for both vector and insert and the average value was used as the DNA concentration. TE buffer was applied as a control to "zero" the reading of spectrophotometer before the concentration of the DNA samples was measured.

4.1.8 Preparation of competent BL21 cells

The bacterial host cells were subjected to Ca^{2+} which interacted with their cellular envelopes and made the cells "competent" in order to take up DNA from their adjacent environments. This method increases the cell envelopes permeability for DNA uptake. The *E.*

coli cells were cultured in LB medium and a single bacterial colony was used to inoculate 5 ml sterile LB medium and the reaction mixture was incubated at 37 °C, 180 rpm overnight. One milliliter of this culture was used to inoculate 29 ml of fresh LB medium and shaken at 37 °C, 180 rpm until it reached an OD of 0.375 at 600 nm. The culture was immediately placed on ice and kept cold for the duration of the procedure. Then the cell solution was centrifuged at 4000 rpm for 10 min and the cells were collected from the pellet. The cells were re-suspended in 10 ml cold 100 mM $CaCl_2$ and centrifuged at the same speed and re-suspended in 10 ml cold 100 mM $CaCl_2$. The produced mixture was then subjected to the ice for 20 min. Then the mixture was centrifuged and the competent cells were then re-suspended in 2 ml 100 mM $CaCl_2$ containing 10 % glycerol. The 100 µL of the prepared competent cells were dispensed and incubated at 4 °C overnight, and then stored at -70 °C. According to the protocol followed (Ausubel *et al*., 1989), it is suggested that competent cells are most efficient when prepared 24 h prior to transformation. Therefore, the produced competent cells were incubated at 4 °C overnight in order to enhance the effectiveness of the subsequent transformation procedure. 1 µL of the plasmid DNA was added to 100 µL of the competent cells and incubated on ice for 20 min and thereafter subjected to heat shock for 1 min at 42 °C. After the heat shock, the cells were immediately incubated on ice for 5 min. Then 900 µL of fresh LB medium was added to each of the mixtures which were then incubated at 37 °C for 1 h at 180 rpm.

4.1.9 Cloning and Expression of *Rv3906c* gene

The 560-bp coding region of the *Rv3906c* gene was amplified by PCR using forward primer 5’GCTAGCATGGAGTACTGCATAGCCGGCG 3’ and reverse primer 5’ CTCGAGAA GGCTGCTGGCGCCGTCGGC 3’. This amplification generated *Nhe1* and *Xho1* sites on each end of the amplified fragment. The *Rv3906c* gene was sub-cloned into a pET21c vector with the

C-terminal His6 tag. The constructed expression vector (pET21c- Rv3906c) containing a coding region of the *Rv3906c* gene was transformed into host cells of *E. coli* BL21 (DE3) on the basis of the standard protocol. The culture of the competent cells from a newly transformed plate was transferred in LB media (containing 100µg/µl ampicillin) for the inoculation as well as incubation was carried out at 37 ^{o}C, with constant agitation at 220 rpm in an incubator shaker till the absorbance of the solution reaches the magnitude of 0.6 at 600 nm. The induction of the cultured cells was carried out by using 0.25 mM IPTG (Sigma, Saint Louis, USA) followed by 15 hours of incubation at 16 °C. The resulted cells were centrifuged at 7000 rpm for 10 min at 4 °C. Then the collected cells were dissolved in the solution containing 50 mM Tris-HCl buffer, 500 mM NaCl, 5 mM β-mercaptoethanol, 5% (v/v) glycerol, 100 mM phenylmethanesulfonyl fluoride (PMSF), 0.1mg/ml lysozyme and 1% (v/v) triton X-100 (U. S. Biochemical Corp). The lysate cells were sonicated on ice and then centrifuged for 30 min at 13, 000 rpm at 4°C. The supernatant was collected for purification of Rv3906c protein by the nickel affinity chromatography.

4.1.10 Purification of Rv3906c

During affinity chromatography, a transparent supernatant was passed through the pre-equilibrated Ni-NTA column containing buffer (50 mM Tris-HCl, pH 8.0, 500 mM NaCl, 5% (v/v glycerol, 5 mM β-mercaptoethanol and 10 mM imidazole). This resulted in the binding of protein and the column, which was washed at 4 °C with 50 ml of washing buffer (50 mM Tris-HCl, pH 8.0, 500 mM NaCl, 5% (v/v) glycerol, 5 mM β-mercaptoethanol and 20 mM imidazole). The elution of the attached protein was carried out by using 300 mM imidazole. The fractions were concentrated using Amicon Ultra 10K device (Merck Darmstadt, Germany) and dialyzed against 50 mM Tris-HCl pH 8.0 buffer and stored for further study. The purity of the

protein was confirmed by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE).

4.2 Results and Discussions

After functional annotations of 1000 HPs in the genome of *M. tuberculosis*, around 28 HPs were classified as "virulent". In order to perform their experimental validations, the availability of clones for 28 classified virulent HPs was performed in a variety of plasmid repositories. In the DNASU plasmid repository (https://dnasu.org/DNASU/), around 291 clones of HPs was available. Among this set of plasmids, five clones of interest were selected (Table 4.1). Due to ease of the availability, the plasmid for Rv3906c (Uniprot ID - O05439) was selected for the current study. The outcomes of the experimental approaches are explained below in details:

Table 4.1: List of predicted virulence factors in the genome of *M. tuberculosis* with clones available in DNASU repository

S. No	Uniprot ID	DNASU Plasmid ID	Gene Symbol	Vector	Selection markers
1)	I6Y4V2	MtCD00544376	Rv0804	pANT7_cGST	bacterial:ampicillin
2)	P9WM39	MtCD00532132	Rv1288	pSGX3	bacterial:kanamycin
3)	L0T9Q6	MtCD00544010	Rv1949c	pANT7_cGST	bacterial:ampicillin
4)	P95115	MtCD00544192	Rv2980	pANT7_cGST	bacterial:ampicillin
5)	O05439	MtCD00544327	Rv3906c	pANT7_cGST	bacterial:ampicillin

Blue color – Virulent HP with clone available in the DNASU plasmid repository (https://dnasu.org/DNASU/) and selected for the experimental studies.

4.2.1 Cloning and expression of *Rv3906c*

The *Rv3906c* gene obtained from plasmid pANT7_cGST was subjected to PCR-based amplification, which allows the insertion of *Nhe1* and *Xho1* site at the 5′ and 3′ end, respectively. The amplified gene product was 500 bp long **(**Figure 4.2A**)**. The resulted gene construct and the pET21c vector were digested with same restriction enzymes *Nhe1* and *Xho1*, and the digested gene was subsequently ligated to the pET21c *Nhe1* and *Xho1* backbone fragment. The resulted

clone was verified by colony PCR, restriction digestion with *Nhe1* and *Xho1* endonucleases as well as by DNA sequencing **(**Figure 4.2B**)**. After confirmation, the constructed pET21c-Rv3906c plasmid was transformed into *E. coli* BL21(DE3) competent cells. The recombinant Rv3906c protein was expressed at 16 ^{o}C by inducing with 0.25 mM IPTG. The over-expression of Rv3906c was clearly revealed on SDS-PAGE with a noticeable molecular weight of ~18 kDa (Figure 4.3A).

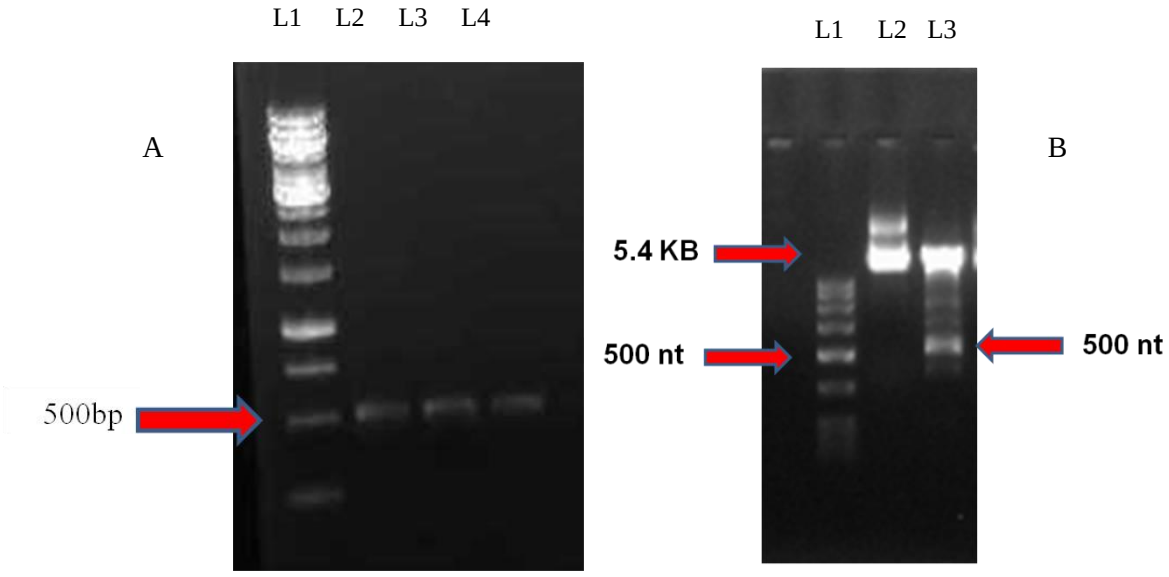


Figure 4.2: (A) Cloning of *Rv3906c* gene: amplified *Rv3906c* gene. Lane 1: Marker while Lane 2, 3 and 4: amplified gene product of 500 bp. (**B**) Restriction digestion of *Rv3906c* and pET21c. Lane 1: *Rv3906c* nucleotide, Lane 2: pET21c nucleotide and Lane 3: Marker.

4.2.2 Purification of Rv3906c

The recombinant Rv3906c protein was isolated and purified to homogeneity at 4 ^{o}C from cell extract under the innate conditions. The Rv3906c protein was detected in the soluble bacterial extract and separated after the centrifugation of sonicated *E. coli* cells at 20,000g. Rv3906c was purified to homogeneity by one chromatographic step on Ni-NTA affinity chromatography. Eluted fractions of Ni-NTA chromatography is shown in Figure 4.3B. The size

of the Rv3906c was observed to be little bigger than expected because the sequence based analyses showed the presence of six phosphorylation sites, one tyrosine sulfation site, and one palmitoylation site. The protein fractions obtained from Ni-NTA chromatography were subjected to SDS-PAGE in order to analyze its purity. A single distinct band was detected on SDS-PAGE demonstrating the purity of Rv3906c (Figure 4.3C)**.**

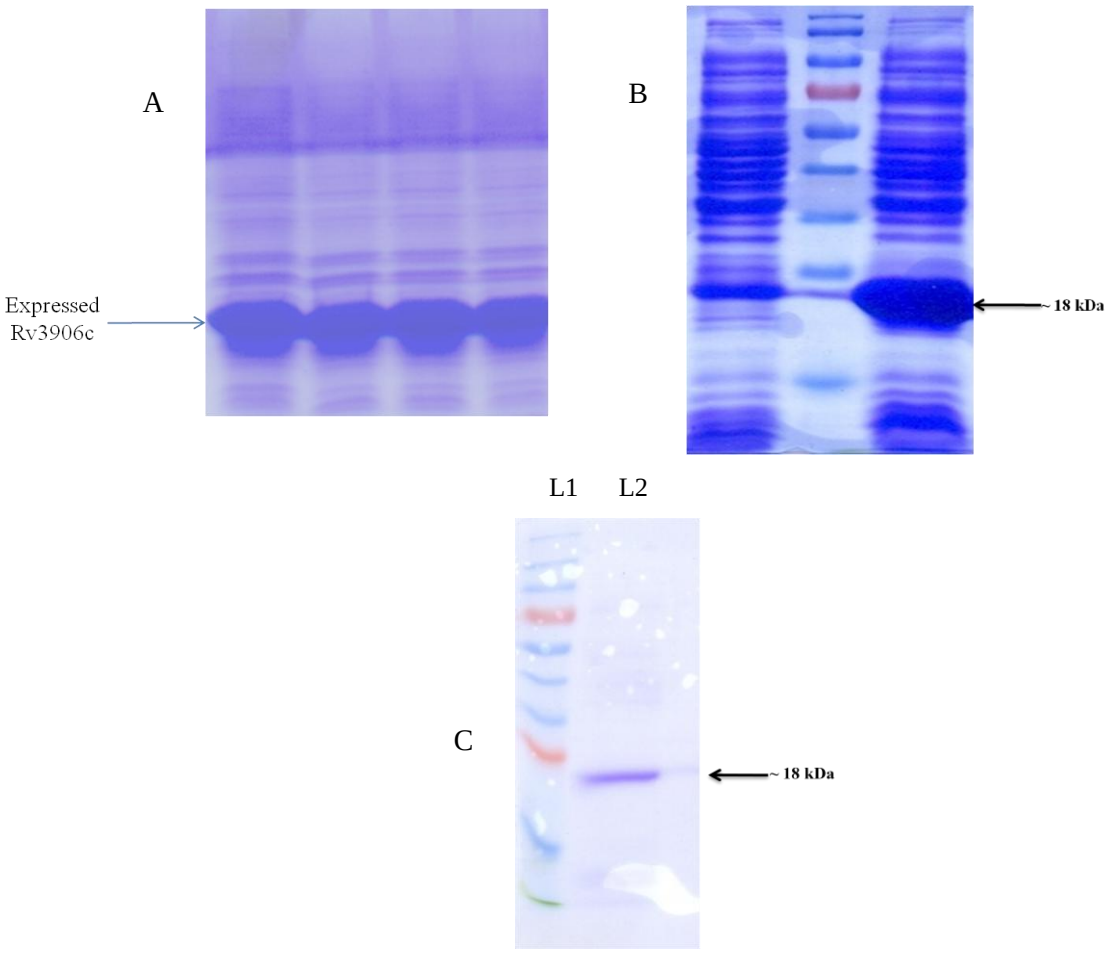


Figure 4.3: (A) SDS-PAGE illustrating the expression of the recombinant Rv3906c protein **(B)** The eluent resulted from Ni–NTA column chromatography showed 17.80 kDa (~18 kD) **(C)** SDS-PAGE of purified Rv3906c protein. Lane 1: Marker and Lane 2: purified Rv3906c

4.2.3 Role of Rv3906c in pathogenesis

Due to unavailability of any structural homolog in the protein databases, the structure of Rv3906c was predicted by using the *ab initio* algorithm implemented in the I-TASSER server. The predicted structure of Rv3906c showed all beta topology with 15 β-strands (Figure 4.4) and after analyzing its sub-cellular localizations, the HP was predicted to be a peripheral protein.

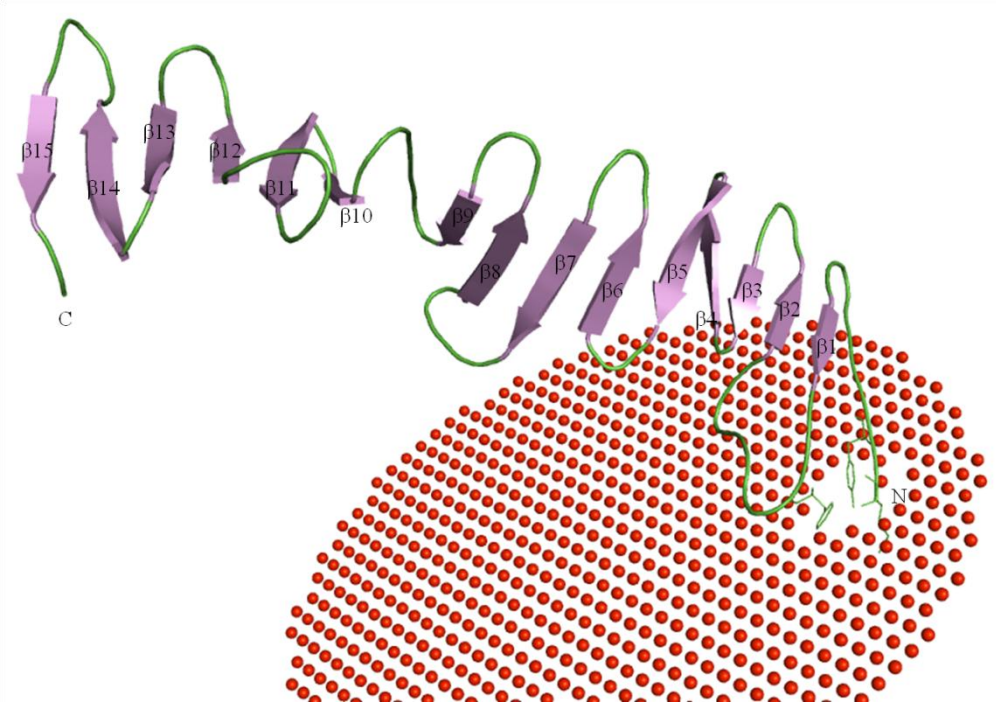


Figure 4.4: The predicted structure of Rv3906c protein showing all beta topology

Moreover, the Rv3906c protein was predicted to be a pullulanase enzyme, which catalyzes the hydrolysis of the polysaccharides and may be crucial for the intracellular survival of the pathogenic bacteria (Uda et al., 2016). The pullulanase enzymes are thermostable in nature (Koch et al., 1997, Duffner et al., 2000), therefore, the Rv3906c may be involved in the survival of *M. tuberculosis* during high thermal conditions as TB infection resulted in an increment in the body temperature (Bhatt et al., 2012). In order to understand the role of Rv3906c in the elevated fever conditions, its structure was immersed in the 1-Palmitoyl-2-oleoyl-sn-glycero-3-

phosphoethanolamine (POPE) composition membrane which is characteristic to the *M. tuberculosis* (Figure 4.5) by suing GROMACS software package. TIP3P water model was used for the simulations of the membrane-based peripheral protein Rv3906c.

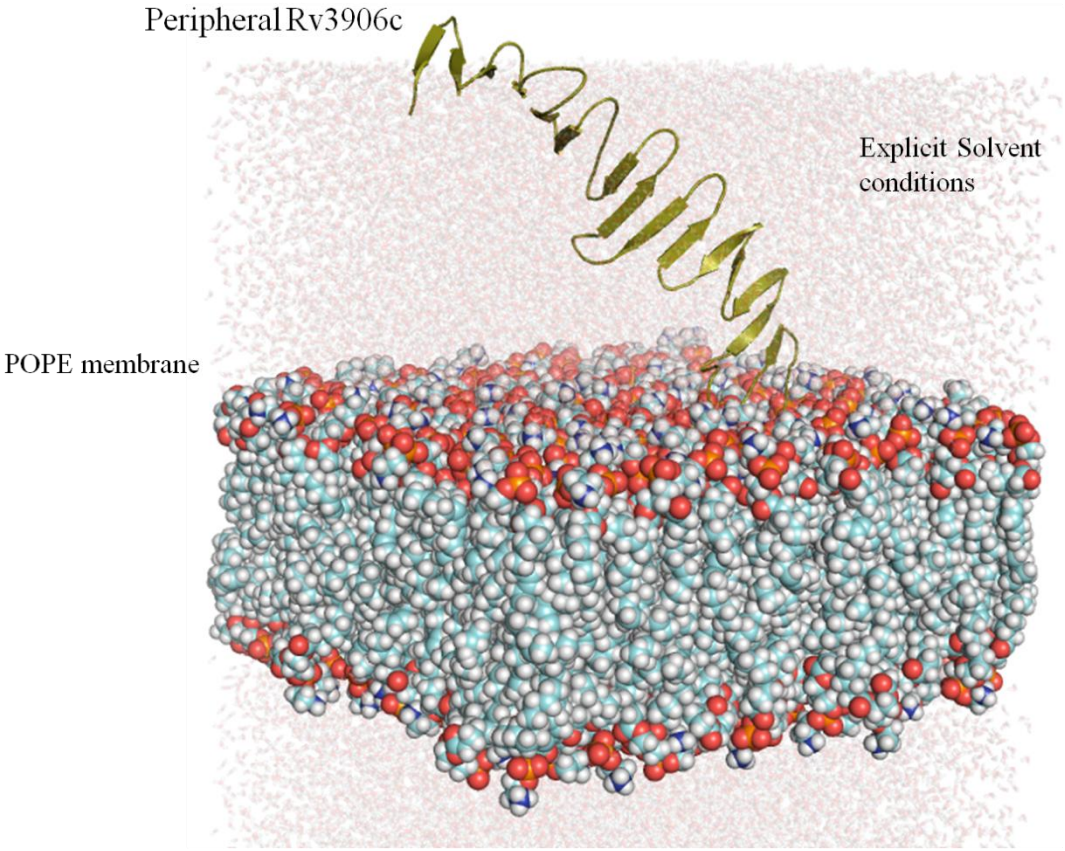


Figure 4.5: The predicted structure of Rv3906c protein immersed in POPE membrane of *M. tuberculosis*

After successful formation of the membrane-protein system, it was subjected to MD simulations each for 100 ns from the temperature ranges from 300 K – 375 K. These analyses showed that the structural features of Rv3906c were maintained even at 375 K. The RMSD plot showed at 350 K, little variation in the RMSD values was observed as compared to the other conditions (Figure 4.6A). While the Rg plots showed that comparable compactness in the structure of Rv3906c was observed for 350 K and 375 K as compared to the other studied

conditions (Figure 4.6B). Furthermore, the RMSF plots indicated the similar behavior of constituent residues in all temperature conditions ranging from 300 K – 375 K, which is attributed to the presence of high flexible regions in the structure of Rv3906c (Figure 4.6C). The MD simulation studies confirm that the Rv3906 can maintain its structural integrity even at high-temperature conditions and can facilitate the survival of the *M. tuberculosis* inside the human host.

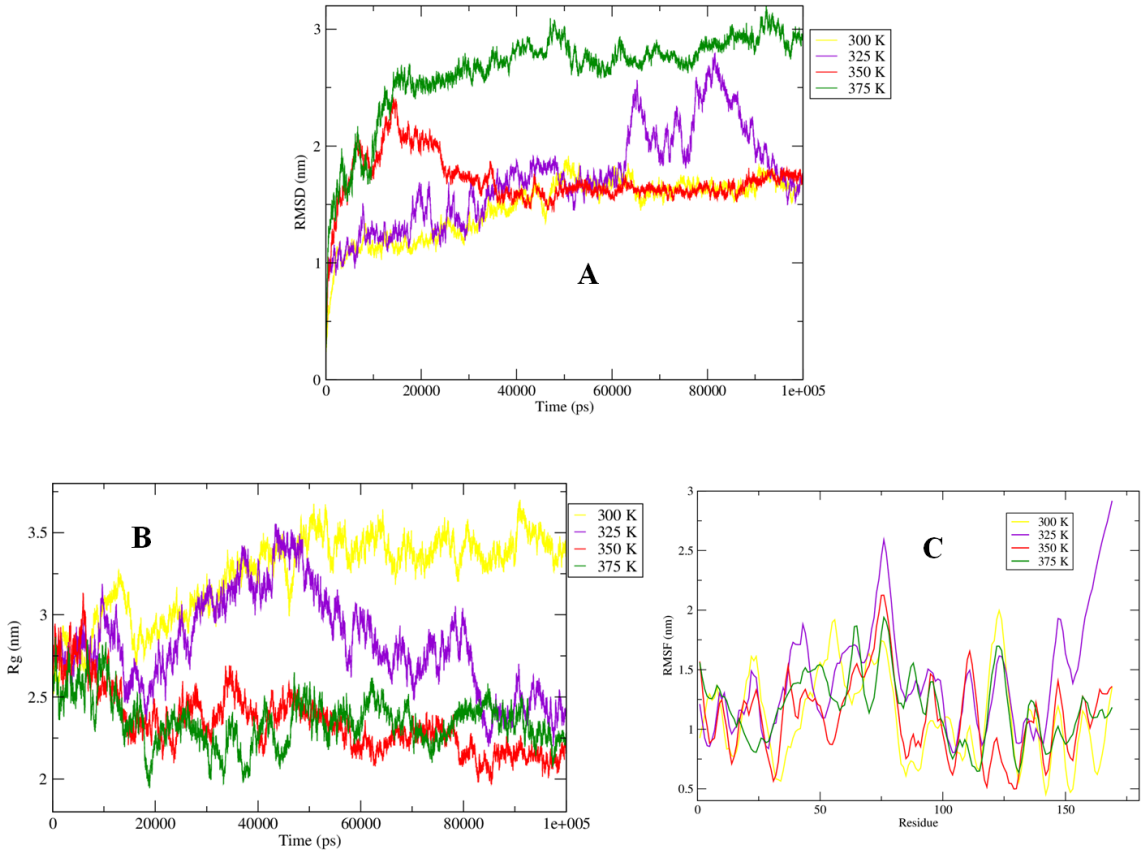


Figure 4.6: (A) The RMSD plots showing relatively higher stability at 350 K as compared to the other conditions after performing MD simulation for 100 ns each with varied temperature. (**B**) The Rg plots showing the presence of relatively higher compactness in the structure of Rv3906c at 350 K and 375 K. **(C)** RMSF plots showing higher fluctuation in the constituent residues in all the studied conditions.

4.3 Conclusions

The *Rv3906c* gene was successfully amplified by using PCR with template cDNA was purified by agarose gel electrophoresis. The size of the product was 500 bp and the gene was successfully cloned into a pET21c expression vector which was validated by colony PCR as well as enzymatic digestion. The six phosphorylation sites, one tyrosine sulfation site, and one palmitoylation site were predicted in the Rv3906c. Due to unavailability of solved 3-D structure of the Rv3906c protein and lack of any structural homolog in the protein databases, its structure was predicted by using *ab initio* techniques. This predicted structure subjected to MD simulations at temperatures ranging from 300 K – 375 K, showed that the Rv3906 can maintain its structural identity at elevated temperature conditions. Therefore, may be involved in the development of temperature tolerance in *M. tuberculosis*.

CHAPTER 5

CONCLUDING REMARKS

This work involved extensive computational studies supported with experimental validations to better understand the conformational profile of virulent proteins of *Mycobacterium tuberculosis*. The computational analyses of multi-drug resistance, acid tolerance and sequence-based functional annotations performed in this study have been connected with the experimentally derived data present in the literature in order to obtain a synergy between the *in silico* and *in vitro* approaches. Moreover, the studies performed in this thesis enabled the discovery of novel hidden virulence factors in the genome of *M. tuberculosis*. The proposed scheme in Figures 3.27 and 3.30 provides a step-by-step guidance regarding the functional annotations of hypothetical proteins (HPs) and molecular modelling strategies adopted in this study.

Drug-resistant tuberculosis (TB) is a pathological condition in which the bacteria developed the resistance against anti-TB drugs and considered as the biggest challenge to the current therapy. An extensive literature search showed that arabinosyltransferases, catalase - peroxidase enzyme KatG, β-subunit of RNA polymerase RpoB and ribosomal protein rpsL illustrated in section 3.2 are the key proteins involved in the development of multi-drug resistance (MDR) against the first line drugs. The isolated MDR-TB strains showed a variety of point mutations in these studied proteins. Therefore, the disease-associated point mutations were classified and their effects on the structure of the studied proteins were extensive analyzed. The conformational behaviors of the predicted disease-associated mutations were compared with the experimentally validated substitutions available in the literature. The outcomes obtained through a comparative study were used for the prediction of MDR-associated point mutations. The

structural basis of the predicted mutations were analyzed using the principles of DFT based techniques and MD simulations. Furthermore, the effects of mutations on the binding of the drugs were analyzed using molecular docking techniques. These extensive computational studies on the proteins of *M. tuberculosis* revealed 62 new point mutations associated with ethambutol resistance in addition to 61 isoniazid, 104 rifampicin, and 64 streptomycin tolerant substitutions. The structural analyses showed that the effects of these computationally predicted mutations were comparable to the experimentally validated substitutions available in the literature. These new findings can be utilized in the understanding of the drug resistance mechanisms present in *M. tuberculosis* and in current drug therapy.

In addition to the drug resistance, the *M. tuberculosis* possesses inherent resistance mechanisms which enable its survival in the limiting conditions inside a human host. The acid tolerance was considered to be the major resistance mechanism which facilitates its survival inside the macrophage cells. Among the diversity of proteins involved in the acid resistance, seven proteins (explained in section 3.3) were found to be the primary cause of the acid tolerance. Literature revealed scarce information about the structural and molecular basis of the proteins involved in such resistance. Therefore, constant-pH based MD simulations were performed for each protein highlighted in section 3.3 and their behavior was studied in the pH range of 3-6. The analyses of the generated outcomes showed that the studied proteins were able to maintain their structure integrity in acidic conditions, particularly lipF, PhoP, Rv2136c, and Rv3671c. The lipF and PhoP protein showed highest structural integrity at pH 4 as compared to the other proteins and therefore, may be the primary protein involved in the pathogen survival. These findings can be utilized in other studies aimed at the development of novel therapeutic agents against the TB infection.

The knowledge developed after analyzing the conformational behaviors of the virulent proteins of *M. tuberculosis* were utilized in the classification of the novel undiscovered virulent proteins in the set of Hypothetical proteins (HPs). The genomes of the studied model pathogens (Discussed in section 3.4.1.6) contains 902 HPs collectively, of which 446 proteins were successfully annotated and the virulence characteristics associated with each protein were assessed (Mohd et al., 2016; Shahbaaz et al., 2015a; Shahbaaz et al., 2015b). The accuracy achieved by the *in silico* protocols were evaluated on the basis of the receiver operating characteristic (ROC) method. Here the accuracy of the protocol and the average area covered by the ROC curve were obtained at 96% and 0.704, respectively, indicating the reliability of the methods (Mohd et al., 2016; Shahbaaz et al., 2015a; Shahbaaz et al., 2015b).

The newly developed protocol in this study was used for the prediction of the functionalities and virulence characteristics of 1000 HPs present in the genome of *M. tuberculosis*. The functions of 662 HPs were successfully annotated because they are showing high homology with the proteins of known functions. These annotated HPs were further classified into 483 enzymes, 141 HPs of cellular processes and 38 proteins were found to be involved in the transportation associated mechanisms (discussed in section 3.4.2). Furthermore, 28 HPs were classified to be putative virulence factors and six HPs with highest predicted scores were subjected to structural analyses and were observed to be unstable in the explicit solvent conditions.

Moreover, the clones for virulent HPs were searched in diverse publically available repositories and on the basis of ease of availability the Rv3906c was selected for further experimental studies. The Rv3906c gene was successfully amplified, cloned and expressed in the pET21c vector. The size of the PCR amplified product was around 500 bp and its expression in

bacterial systems was validated by using colony PCR as well as enzymatic digestion. The six phosphorylation sites, one tyrosine sulfation site, and one palmitoylation site were predicted in the Rv3906c. The sequence-based functional annotations predicted its function to be pullulanase, is a category of polysaccharides hydrolyzing enzymes which are thermostable in nature. In order to validate its role in temperature based tolerance, the structure of Rv3906c was predicted using *ab initio* techniques and simulated at temperatures ranging from 300 K – 375 K. These observations showed that the Rv3906 can maintain its structural integrity even at higher temperature conditions and therefore, may be involved in the development of temperature tolerance in *M. tuberculosis*.

The outcomes obtained from the current study provide new insights into the resistance mechanisms as well as uncovering of new uncharacterized drug targets. These finding can be utilized in the development of novel therapeutic approaches regarding the treatment of the TB infections. It can also be utilized in other biochemical studies regarding the discovery of hidden pathogenic pathways in *M. tuberculosis*.

CHAPTER 6

FUTURE WORK

The extensive analyses performed in this study enlightens the hidden pathogenic pathways present in the *Mycobacterium tuberculosis* which enable its survival in the limiting environment of the human host and also facilitates the development of drug resistance in the bacteria. Furthermore, several novel drug targets were classified in this study and their existence is experimentally validated using the techniques of molecular cloning. The outcomes of this study will improve the process of drug designing, as a novel *in silico* protocol was formulated for the annotation of the uncharacterized proteins. The adopted pipeline can improve the predicatively and efficiency of new drugs from a developmental stage in the laboratory to the commercialization of new products following clinical trials.

The future prospect of this research is based entirely on the experimental validation of the classified putative virulent proteins and to establish their roles in the virulence through the available activity assays. Furthermore, the mechanisms of a variety of transport proteins will be studied using Molecular Dynamics simulations as this category of proteins are involved in the exchange of the nutrients from the surrounding environment enabling the survival of the *M. tuberculosis*. Moreover, the mutation based drug-resistance analyses will further be studied regarding the second and third line of drugs and a general hypothesis will be formulated based on these observations. These results will be utilized in the formation of novel pharmacophore models which will facilitate the process of drug design and discovery against the drug resistant strains of the *M. tuberculosis*.

REFERENCES

ABDALLAH, A. M., GEY VAN PITTIUS, N. C., CHAMPION, P. A., COX, J., LUIRINK, J., VANDENBROUCKE-GRAULS, C. M., APPELMELK, B. J. & BITTER, W. 2007. Type VII secretion--mycobacteria show the way. *Nat Rev Microbiol,* 5**,** 883-91.

ADHIKARI, A. N., PENG, J., WILDE, M., XU, J., FREED, K. F. & SOSNICK, T. R. 2012. Modeling large regions in proteins: applications to loops, termini, and folding. *Protein Sci,* 21**,** 107-21.

AGARWAL, N., LAMICHHANE, G., GUPTA, R., NOLAN, S. & BISHAI, W. R. 2009. Cyclic AMP intoxication of macrophages by a Mycobacterium tuberculosis adenylate cyclase. *Nature,* 460**,** 98-102.

ALDER, B. J. & WAINWRIGHT, T. E. 1957. Phase Transition for a Hard Sphere System. *The Journal of Chemical Physics,* 27**,** 1208-1209.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

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**,** 3389-402.

APOSTOLICO, A. & GIANCARLO, R. 1998. Sequence alignment in molecular biology. *J Comput Biol,* 5**,** 173-96.

ARIYACHET, C., SOLIS, N. V., LIU, Y., PRASADARAO, N. V., FILLER, S. G. & MCBRIDE, A. E. 2013. SR-like RNA-binding protein Slr1 affects Candida albicans filamentation and virulence. *Infect Immun,* 81**,** 1267-76.

ASSIS, P. A., ESPÍNDOLA, M. S., PAULA-SILVA, F. W., RIOS, W. M., PEREIRA, P. A., LEÃO, S. C., SILVA, C. L. & FACCIOLI, L. H. 2014. Mycobacterium tuberculosis expressing phospholipase C subverts PGE2 synthesis and induces necrosis in alveolar macrophages. *BMC Microbiology,* 14**,** 1-10.

ASZODI, A. & TAYLOR, W. R. 1996. Homology modelling by distance geometry. *Fold Des,* 1**,** 325-34.

AZAD, A. K., SIRAKOVA, T. D., FERNANDES, N. D. & KOLATTUKUDY, P. E. 1997. Gene knockout reveals a novel gene cluster for the synthesis of a class of cell wall lipids unique to pathogenic mycobacteria. *J Biol Chem,* 272**,** 16741-5.

BAIROCH, A. 1991. PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res,* 19 Suppl**,** 2241-5.

BAIROCH, A. & APWEILER, R. 1999. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res,* 27**,** 49-54.

BAKER, D. & SALI, A. 2001. Protein structure prediction and structural genomics. *Science,* 294**,** 93-6.

BAKU, #X142, A, Z., NAPI, #XF3, RKOWSKA, A., BIELECKI, J., AUGUSTYNOWICZ-KOPE, #X107, , E., ZWOLSKA, Z. & JAGIELSKI, T. 2013. Mutations in the embB Gene and Their Association with Ethambutol Resistance in Multidrug-Resistant Mycobacterium tuberculosis Clinical Isolates from Poland. *BioMed Research International,* 2013**,** 5.

BAN, F., RANKIN, N. K., GAULD, W. J. & BOYD, J. R. 2002. Recent applications of density functional theory calculations to biomolecules. *Theoretical Chemistry Accounts,* 108**,** 1-11.

BANERJEE, S., NANDYALA, A. K., RAVIPRASAD, P., AHMED, N. & HASNAIN, S. E. 2007. Iron-Dependent RNA-Binding Activity of Mycobacterium tuberculosis Aconitase. *Journal of Bacteriology,* 189**,** 4046-4052.

BARON, C. & COOMBES, B. 2007. Targeting bacterial secretion systems: benefits of disarmament in the microcosm. *Infect Disord Drug Targets,* 7**,** 19-27.

BATEMAN, A., BIRNEY, E., CERRUTI, L., DURBIN, R., ETWILLER, L., EDDY, S. R., GRIFFITHS-JONES, S., HOWE, K. L., MARSHALL, M. & SONNHAMMER, E. L. 2002. The Pfam protein families database. *Nucleic Acids Res,* 30**,** 276-80.

BAXEVANIS, A. D. 1998. Practical aspects of multiple sequence alignment. *Methods Biochem Anal,* 39**,** 172-88.

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**,** 1420-2.

BELLINZONI, M., BURONI, S., PASCA, M. R., GUGLIERAME, P., ARCESI, F., DE ROSSI, E. & RICCARDI, G. 2005. Glutamine amidotransferase activity of NAD+ synthetase from Mycobacterium tuberculosis depends on an amino-terminal nitrilase domain. *Res Microbiol,* 156**,** 173-7.

BERG, S., KAUR, D., JACKSON, M. & BRENNAN, P. J. 2007. The glycosyltransferases of Mycobacterium tuberculosis - roles in the synthesis of arabinogalactan, lipoarabinomannan, and other glycoconjugates. *Glycobiology,* 17**,** 35-56R.

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**,** 235-42.

BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J., MEYER, E. F., JR., BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCHI, T. & TASUMI, M. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol,* 112**,** 535-42.

BERTHET, F. X., RAUZIER, J., LIM, E. M., PHILIPP, W., GICQUEL, B. & PORTNOI, D. 1995. Characterization of the Mycobacterium tuberculosis erp gene encoding a potential cell surface protein with repetitive structures. *Microbiology,* 141 (Pt 9)**,** 2123-30.

BHATT, M. L. B., KANT, S. & BHASKAR, R. 2012. Pulmonary tuberculosis as differential diagnosis of lung cancer. *South Asian Journal of Cancer,* 1**,** 36-42.

BILOFSKY, H. S., BURKS, C., FICKETT, J. W., GOAD, W. B., LEWITTER, F. I., RINDONE, W. P., SWINDELL, C. D. & TUNG, C. S. 1986. The GenBank genetic sequence databank. *Nucleic Acids Res,* 14**,** 1-4.

BISWAS, T., SMALL, J., VANDAL, O., ODAIRA, T., DENG, H., EHRT, S. & TSODIKOV, O. V. 2010a. Structural insight into serine protease Rv3671c that Protects M. tuberculosis from oxidative and acidic stress. *Structure,* 18**,** 1353-63.

BISWAS, T., SMALL, J., VANDAL, O., ODAIRA, T., DENG, H., EHRT, S. & TSODIKOV, O. V. 2010b. Structural Insight into Serine Protease Rv3671c that Protects M. tuberculosis from Oxidative and Acidic Stress. *Structure,* 18**,** 1353-1363.

BJORNSON, H. S. 1984. Enzymes associated with the survival and virulence of gram-negative anaerobes. *Rev Infect Dis,* 6 Suppl 1**,** S21-4.

BLANC-POTARD, A. B. & LAFAY, B. 2003. MgtC as a horizontally-acquired virulence factor of intracellular bacterial pathogens: evidence from molecular phylogeny and comparative genomics. *J Mol Evol,* 57**,** 479-86.

BLOMBERG, M. R. A. & SIEGBAHN, P. E. M. 2001. A Quantum Chemical Approach to the Study of Reaction Mechanisms of Redox-Active Metalloenzymes. *The Journal of Physical Chemistry B,* 105**,** 9375-9386.

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**,** 347-52.

BOEHME , C. C., NABETA , P., HILLEMANN , D., NICOL , M. P., SHENAI , S., KRAPP , F., ALLEN , J., TAHIRLI , R., BLAKEMORE , R., RUSTOMJEE , R., MILOVIC , A., JONES , M., O'BRIEN , S. M., PERSING , D. H., RUESCH-GERDES , S., GOTUZZO , E., RODRIGUES , C., ALLAND , D. & PERKINS , M. D. 2010. Rapid Molecular Detection of Tuberculosis and Rifampin Resistance. *New England Journal of Medicine,* 363**,** 1005-1015.

BOISSIER, F., BARDOU, F., GUILLET, V., UTTENWEILER-JOSEPH, S., DAFFÉ, M., QUÉMARD, A. & MOUREY, L. 2006. Further Insight into S-Adenosylmethionine-dependent Methyltransferases: STRUCTURAL CHARACTERIZATION OF Hma, AN ENZYME ESSENTIAL FOR THE BIOSYNTHESIS OF OXYGENATED MYCOLIC ACIDS IN MYCOBACTERIUM TUBERCULOSIS. *Journal of Biological Chemistry,* 281**,** 4434-4445.

BORADIA, V. M., MALHOTRA, H., THAKKAR, J. S., TILLU, V. A., VUPPALA, B., PATIL, P., SHEOKAND, N., SHARMA, P., CHAUHAN, A. S., RAJE, M. & RAJE, C. I. 2014. Mycobacterium tuberculosis acquires iron by cell-surface sequestration and internalization of human holo-transferrin. *Nat Commun,* 5.

BORTOLUZZI, A., MUSKETT, F. W., WATERS, L. C., ADDIS, P. W., RIECK, B., MUNDER, T., SCHLEIER, S., FORTI, F., GHISOTTI, D., CARR, M. D. & O'HARE, H. M. 2013. Mycobacterium tuberculosis RNA polymerase-binding protein A (RbpA) and its interactions with sigma factors. *J Biol Chem,* 288**,** 14438-50.

BOTELLA, H., PEYRON, P., LEVILLAIN, F., POINCLOUX, R., POQUET, Y., BRANDLI, I., WANG, C., TAILLEUX, L., TILLEUL, S., CHARRIERE, G. M., WADDELL, S. J., FOTI, M., LUGO-VILLARINO, G., GAO, Q., MARIDONNEAU-PARINI, I., BUTCHER, P. D., CASTAGNOLI, P. R., GICQUEL, B., DE CHASTELLIER, C. & NEYROLLES, O. 2011. Mycobacterial p(1)-type ATPases mediate resistance to zinc poisoning in human macrophages. *Cell Host Microbe,* 10**,** 248-59.

BOWIE, J. U., LUTHY, R. & EISENBERG, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science,* 253**,** 164-70.

BRAIBANT, M., GILOT, P. & CONTENT, J. 2000. The ATP binding cassette (ABC) transport systems of Mycobacterium tuberculosis. *FEMS Microbiol Rev,* 24**,** 449-67.

BRANDEEN, C. I. 1980. Relation between structure and function of alpha/beta-proteins. *Q Rev Biophys,* 13**,** 317-38.

BRITTEN, R. J. & DAVIDSON, E. H. 1969. Gene regulation for higher cells: a theory. *Science,* 165**,** 349-57.

BRITTON, K. L., ABEYSINGHE, I. S., BAKER, P. J., BARYNIN, V., DIEHL, P., LANGRIDGE, S. J., MCFADDEN, B. A., SEDELNIKOVA, S. E., STILLMAN, T. J., WEERADECHAPON, K. & RICE, D. W. 2001. The structure and domain organization of Escherichia coli isocitrate lyase. *Acta Crystallogr D Biol Crystallogr,* 57**,** 1209-18.

BROOKS, B. R., BROOKS, C. L., 3RD, MACKERELL, A. D., JR., NILSSON, L., PETRELLA, R. J., ROUX, B., WON, Y., ARCHONTIS, G., BARTELS, C., BORESCH, S., CAFLISCH, A., CAVES, L., CUI, Q., DINNER, A. R., FEIG, M., FISCHER, S., GAO, J., HODOSCEK, M., IM, W., KUCZERA, K., LAZARIDIS, T., MA, J., OVCHINNIKOV, V., PACI, E., PASTOR, R. W., POST, C. B., PU, J. Z., SCHAEFER, M., TIDOR, B., VENABLE, R. M., WOODCOCK, H. L., WU, X., YANG, W., YORK, D. M. & KARPLUS, M. 2009. CHARMM: the biomolecular simulation program. *J Comput Chem,* 30**,** 1545-614.

BRUNING, J. B., MURILLO, A. C., CHACON, O., BARLETTA, R. G. & SACCHETTINI, J. C. 2011. Structure of the Mycobacterium tuberculosis D-alanine:D-alanine ligase, a target of the antituberculosis drug D-cycloserine. *Antimicrob Agents Chemother,* 55**,** 291-301.

BRUST, B., LECOUFLE, M., TUAILLON, E., DEDIEU, L., CANAAN, S., VALVERDE, V. & KREMER, L. 2011. <italic>Mycobacterium tuberculosis</italic> Lipolytic Enzymes as Potential Biomarkers for the Diagnosis of Active Tuberculosis. *PLoS One,* 6**,** e25078.

C NOLAN, A. & MARGOLIASH, E. 1968. Comparative Aspects of Primary Structures of Proteins. *Annual Review of Biochemistry,* 37**,** 727-791.

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**,** 257-67.

CAPRIOTTI, E., FARISELLI, P. & CASADIO, R. 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res,* 33**,** W306-10.

CAR, R. & PARRINELLO, M. 1985. Unified approach for molecular dynamics and density-functional theory. *Phys Rev Lett,* 55**,** 2471-2474.

CARRILLO, H. & LIPMAN, D. 1988. The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.,* 48**,** 1073-1082.

CASE, D. A., CHEATHAM, T. E., 3RD, DARDEN, T., GOHLKE, H., LUO, R., MERZ, K. M., JR., ONUFRIEV, A., SIMMERLING, C., WANG, B. & WOODS, R. J. 2005. The Amber biomolecular simulation programs. *J Comput Chem,* 26**,** 1668-88.

CASE, D. A., DARDEN, T. A., CHEATHAM, T. E., SIMMERLING, C. L., WANG, J., DUKE, R. E., LUO, R., WALKER, R. C., ZHANG, W., MERZ, K. M., ROBERTS, B., HAYIK, S., ROITBERG, A., SEABRA, G., SWAILS, J., GOETZ, A. W., KOLOSSVÁRY, I., WONG, K. F., PAESANI, F., VANICEK, J., WOLF, R. M., LIU, J., WU, X., BROZELL, S. R., STEINBRECHER, T., GOHLKE, H., CAI, Q., YE, X., HSIEH, M. J., CUI, G., ROE, D. R., MATHEWS, D. H., SEETIN, M. G., SALOMON-FERRER, R., SAGUI, C., BABIN, V., LUCHKO, T., GUSAROV, S., KOVALENKO, A. & KOLLMAN, P. A. 2012. AMBER 12. University of California, San Francisco.

CELLITTI, SUSAN E., SHAFFER, J., JONES, DAVID H., MUKHERJEE, T., GURUMURTHY, M., BURSULAYA, B., BOSHOFF, HELENA I., CHOI, I., NAYYAR, A., LEE, YONG S., CHERIAN, J., NIYOMRATTANAKIT, P., DICK, T., MANJUNATHA, UJJINI H., BARRY, CLIFTON E., SPRAGGON, G. & GEIERSTANGER, BERNHARD H. 2012. Structure of Ddn,

the Deazaflavin-Dependent Nitroreductase from Mycobacterium tuberculosis Involved in Bioreductive Activation of PA-824. *Structure(London, England:1993),* 20**,** 101-112.

CHAITIN, G. J. 1966. On the Length of Programs for Computing Finite Binary Sequences. *J. ACM,* 13**,** 547-569.

CHAKRABORTY, S., GOGOI, M. & CHAKRAVORTTY, D. 2015. Lactoylglutathione lyase, a critical enzyme in methylglyoxal detoxification, contributes to survival of Salmonella in the nutrient rich environment. *Virulence,* 6**,** 50-65.

CHAUDHURI, B. N., CHAN, S., PERRY, L. J. & YEATES, T. O. 2004. Crystal structure of the apo forms of psi 55 tRNA pseudouridine synthase from Mycobacterium tuberculosis: a hinge at the base of the catalytic cleft. *J Biol Chem,* 279**,** 24585-91.

CHEN, X., HU, Y., YANG, B., GONG, X., ZHANG, N., NIU, L., WU, Y. & GE, H. 2015. Structure of lpg0406, a carboxymuconolactone decarboxylase family protein possibly involved in antioxidative response from Legionella pneumophila. *Protein Sci,* 24**,** 2070-5.

CHENE, P. 2002. ATPases as drug targets: learning from their structure. *Nat Rev Drug Discov,* 1**,** 665-73.

CHOMSKY, N. 1959. On certain formal properties of grammars. *Information and Control,* 2**,** 137-167.

CHOTHIA, C. 1975. Structural invariants in protein folding. *Nature,* 254**,** 304-8.

CHOTHIA, C. & JANIN, J. 1981. Relative orientation of close-packed beta-pleated sheets in proteins. *Proc Natl Acad Sci U S A,* 78**,** 4146-50.

CHOTHIA, C. & LESK, A. M. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J,* 5**,** 823-6.

CHOTHIA, C., LEVITT, M. & RICHARDSON, D. 1977. Structure of proteins: packing of alpha-helices and pleated sheets. *Proc Natl Acad Sci U S A,* 74**,** 4130-4.

CHOU, P. Y. & FASMAN, G. D. 1974. Prediction of protein conformation. *Biochemistry,* 13**,** 222-45.

CHURCHILL, G. A. 1989. Stochastic models for heterogeneous DNA sequences. *Bull Math Biol,* 51**,** 79-94.

CLAESSENS, M., VAN CUTSEM, E., LASTERS, I. & WODAK, S. 1989. Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng,* 2**,** 335-45.

CLARKE, B. 1970. Selective Constraints on Amino-acid Substitutions during the Evolution of Proteins. *Nature,* 228**,** 159-160.

COHEN, C. & PARRY, D. A. D. 1986. α-Helical coiled coils — a widespread motif in proteins. *Trends in Biochemical Sciences,* 11**,** 245-248.

COHEN, F. E., STERNBERG, M. J. & TAYLOR, W. R. 1981. Analysis of the tertiary structure of protein beta-sheet sandwiches. *J Mol Biol,* 148**,** 253-72.

COHEN, F. E. & STERNBERG, M. J. E. 1980. On the prediction of protein structure: The significance of the root-mean-square deviation. *Journal of Molecular Biology,* 138**,** 321-333.

COHEN, K. A., ABEEL, T., MANSON MCGUIRE, A., DESJARDINS, C. A., MUNSAMY, V., SHEA, T. P., WALKER, B. J., BANTUBANI, N., ALMEIDA, D. V., ALVARADO, L., CHAPMAN, S. B., MVELASE, N. R., DUFFY, E. Y., FITZGERALD, M. G., GOVENDER, P., GUJJA, S., HAMILTON, S., HOWARTH, C., LARIMER, J. D., MAHARAJ, K., PEARSON, M. D., PRIEST, M. E., ZENG, Q., PADAYATCHI, N., GROSSET, J., YOUNG, S. K., WORTMAN, J., MLISANA, K. P., O'DONNELL, M. R., BIRREN, B. W., BISHAI, W. R., PYM, A. S. & EARL, A. M. 2015. Evolution of Extensively Drug-Resistant Tuberculosis over Four Decades: Whole Genome Sequencing and Dating Analysis of <italic>Mycobacterium tuberculosis</italic> Isolates from KwaZulu-Natal. *PLoS Med,* 12**,** e1001880.

COKER, O. O., WARIT, S., RUKSEREE, K., SUMMPUNN, P., PRAMMANANAN, T. & PALITTAPONGARNPIM, P. 2013. Functional characterization of two members of histidine phosphatase superfamily in Mycobacterium tuberculosis. *BMC Microbiol,* 13**,** 292.

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**,** 537-44.

COLOVOS, C. & YEATES, T. O. 1993. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci,* 2**,** 1511-9.

COMAS, I., BORRELL, S., ROETZER, A., ROSE, G., MALLA, B., KATO-MAEDA, M., GALAGAN, J., NIEMANN, S. & GAGNEUX, S. 2012. Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat Genet,* 44**,** 106-110.

CONNOLLY, M. 1983. Solvent-accessible surfaces of proteins and nucleic acids. *Science,* 221**,** 709-713.

COTES, K., DHOUIB, 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**,** 417-27.

CRICK, F. 1953. The packing of [alpha]-helices: simple coiled-coils. *Acta Crystallographica,* 6**,** 689-697.

CRICK, F. 1970. Central Dogma of Molecular Biology. *Nature,* 227**,** 561-563.

CRIPPEN, G. M. 1977. A novel approach to calculation of conformation: Distance geometry. *Journal of Computational Physics,* 24**,** 96-107.

DANG, G., CAO, J., CUI, Y., SONG, N., CHEN, L., PANG, H. & LIU, S. 2016. Characterization of Rv0888, a Novel Extracellular Nuclease from Mycobacterium tuberculosis. *Scientific Reports,* 6**,** 19033.

DARBY, C. M., VENUGOPAL, A., EHRT, S. & NATHAN, C. F. 2011. Mycobacterium tuberculosis gene Rv2136c is dispensable for acid resistance and virulence in mice. *Tuberculosis (Edinb),* 91**,** 343-7.

DAS, R., QIAN, B., RAMAN, S., VERNON, R., THOMPSON, J., BRADLEY, P., KHARE, S., TYKA, M. D., BHAT, D., CHIVIAN, D., KIM, D. E., SHEFFLER, W. H., MALMSTROM, L., WOLLACOTT, A. M., WANG, C., ANDRE, I. & BAKER, D. 2007. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins,* 69 Suppl 8**,** 118-28.

DAYHOFF, M. O. 1978. *Atlas of Protein Sequence and Structure,* National Biomedical Research Foundation.

DE WACHTER, R. 1981. The number of repeats expected in random nucleic acid sequences and found in genes. *J Theor Biol,* 91**,** 71-98.

DEMAIO, J., ZHANG, Y., KO, C., YOUNG, D. B. & BISHAI, W. R. 1996. A stationary-phase stress-response sigma factor from Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A,* 93**,** 2790-4.

DOERN, C. D., HOLDER, R. C. & REID, S. D. 2008. Point mutations within the streptococcal regulator of virulence (Srv) alter protein-DNA interactions and Srv function. *Microbiology,* 154**,** 1998-2007.

DORN, M., MB, E. S., BURIOL, L. S. & LAMB, L. C. 2014. Three-dimensional protein structure prediction: Methods and computational strategies. *Comput Biol Chem,* 53PB**,** 251-276.

DUBEY, V. S., SIRAKOVA, T. D. & KOLATTUKUDY, P. E. 2002. Disruption of msl3 abolishes the synthesis of mycolipanoic and mycolipenic acids required for polyacyltrehalose synthesis in Mycobacterium tuberculosis H37Rv and causes cell aggregation. *Mol Microbiol,* 45**,** 1451-9.

DUBNAU, E., LANEELLE, M. A., SOARES, S., BENICHOU, A., VAZ, T., PROME, D., PROME, J. C., DAFFE, M. & QUEMARD, A. 1997. Mycobacterium bovis BCG genes involved in the biosynthesis of cyclopropyl keto- and hydroxy-mycolic acids. *Mol Microbiol,* 23**,** 313-22.

DUFFNER, F., BERTOLDO, C., ANDERSEN, J. T., WAGNER, K. & ANTRANIKIAN, G. 2000. A New Thermoactive Pullulanase from Desulfurococcus mucosus: Cloning, Sequencing, Purification, and Characterization of the Recombinant Enzyme after Expression in Bacillus subtilis. *Journal of Bacteriology,* 182**,** 6331-6338.

DUMAS, J. P. & NINIO, J. 1982. Efficient algorithms for folding and comparing nucleic acid sequences. *Nucleic Acids Res,* 10**,** 197-206.

DUNNILL, P. 1968. The use of helical net-diagrams to represent protein structures. *Biophys J,* 8**,** 865-75.

EDAGWA, B., WANG, Y. & NARAYANASAMY, P. 2013. Synthesis of azide derivative and discovery of glyoxalase pathway inhibitor against pathogenic bacteria(). *Bioorganic & medicinal chemistry letters,* 23**,** 10.1016/j.bmcl.2013.09.011.

EICHINGER, V., NUSSBAUMER, T., PLATZER, A., JEHL, M. A., ARNOLD, R. & RATTEI, T. 2016. EffectiveDB--updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems. *Nucleic Acids Res,* 44**,** D669-74.

EISENBERG, D., LUTHY, R. & BOWIE, J. U. 1997. VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol,* 277**,** 396-404.

EJIM, L. J., D'COSTA, V. M., ELOWE, N. H., LOREDO-OSTI, J. C., MALO, D. & WRIGHT, G. D. 2004. Cystathionine beta-lyase is important for virulence of Salmonella enterica serovar Typhimurium. *Infect Immun,* 72**,** 3310-4.

ELDHOLM, V., MONTESERIN, J., RIEUX, A., LOPEZ, B., SOBKOWIAK, B., RITACCO, V. & BALLOUX, F. 2015. Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nat Commun,* 6.

ELIAS, R. 2012. Difficulties in applying pure Kohn–Sham density functional theory electronic structure methods to protein molecules. *Journal of Physics: Condensed Matter,* 24**,** 072202.

ENG, W. S., HOCKOVA, D., SPACEK, P., JANEBA, Z., WEST, N. P., WOODS, K., NAESENS, L. M., KEOUGH, D. T. & GUDDAT, L. W. 2015. First Crystal Structures of Mycobacterium tuberculosis 6-Oxopurine Phosphoribosyltransferase: Complexes with GMP and Pyrophosphate and with Acyclic Nucleoside Phosphonates Whose Prodrugs Have Antituberculosis Activity. *J Med Chem,* 58**,** 4822-38.

ESWAR, N., WEBB, B., MARTI-RENOM, M. A., MADHUSUDHAN, M. S., ERAMIAN, D., SHEN, M. Y., PIEPER, U. & SALI, A. 2006. Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics,* Chapter 5**,** Unit 5 6.

EWANN, F., JACKSON, M., PETHE, K., COOPER, A., MIELCAREK, N., ENSERGUEIX, D., GICQUEL, B., LOCHT, C. & SUPPLY, P. 2002. Transient requirement of the PrrA-PrrB two-component system for early intracellular multiplication of Mycobacterium tuberculosis. *Infect Immun,* 70**,** 2256-63.

FANG, H. & GOUGH, J. 2013. DcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic Acids Res,* 41**,** D536-44.

FELDMANN, R. J. 1976. The Design of Computing Systems for Molecular Modeling. *Annual Review of Biophysics and Bioengineering,* 5**,** 477-510.

FELSENSTEIN, J. 1989. Computational molecular biology: Sources and methods for sequence analysis. *Trends in Genetics,* 5**,** 419.

FELSENSTEIN, J. 1982. Numerical Methods for Inferring Evolutionary Trees. *The Quarterly Review of Biology,* 57**,** 379-404.

FENNELL, C. J. & GEZELTER, J. D. 2006. Is the Ewald summation still necessary? Pairwise alternatives to the accepted standard for long-range electrostatics. *J Chem Phys,* 124**,** 234104.

FERRER-COSTA, C., GELPI, J. L., ZAMAKOLA, L., PARRAGA, I., DE LA CRUZ, X. & OROZCO, M. 2005. PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics,* 21**,** 3176-8.

FICKETT, J. W. 1982. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res,* 10**,** 5303-18.

FILLOUX, A. 2010. Secretion signal and protein targeting in bacteria: a biological puzzle. *J Bacteriol,* 192**,** 3847-9.

FINN, R. D., CLEMENTS, J. & EDDY, S. R. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res,* 39**,** W29-37.

FINN, R. D., COGGILL, P., EBERHARDT, R. Y., EDDY, S. R., MISTRY, J., MITCHELL, A. L., POTTER, S. C., PUNTA, M., QURESHI, M., SANGRADOR-VEGAS, A., SALAZAR, G. A., TATE, J. & BATEMAN, A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res,* 44**,** D279-85.

FISCHER, D. & EISENBERG, D. 1997. Assigning folds to the proteins encoded by the genome of Mycoplasma genitalium. *Proc Natl Acad Sci U S A,* 94**,** 11929-34.

FISER, A. & SALI, A. 2003. Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol,* 374**,** 461-91.

FITCH, W. M. 1971. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Zoology,* 20**,** 406-416.

FITCH, W. M. & MARGOLIASH, E. 1967. Construction of phylogenetic trees. *Science,* 155**,** 279-84.

FLOCKNER, H., BRAXENTHALER, M., LACKNER, P., JARITZ, M., ORTNER, M. & SIPPL, M. J. 1995. Progress in fold recognition. *Proteins,* 23**,** 376-86.

FLORKIN, M. 1962. *Isologie, homologie, analogie et convergence en biochimie comparée,* Bruxelles, Académie royale.

FOLKMAN, L., STANTIC, B., SATTAR, A. & ZHOU, Y. 2016. EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models. *J Mol Biol,* 428**,** 1394-405.

FORRELLAD, M. A., KLEPP, L. I., GIOFFRE, A., SABIO Y GARCIA, J., MORBIDONI, H. R., DE LA PAZ SANTANGELO, M., CATALDI, A. A. & BIGI, F. 2013. Virulence factors of the Mycobacterium tuberculosis complex. *Virulence,* 4**,** 3-66.

FREEMAN, Z. N., DORUS, S. & WATERFIELD, N. R. 2013. The KdpD/KdpE two-component system: integrating K(+) homeostasis and virulence. *PLoS Pathog,* 9**,** e1003201.

FRISCH, M. J., TRUCKS, G. W., SCHLEGEL, H. B., SCUSERIA, G. E., ROBB, M. A., CHEESEMAN, J. R., SCALMANI, G., BARONE, V., MENNUCCI, B., PETERSSON, G. A., NAKATSUJI, H., CARICATO, M., LI, X., HRATCHIAN, H. P., IZMAYLOV, A. F., BLOINO, J., ZHENG, G., SONNENBERG, J. L., HADA, M., EHARA, M., TOYOTA, K., FUKUDA, R., HASEGAWA, J., ISHIDA, M., NAKAJIMA, T., HONDA, Y., KITAO, O., NAKAI, H., VREVEN, T., MONTGOMERY JR., J. A., PERALTA, J. E., OGLIARO, F., BEARPARK, M. J., HEYD, J., BROTHERS, E. N., KUDIN, K. N., STAROVEROV, V. N., KOBAYASHI, R., NORMAND, J., RAGHAVACHARI, K., RENDELL, A. P., BURANT, J. C., IYENGAR, S. S., TOMASI, J., COSSI, M., REGA, N., MILLAM, N. J., KLENE, M., KNOX, J. E., CROSS, J. B., BAKKEN, V., ADAMO, C., JARAMILLO, J., GOMPERTS, R., STRATMANN, R. E., YAZYEV, O., AUSTIN, A. J., CAMMI, R., POMELLI, C., OCHTERSKI, J. W., MARTIN, R. L., MOROKUMA, K., ZAKRZEWSKI, V. G., VOTH, G. A., SALVADOR, P., DANNENBERG, J. J., DAPPRICH, S., DANIELS, A. D., FARKAS, Ö., FORESMAN, J. B., ORTIZ, J. V., CIOSLOWSKI, J. & FOX, D. J. 2009. Gaussian 09. Wallingford, CT, USA: Gaussian, Inc.

FRISTENSKY, B. 1986. Improving the efficiency of dot-matrix similarity searches through use of an oligomer table. *Nucleic Acids Res,* 14**,** 597-610.

FUZO, C. A. & DEGREVE, L. 2012. Effect of the thermostat in the molecular dynamics simulation on the folding of the model protein chignolin. *J Mol Model,* 18**,** 2785-94.

GALPERIN, M. Y. & KOONIN, E. V. 2004. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res,* 32**,** 5452-63.

GAMOW, G., RICH, A. & YCAS, M. 1956. The problem of information transfer from the nucleic acids to proteins. *Advances in biological and medical physics,* 4**,** 23-68.

GARG, A. & GUPTA, D. 2008. VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinformatics,* 9**,** 62.

GARG, R., TRIPATHI, D., KANT, S., CHANDRA, H., BHATNAGAR, R. & BANERJEE, N. 2014. A Conserved Hypothetical Protein Rv0574c Is Required for Cell Wall Integrity, Stress Tolerance and Virulence of Mycobacterium tuberculosis. *Infection and Immunity*.

GATLIN, L. L. 1966. The information content of DNA. *J Theor Biol,* 10**,** 281-300.

GAULTON, A., BELLIS, L. J., BENTO, A. P., CHAMBERS, J., DAVIES, M., HERSEY, A., LIGHT, Y., MCGLINCHEY, S., MICHALOVICH, D., AL-LAZIKANI, B. & OVERINGTON, J. P. 2012. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res,* 40**,** D1100-7.

GIBBS, A. J. & MCINTYRE, G. A. 1970. The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur J Biochem,* 16**,** 1-11.

GIFFIN, M. M., MODESTI, L., RAAB, R. W., WAYNE, L. G. & SOHASKEY, C. D. 2012. ald of Mycobacterium tuberculosis Encodes both the Alanine Dehydrogenase and the Putative Glycine Dehydrogenase. *Journal of Bacteriology,* 194**,** 1045-1054.

GIFFIN, M. M., SHI, L., GENNARO, M. L. & SOHASKEY, C. D. 2016. Role of Alanine Dehydrogenase of Mycobacterium tuberculosis during Recovery from Hypoxic Nonreplicating Persistence. *PLoS One,* 11**,** e0155522.

GILL, S. C. & VON HIPPEL, P. H. 1989. Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem,* 182**,** 319-26.

GLICKMAN, M. S., COX, J. S. & JACOBS, W. R., JR. 2000. A novel mycolic acid cyclopropane synthetase is required for cording, persistence, and virulence of Mycobacterium tuberculosis. *Mol Cell,* 5**,** 717-27.

GOMEZ, M., DOUKHAN, L., NAIR, G. & SMITH, I. 1998. sigA is an essential gene in Mycobacterium smegmatis. *Mol Microbiol,* 29**,** 617-28.

GOTOH, O. 1987. Pattern matching of biological sequences with limited storage. *Comput Appl Biosci,* 3**,** 17-20.

GRIBSKOV, M. & BURGESS, R. R. 1986. Sigma factors from E. coli, B. subtilis, phage SP01, and phage T4 are homologous proteins. *Nucleic Acids Res,* 14**,** 6745-63.

GRINDON, C., HARRIS, S., EVANS, T., NOVIK, K., COVENEY, P. & LAUGHTON, C. 2004. Large-scale molecular dynamics simulation of DNA: implementation and validation of the AMBER98 force field in LAMMPS. *Philos Trans A Math Phys Eng Sci,* 362**,** 1373-86.

GROISMAN, E. A. 2001. The pleiotropic two-component regulatory system PhoP-PhoQ. *J Bacteriol,* 183**,** 1835-42.

GRUNBERG, R., NILGES, M. & LECKNER, J. 2007. Biskit--a software platform for structural bioinformatics. *Bioinformatics,* 23**,** 769-70.

GRUNDNER, C., NG, H. L. & ALBER, T. 2005. Mycobacterium tuberculosis protein tyrosine phosphatase PtpB structure reveals a diverged fold and a buried active site. *Structure,* 13**,** 1625-34.

GUIBAS, L. J. & ODLYZKO, A. M. 1980. Long repetitive patterns in random sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete,* 53**,** 241-262.

GUPTA, A., KAPIL, R., DHAKAN, D. B. & SHARMA, V. K. 2014. MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. *PLoS One,* 9**,** e93907.

GURUPRASAD, K., REDDY, B. V. & PANDIT, M. W. 1990. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng,* 4**,** 155-61.

HAFNER, J. 2008. Ab-initio simulations of materials using VASP: Density-functional theory and beyond. *J Comput Chem,* 29**,** 2044-78.

HAMM, G. H. & CAMERON, G. N. 1986. The EMBL data library. *Nucleic Acids Res,* 14**,** 5-9.

HARTH, G. & HORWITZ, M. A. 1999. An inhibitor of exported Mycobacterium tuberculosis glutamine synthetase selectively blocks the growth of pathogenic mycobacteria in axenic culture and in human monocytes: extracellular proteins as potential novel drug targets. *J Exp Med,* 189**,** 1425-36.

HARTH, G., MASLESA-GALIC, S., TULLIUS, M. V. & HORWITZ, M. A. 2005. All four Mycobacterium tuberculosis glnA genes encode glutamine synthetase activities but only GlnA1 is abundantly expressed and essential for bacterial homeostasis. *Mol Microbiol,* 58**,** 1157-72.

HEITHOFF, D. M., SINSHEIMER, R. L., LOW, D. A. & MAHAN, M. J. 1999. An essential role for DNA adenine methylation in bacterial virulence. *Science,* 284**,** 967-70.

HODGMAN, T. C. 1988. A new superfamily of replicative proteins. *Nature,* 333**,** 22-3.

HODGMAN, T. C. 2000. A historical perspective on gene/protein functional assignment. *Bioinformatics,* 16**,** 10-5.

HOLDEN, Z. C., RICHARD, R. M. & HERBERT, J. M. 2013. Periodic boundary conditions for QM/MM calculations: Ewald summation for extended Gaussian basis sets. *J Chem Phys,* 139**,** 244108.

HOLM, L. & ROSENSTROM, P. 2010. Dali server: conservation mapping in 3D. *Nucleic Acids Res,* 38**,** W545-9.

HOLM, L. & SANDER, C. 1996. Mapping the protein universe. *Science,* 273**,** 595-603.

HONDALUS, M. K., BARDAROV, S., RUSSELL, R., CHAN, J., JACOBS, W. R., JR. & BLOOM, B. R. 2000. Attenuation of and protection induced by a leucine auxotroph of Mycobacterium tuberculosis. *Infect Immun,* 68**,** 2888-98.

HOOFT, R. W., VRIEND, G., SANDER, C. & ABOLA, E. E. 1996. Errors in protein structures. *Nature,* 381**,** 272.

HOPP, T. P. & WOODS, K. R. 1981. Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci U S A,* 78**,** 3824-8.

HUET, G., DAFFÉ, M. & SAVES, I. 2005. Identification of the Mycobacterium tuberculosis SUF Machinery as the Exclusive Mycobacterial System of [Fe-S] Cluster Assembly: Evidence for Its Implication in the Pathogen's Survival. *Journal of Bacteriology,* 187**,** 6137-6146.

HUNT, M. C. & ALEXSON, S. E. 2002. The role Acyl-CoA thioesterases play in mediating intracellular lipid metabolism. *Prog Lipid Res,* 41**,** 99-130.

HUNTER, S., APWEILER, R., ATTWOOD, T. K., BAIROCH, A., BATEMAN, A., BINNS, D., BORK, P., DAS, U., DAUGHERTY, L., DUQUENNE, L., FINN, R. D., GOUGH, J., HAFT, D., HULO, N., KAHN, D., KELLY, E., LAUGRAUD, A., LETUNIC, I., LONSDALE, D., LOPEZ, R., MADERA, M., MASLEN, J., MCANULLA, C., MCDOWALL, J., MISTRY, J., MITCHELL, A., MULDER, N., NATALE, D., ORENGO, C., QUINN, A. F., SELENGUT, J. D., SIGRIST, C. J., THIMMA, M., THOMAS, P. D., VALENTIN, F., WILSON, D., WU, C. H. & YEATS, C. 2009. InterPro: the integrative protein signature database. *Nucleic Acids Res,* 37**,** D211-5.

HUNTER, S. W., GAYLORD, H. & BRENNAN, P. J. 1986. Structure and antigenicity of the phosphorylated lipopolysaccharide antigens from the leprosy and tubercle bacilli. *J Biol Chem,* 261**,** 12345-51.

IKAI, A. 1980. Thermostability and aliphatic index of globular proteins. *J Biochem,* 88**,** 1895-8.

ISLAM, S. A. & STERNBERG, M. J. 1989. A relational database of protein structures designed for flexible enquiries about conformation. *Protein Eng,* 2**,** 431-42.

JACKSON, M., PHALEN, S. W., LAGRANDERIE, M., ENSERGUEIX, D., CHAVAROT, P., MARCHAL, G., MCMURRAY, D. N., GICQUEL, B. & GUILHOT, C. 1999. Persistence and protective efficacy of a Mycobacterium tuberculosis auxotroph vaccine. *Infect Immun,* 67**,** 2867-73.

JADHAV, A. A., BARDAPURKAR, J. S. & JAIN, A. 2009. Alkaline phosphatase: Distinguishing between tuberculous and nontuberculous pleural effusion. *Lung India : Official Organ of Indian Chest Society,* 26**,** 77-80.

JAFRI, R. Z., ALI, A., MESSONNIER, N. E., TEVI-BENISSAN, C., DURRHEIM, D., ESKOLA, J., FERMON, F., KLUGMAN, K. P., RAMSAY, M., SOW, S., ZHUJUN, S., BHUTTA, Z. A. & ABRAMSON, J. 2013. Global epidemiology of invasive meningococcal disease. *Popul Health Metr,* 11**,** 17.

JAYARAM, B., DHINGRA, P., MISHRA, A., KAUSHIK, R., MUKHERJEE, G., SINGH, A. & SHEKHAR, S. 2014. Bhageerath-H: a homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins. *BMC Bioinformatics,* 15 Suppl 16**,** S7.

JOHN, B. & SALI, A. 2003. Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res,* 31**,** 3982-92.

JOHNSTON, M. A., GALVAN, I. F. & VILLA-FREIXA, J. 2005. Framework-based design of a new all-purpose molecular simulation application: the Adun simulator. *J Comput Chem,* 26**,** 1647-59.

JORGENSEN, W. L. & TIRADO-RIVES, J. 2005a. Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. *J Comput Chem,* 26**,** 1689-700.

JORGENSEN, W. L. & TIRADO-RIVES, J. 2005b. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc Natl Acad Sci U S A,* 102**,** 6665-70.

JUNG, S. H., PASTAN, I. & LEE, B. 1994. Design of interchain disulfide bonds in the framework region of the Fv fragment of the monoclonal antibody B3. *Proteins,* 19**,** 35-47.

KAITO, C., MORISHITA, D., MATSUMOTO, Y., KUROKAWA, K. & SEKIMIZU, K. 2006. Novel DNA binding protein SarZ contributes to virulence in Staphylococcus aureus. *Mol Microbiol,* 62**,** 1601-17.

KALLBERG, M., MARGARYAN, G., WANG, S., MA, J. & XU, J. 2014. RaptorX server: a resource for template-based protein structure modeling. *Methods Mol Biol,* 1137**,** 17-27.

KAPLAN, W. & LITTLEJOHN, T. G. 2001. Swiss-PDB Viewer (Deep View). *Brief Bioinform,* 2**,** 195-7.

KARPLUS, K. 2009. SAM-T08, HMM-based protein structure prediction. *Nucleic Acids Res,* 37**,** W492-7.

KARPLUS, M. & KURIYAN, J. 2005. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 6679-6685.

KARPLUS, M. & MCCAMMON, J. A. 2002. Molecular dynamics simulations of biomolecules. *Nat Struct Biol,* 9**,** 646-52.

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**,** 845-58.

KELLY, J. M. & MEYER, E. F. 1980. Storage and retrieval of nucleic acid sequence data. *Computers & Chemistry,* 4**,** 107-111.

KESHAVJEE , S. & FARMER , P. E. 2012. Tuberculosis, Drug Resistance, and the History of Modern Medicine. *New England Journal of Medicine,* 367**,** 931-936.

KHOSRAVI, A. D., GOODARZI, H. & ALAVI, S. M. 2012. Detection of genomic mutations in katG, inhA and rpoB genes of Mycobacterium tuberculosis isolates using polymerase chain reaction and multiplex allele-specific polymerase chain reaction. *Braz J Infect Dis,* 16**,** 57-62.

KIM, D. E., CHIVIAN, D. & BAKER, D. 2004. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res,* 32**,** W526-31.

KIMURA, M. & OTA, T. 1972. On the stochastic model for estimation of mutational distance between homologous proteins. *J Mol Evol,* 2**,** 87-90.

KINI, R. M. & EVANS, H. J. 1991. Molecular modeling of proteins: a strategy for energy minimization by molecular mechanics in the AMBER force field. *J Biomol Struct Dyn,* 9**,** 475-88.

KLEIN, P. 1986. Prediction of protein structural class by discriminant analysis. *Biochim Biophys Acta,* 874**,** 205-15.

KLUG, A. & RHODES, D. 1987. 'Zinc fingers': a novel protein motif for nucleic acid recognition. *Trends in Biochemical Sciences,* 12**,** 464-469.

KOCH, R., CANGANELLA, F., HIPPE, H., JAHNKE, K. D. & ANTRANIKIAN, G. 1997. Purification and Properties of a Thermostable Pullulanase from a Newly Isolated Thermophilic Anaerobic Bacterium, Fervidobacterium pennavorans Ven5. *Applied and Environmental Microbiology,* 63**,** 1088-1094.

KOCH, R. E. 1971. The influence of neighboring base pairs upon base-pair substitution mutation rates. *Proc Natl Acad Sci U S A,* 68**,** 773-6.

KOEHL, P. & LEVITT, M. 1999. A brighter future for protein structure prediction. *Nat Struct Biol,* 6**,** 108-11.

KONDO, Y., OHARA, N., SATO, K., YOSHIMURA, M., YUKITAKE, H., NAITO, M., FUJIWARA, T. & NAKAYAMA, K. 2010. Tetratricopeptide repeat protein-associated proteins contribute to the virulence of Porphyromonas gingivalis. *Infect Immun,* 78**,** 2846-56.

KOVACS-SIMON, A., TITBALL, R. W. & MICHELL, S. L. 2011. Lipoproteins of bacterial pathogens. *Infect Immun,* 79**,** 548-61.

KRISTOFFERSON, D. 1987. The BIONET electronic network. *Nature,* 325**,** 555-556.

KROGH, A., LARSSON, B., VON HEIJNE, G. & SONNHAMMER, E. L. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol,* 305**,** 567-80.

KRZYWICKI, A. & SLONIMSKI, P. P. 1967. Formal analysis of protein sequences: I. Specific long-range constraints in pair associations of amino acids. *Journal of Theoretical Biology,* 17**,** 136-158.

KUMAR, S. & VARELA, M. F. 2012. Biochemistry of bacterial multidrug efflux pumps. *Int J Mol Sci,* 13**,** 4484-95.

KYTE, J. & DOOLITTLE, R. F. 1982a. A simple method for displaying the hydropathic character of a protein. *J Mol Biol,* 157**,** 105-32.

KYTE, J. & DOOLITTLE, R. F. 1982b. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology,* 157**,** 105-132.

LANDSCHULZ, W. H., JOHNSON, P. F. & MCKNIGHT, S. L. 1988. The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science,* 240**,** 1759-64.

LASKOWSKI, R. A., RULLMANNN, J. A., MACARTHUR, M. W., KAPTEIN, R. & THORNTON, J. M. 1996. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR,* 8**,** 477-86.

LASTERS, I., WODAK, S. J., ALARD, P. & VAN CUTSEM, E. 1988. Structural principles of parallel beta-barrels in proteins. *Proc Natl Acad Sci U S A,* 85**,** 3338-42.

LAURENZI, M., GINSBERG, A. & SPIGELMAN, M. 2007. Challenges associated with current and future TB treatment. *Infect Disord Drug Targets,* 7**,** 105-19.

LAWRENCE, C., ALTSCHUL, S., BOGUSKI, M., LIU, J., NEUWALD, A. & WOOTTON, J. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science,* 262**,** 208-214.

LEBLANC, C., PRUDHOMME, T., TABOURET, G., RAY, A., BURBAUD, S., CABANTOUS, S., MOUREY, L., GUILHOT, C. & CHALUT, C. 2012. 4′-Phosphopantetheinyl Transferase PptT, a New Drug Target Required for Mycobacterium tuberculosis Growth and Persistence In Vivo. *PLoS Pathogens,* 8**,** e1003097.

LEE, B. & RICHARDS, F. M. 1971. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology,* 55**,** 379-IN4.

LEE, C. 1996. Testing homology modeling on mutant proteins: predicting structural and thermodynamic effects in the Ala98-->Val mutants of T4 lysozyme. *Fold Des,* 1**,** 1-12.

LEE, D., REDFERN, O. & ORENGO, C. 2007. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol,* 8**,** 995-1005.

LEE, Y.-V., WAHAB, H. A. & CHOONG, Y. S. 2015. Potential Inhibitors for Isocitrate Lyase of Mycobacterium tuberculosis and Non-M. tuberculosis: A Summary. *BioMed Research International,* 2015**,** 20.

LESK, A. M. & CHOTHIA, C. 1980. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol,* 136**,** 225-70.

LESZCZYNSKI, J. & ROSE, G. 1986. Loops in globular proteins: a novel category of secondary structure. *Science,* 234**,** 849-855.

LETUNIC, I., DOERKS, T. & BORK, P. 2015. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res,* 43**,** D257-60.

LI, W., CHIANG, Y. H. & COAKER, G. 2013. The HopQ1 effector's nucleoside hydrolase-like domain is required for bacterial virulence in arabidopsis and tomato, but not host recognition in tobacco. *PLoS One,* 8**,** e59684.

LI, X.-X. & ZHOU, X.-N. 2013. Co-infection of tuberculosis and parasitic diseases in humans: a systematic review. *Parasites & Vectors,* 6**,** 79-79.

LIAO, R. Z., YU, J. G. & HIMO, F. 2010. Reaction mechanism of the trinuclear zinc enzyme phospholipase C: a density functional theory study. *J Phys Chem B,* 114**,** 2533-40.

LIN, Y., LI, Q., XIE, L. & XIE, J. 2016. Mycobacterium tuberculosis rv1400c encodes functional lipase/esterase. *Protein Expr Purif.*

LINDAHL, E. R. 2008. Molecular dynamics simulations. *Methods Mol Biol,* 443**,** 3-23.

LIPMAN, D. J. & PEARSON, W. R. 1985. Rapid and sensitive protein similarity searches. *Science,* 227**,** 1435-41.

LIU, J. S., NEUWALD, A. F. & LAWRENCE, C. E. 1995. Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies. *Journal of the American Statistical Association,* 90**,** 1156-1170.

LOBLEY, A., SADOWSKI, M. I. & JONES, D. T. 2009. pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics,* 25**,** 1761-7.

LU, L.-D., SUN, Q., FAN, X.-Y., ZHONG, Y., YAO, Y.-F. & ZHAO, G.-P. 2010. Mycobacterial MazG Is a Novel NTP Pyrophosphohydrolase Involved in Oxidative Stress Response. *The Journal of Biological Chemistry,* 285**,** 28076-28085.

MANCA, C., PAUL, S., BARRY, C. E., 3RD, FREEDMAN, V. H. & KAPLAN, G. 1999. Mycobacterium tuberculosis catalase and peroxidase activities and resistance to oxidative killing in human monocytes in vitro. *Infect Immun,* 67**,** 74-9.

MANGANELLI, R., DUBNAU, E., TYAGI, S., KRAMER, F. R. & SMITH, I. 1999. Differential expression of 10 sigma factor genes in Mycobacterium tuberculosis. *Mol Microbiol,* 31**,** 715-24.

MARCHLER-BAUER, A., ANDERSON, J. B., CHERUKURI, P. F., DEWEESE-SCOTT, C., GEER, L. Y., GWADZ, M., HE, S., HURWITZ, D. I., JACKSON, J. D., KE, Z., LANCZYCKI, C. J., LIEBERT, C. A., LIU, C., LU, F., MARCHLER, G. H., MULLOKANDOV, M., SHOEMAKER, B. A., SIMONYAN, V., SONG, J. S., THIESSEN, P. A., YAMASHITA, R. A., YIN, J. J., ZHANG, D. & BRYANT, S. H. 2005. CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res,* 33**,** D192-6.

MARCHLER-BAUER, A., DERBYSHIRE, M. K., GONZALES, N. R., LU, S., CHITSAZ, F., GEER, L. Y., GEER, R. C., HE, J., GWADZ, M., HURWITZ, D. I., LANCZYCKI, C. J., LU, F., MARCHLER, G. H., SONG, J. S., THANKI, N., WANG, Z., YAMASHITA, R. A., ZHANG, D., ZHENG, C. & BRYANT, S. H. 2015. CDD: NCBI's conserved domain database. *Nucleic Acids Res,* 43**,** D222-6.

MARSHALL, R. D. 1972. Glycoproteins. *Annu Rev Biochem,* 41**,** 673-702.

MARSILI, S., SIGNORINI, G. F., CHELLI, R., MARCHI, M. & PROCACCI, P. 2010. ORAC: a molecular dynamics simulation program to explore free energy surfaces in biomolecular systems at the atomistic level. *J Comput Chem,* 31**,** 1106-16.

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.

MARTIN-LÖF, P. 1966. The definition of random sequences. *Information and Control,* 9**,** 602-619.

MARTTILA, H. J., SOINI, H., HUOVINEN, P. & VILJANEN, M. K. 1996. katG mutations in isoniazid-resistant Mycobacterium tuberculosis isolates recovered from Finnish patients. *Antimicrobial Agents and Chemotherapy,* 40**,** 2187-2189.

MASHALIDIS, E. H., MUKHERJEE, T., ŚLEDŹ, P., MATAK-VINKOVIĆ, D., BOSHOFF, H., ABELL, C. & BARRY, C. E. 2011. Rv2607 from Mycobacterium tuberculosis Is a Pyridoxine 5′-Phosphate Oxidase with Unusual Substrate Specificity. *PLoS One,* 6**,** e27643.

MCCAMMON, J. A., GELIN, B. R. & KARPLUS, M. 1977. Dynamics of folded proteins. *Nature,* 267**,** 585-590.

MEENA, L. S., CHOPRA, P., VISHWAKARMA, R. A. & SINGH, Y. 2013. Biochemical characterization of an S-adenosyl-l-methionine-dependent methyltransferase (Rv0469) of Mycobacterium tuberculosis. *Biol Chem,* 394**,** 871-7.

MI, H., POUDEL, S., MURUGANUJAN, A., CASAGRANDE, J. T. & THOMAS, P. D. 2016. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res,* 44**,** D336-42.

MIRAMONTES, P. 1989. DNA and RNA physicochemical constraints, cellular Automata and Molecular evolution. Los Alamos, New York.

MISURA, K. M., CHIVIAN, D., ROHL, C. A., KIM, D. E. & BAKER, D. 2006. Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proc Natl Acad Sci U S A,* 103**,** 5361-6.

MOHD, S., KRISHNA, B., FAIZAN, A. & MD. IMTAIYAZ, H. 2016. Current Advances in the Identification and Characterization of Putative Drug and Vaccine Targets in the Bacterial Genomes. *Current Topics in Medicinal Chemistry,* 16**,** 1040-1069.

MOLLE, V., SAINT, N., CAMPAGNA, S., KREMER, L., LEA, E., DRAPER, P. & MOLLE, G. 2006. pH-dependent pore-forming activity of OmpATb from Mycobacterium tuberculosis and characterization of the channel by peptidic dissection. *Mol Microbiol,* 61**,** 826-37.

MOODLEY, P., SHAH, N. S., TAYOB, N., CONNOLLY, C., ZETOLA, N., GANDHI, N., FRIEDLAND, G. & STURM, A. W. 2011. Spread of Extensively Drug-Resistant Tuberculosis in KwaZulu-Natal Province, South Africa. *PLoS One,* 6**,** e17513.

MORLOCK, G. P., METCHOCK, B., SIKES, D., CRAWFORD, J. T. & COOKSEY, R. C. 2003. ethA, inhA, and katG Loci of Ethionamide-Resistant Clinical Mycobacterium tuberculosis Isolates. Antimicrobial Agents and Chemotherapy, 47, 3799-3805.

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**,** 2785-91.

MURZIN, A. G., BRENNER, S. E., HUBBARD, T. & CHOTHIA, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol,* 247**,** 536-40.

NAIR, R. & ROST, B. 2008. Protein subcellular localization prediction using artificial intelligence technology. *Methods Mol Biol,* 484**,** 435-63.

NELSON, D. L., COX, M. M. & LEHNINGER, A. L. 2013. *Lehninger principles of biochemistry,* New York :, W.H. Freeman.

NEUDERT, G. & KLEBE, G. 2011. DSX: a knowledge-based scoring function for the assessment of protein-ligand complexes. *J Chem Inf Model,* 51**,** 2731-45.

NEUMANN, J. V. 1966. *Theory of Self-Reproducing Automata*, University of Illinois Press.

NG, P. C. & HENIKOFF, S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res,* 31**,** 3812-4.

NGUYEN, T. T., VIET, M. H. & LI, M. S. 2014. Effects of water models on binding affinity: evidence from all-atom simulation of binding of tamiflu to A/H5N1 neuraminidase. *ScientificWorldJournal,* 2014**,** 536084.

NORBERTO DE SOUZA, O. & ORNSTEIN, R. L. 1999. Molecular dynamics simulations of a protein-protein dimer: particle-mesh Ewald electrostatic model yields far superior results to standard cutoff model. *J Biomol Struct Dyn,* 16**,** 1205-18.

NOSS, E. H., PAI, R. K., SELLATI, T. J., RADOLF, J. D., BELISLE, J., GOLENBOCK, D. T., BOOM, W. H. & HARDING, C. V. 2001. Toll-like receptor 2-dependent inhibition of macrophage class II MHC expression and antigen processing by 19-kDa lipoprotein of Mycobacterium tuberculosis. *J Immunol,* 167**,** 910-8.

NOVOA-APONTE, L. & SOTO OSPINA, C. Y. 2014. Mycobacterium tuberculosis P-Type ATPases: Possible Targets for Drug or Vaccine Development. *BioMed Research International,* 2014**,** 296986.

NOY, T., XU, H. & BLANCHARD, J. S. 2014. Acetylation of acetyl-CoA synthetase from Mycobacterium tuberculosis leads to specific inactivation of the adenylation reaction. *Arch Biochem Biophys,* 550-551**,** 42-9.

OKUGAWA, S., MOAYERI, M., POMERANTSEV, A. P., SASTALLA, I., CROWN, D., GUPTA, P. K. & LEPPLA, S. H. 2012. Lipoprotein biosynthesis by prolipoprotein diacylglyceryl transferase is required for efficient spore germination and full virulence of Bacillus anthracis. *Mol Microbiol,* 83**,** 96-109.

OLEKHNOVICH, I. N. & KADNER, R. J. 2002. DNA-binding activities of the HilC and HilD virulence regulatory proteins of Salmonella enterica serovar Typhimurium. *J Bacteriol,* 184**,** 4148-60.

ORENGO, C. A., MICHIE, A. D., JONES, S., JONES, D. T., SWINDELLS, M. B. & THORNTON, J. M. 1997a. CATH--a hierarchic classification of protein domain structures. *Structure,* 5**,** 1093-108.

ORENGO, C. A., MICHIE, A. D., JONES, S., JONES, D. T., SWINDELLS, M. B. & THORNTON, J. M. 1997b. CATH – a hierarchic classification of protein domain structures. *Structure,* 5**,** 1093-1109.

OUZOUNIS, C. A. & VALENCIA, A. 2003. Early bioinformatics: the birth of a discipline--a personal view. *Bioinformatics,* 19**,** 2176-90.

PAGET, M. S., KANG, J. G., ROE, J. H. & BUTTNER, M. J. 1998. sigmaR, an RNA polymerase sigma factor that modulates expression of the thioredoxin system in response to oxidative stress in Streptomyces coelicolor A3(2). *EMBO J,* 17**,** 5776-82.

PAIN, R. H. & ROBSON, B. 1970. Analysis of the code relating sequence to secondary structure in proteins. *Nature,* 227**,** 62-3.

PARISH, T., GORDHAN, B. G., MCADAM, R. A., DUNCAN, K., MIZRAHI, V. & STOKER, N. G. 1999. Production of mutants in amino acid biosynthesis genes of Mycobacterium tuberculosis by homologous recombination. *Microbiology,* 145 (Pt 12)**,** 3497-503.

PARK, J., KARPLUS, K., BARRETT, C., HUGHEY, R., HAUSSLER, D., HUBBARD, T. & CHOTHIA, C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol,* 284**,** 1201-10.

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.

PEARSON, W. R. 1998. Empirical statistical estimates for sequence similarity searches. *J Mol Biol,* 276**,** 71-84.

PEARSON, W. R. & LIPMAN, D. J. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A,* 85**,** 2444-8.

PEREZ, E., SAMPER, S., BORDAS, Y., GUILHOT, C., GICQUEL, B. & MARTIN, C. 2001. An essential role for phoP in Mycobacterium tuberculosis virulence. *Mol Microbiol,* 41**,** 179-87.

PETERSEN, T. N., BRUNAK, S., VON HEIJNE, G. & NIELSEN, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Meth,* 8**,** 785-786.

PETHE, K., ALONSO, S., BIET, F., DELOGU, G., BRENNAN, M. J., LOCHT, C. & MENOZZI, F. D. 2001. The heparin-binding haemagglutinin of M. tuberculosis is required for extrapulmonary dissemination. *Nature,* 412**,** 190-4.

PHILIPSON, L. 1988. The DNA data libraries. *Nature,* 332**,** 676.

PLINKE, C., WALTER, K., ALY, S., EHLERS, S. & NIEMANN, S. 2011. Mycobacterium tuberculosis embB Codon 306 Mutations Confer Moderately Increased Resistance to Ethambutol In Vitro and In Vivo. *Antimicrobial Agents and Chemotherapy,* 55**,** 2891-2896.

POHL, E., HOLMES, R. K. & HOL, W. G. 1999. Crystal structure of the iron-dependent regulator (IdeR) from Mycobacterium tuberculosis shows both metal binding sites fully occupied. *J Mol Biol,* 285**,** 1145-56.

PONDER, J. W. & RICHARDS, F. M. 1987. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol,* 193**,** 775-91.

PONTIUS, J., RICHELLE, J. & WODAK, S. J. 1996. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J Mol Biol,* 264**,** 121-36.

POOLE, K. 2004. Resistance to beta-lactam antibiotics. *Cell Mol Life Sci,* 61**,** 2200-23.

PRICE, W. N., 2ND, CHEN, Y., HANDELMAN, S. K., NEELY, H., MANOR, P., KARLIN, R., NAIR, R., LIU, J., BARAN, M., EVERETT, J., TONG, S. N., FOROUHAR, F., SWAMINATHAN, S. S., ACTON, T., XIAO, R., LUFT, J. R., LAURICELLA, A., DETITTA, G. T., ROST, B.,

MONTELIONE, G. T. & HUNT, J. F. 2009. Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data. *Nat Biotechnol,* 27**,** 51-7.

PRONK, S., PALL, S., SCHULZ, R., LARSSON, P., BJELKMAR, P., APOSTOLOV, R., SHIRTS, M. R., SMITH, J. C., KASSON, P. M., VAN DER SPOEL, D., HESS, B. & LINDAHL, E. 2013. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics,* 29**,** 845-54.

PROZOROV, A. A., FEDOROVA, I. A., BEKKER, O. B. & DANILENKO, V. N. 2014. The virulence factors of Mycobacterium tuberculosis: Genetic control, new conceptions. *Russian Journal of Genetics,* 50**,** 775-797.

PTITSYN, O. B. 1969. Statistical analysis of the distribution of amino acid residues among helical and non-helical regions in globular proteins. *J Mol Biol,* 42**,** 501-10.

PURKAYASTHA, A., MCCUE, L. A. & MCDONOUGH, K. A. 2002. Identification of a Mycobacterium tuberculosis Putative Classical Nitroreductase Gene Whose Expression Is Coregulated with That of the acr Gene within Macrophages, in Standing versus Shaking Cultures, and under Low Oxygen Conditions. *Infection and Immunity,* 70**,** 1518-1529.

PYM, A. S., SAINT-JOANIS, B. & COLE, S. T. 2002. Effect of katG mutations on the virulence of Mycobacterium tuberculosis and the implication for transmission in humans. *Infect Immun,* 70**,** 4955-60.

QIAN, N. & SEJNOWSKI, T. J. 1988. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol,* 202**,** 865-84.

QUEVILLON, E., SILVENTOINEN, V., PILLAI, S., HARTE, N., MULDER, N., APWEILER, R. & LOPEZ, R. 2005. InterProScan: protein domains identifier. *Nucleic Acids Res,* 33**,** W116-20.

RAHMAN, A. 1964. Correlations in the Motion of Atoms in Liquid Argon. *Physical Review,* 136**,** A405-A411.

RAINWATER, D. L. & KOLATTUKUDY, P. E. 1982. Isolation and Characterization of Acyl Coenzyme A Carboxylases from Mycobacterium tuberculosis and Mycobacterium bovis, Which Produce Multiple Methyl-Branched Mycocerosic Acids. *Journal of Bacteriology,* 151**,** 905-911.

RAMAN, S., HAZRA, R., DASCHER, C. C. & HUSSON, R. N. 2004. Transcription regulation by the Mycobacterium tuberculosis alternative sigma factor SigD and its role in virulence. *J Bacteriol,* 186**,** 6605-16.

RASHID, M., SAHA, S. & RAGHAVA, G. P. 2007. Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. *BMC Bioinformatics,* 8**,** 337.

RASHIN, A. A. 1981. Location of domains in globular proteins. *Nature,* 291**,** 85-87.

RAYNAUD, C., GUILHOT, C., RAUZIER, J., BORDAT, Y., PELICIC, V., MANGANELLI, R., SMITH, I., GICQUEL, B. & JACKSON, M. 2002a. Phospholipases C are involved in the virulence of Mycobacterium tuberculosis. *Mol Microbiol,* 45**,** 203-17.

RAYNAUD, C., PAPAVINASASUNDARAM, K. G., SPEIGHT, R. A., SPRINGER, B., SANDER, P., BOTTGER, E. C., COLSTON, M. J. & DRAPER, P. 2002b. The functions of OmpATb, a pore-forming protein of Mycobacterium tuberculosis. *Mol Microbiol,* 46**,** 191-201.

RAZIN, S., YOGEV, D. & NAOT, Y. 1998. Molecular biology and pathogenicity of mycoplasmas. *Microbiol Mol Biol Rev,* 62**,** 1094-156.

REDDY, M. C., BREDA, A., BRUNING, J. B., SHEREKAR, M., VALLURU, S., THURMAN, C., EHRENFELD, H. & SACCHETTINI, J. C. 2014. Structure, activity, and inhibition of the Carboxyltransferase beta-subunit of acetyl coenzyme A carboxylase (AccD6) from Mycobacterium tuberculosis. *Antimicrob Agents Chemother,* 58**,** 6122-32.

RICHTER, L. & SAVIOLA, B. 2009. The lipF promoter of Mycobacterium tuberculosis is upregulated specifically by acidic pH but not by other stress conditions. *Microbiological research,* 164**,** 228-232.

RIFFO-VASQUEZ, Y., SPINA, D., PAGE, C., TORMAY, P., SINGH, M., HENDERSON, B. & COATES, A. 2004. Effect of Mycobacterium tuberculosis chaperonins on bronchial eosinophilia and hyper-responsiveness in a murine model of allergic inflammation. *Clin Exp Allergy,* 34**,** 712-9.

RIGDEN, DANIEL J. 2008. The histidine phosphatase superfamily: structure and function. *Biochemical Journal,* 409**,** 333-348.

RIGDEN, D. J., FERNÁNDEZ-SUÁREZ, X. M. & GALPERIN, M. Y. 2016. The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection. *Nucleic Acids Research,* 44**,** D1-D6.

RINDI, L., FATTORINI, L., BONANNI, D., IONA, E., FREER, G., TAN, D., DEHO, G., OREFICI, G. & GARZELLI, C. 2002. Involvement of the fadD33 gene in the growth of Mycobacterium tuberculosis in the liver of BALB/c mice. *Microbiology,* 148**,** 3873-80.

ROBERTS, E. A., CLARK, A. & FRIEDMAN, R. L. 2005. Bacterial luciferase is naturally destabilized in Mycobacterium tuberculosis and can be used to monitor changes in gene expression. *FEMS Microbiol Lett,* 243**,** 243-9.

ROJAS, A. V., LIWO, A. & SCHERAGA, H. A. 2007. Molecular dynamics with the United-residue force field: ab initio folding simulations of multichain proteins. *J Phys Chem B,* 111**,** 293-309.

ROOMAN, M. J. & WODAK, S. J. 1988. Identification of predictive sequence motifs limited by protein structure data base size. *Nature,* 335**,** 45-9.

ROSSMANN, M. G. & ARGOS, P. 1980. Three-dimensional coordinates from stereodiagrams of molecular structures. *Acta Crystallographica Section B,* 36**,** 819-823.

ROST, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng,* 12**,** 85-94.

ROY, A., KUCUKURAL, A. & ZHANG, Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc,* 5**,** 725-38.

RUYMGAART, A. P., CARDENAS, A. E. & ELBER, R. 2011. MOIL-opt: Energy-Conserving Molecular Dynamics on a GPU/CPU system. *J Chem Theory Comput,* 7**,** 3072-3082.

RYCHLEWSKI, L., ZHANG, B. & GODZIK, A. 1998. Fold and function predictions for Mycoplasma genitalium proteins. *Fold Des,* 3**,** 229-38.

RYCHLIK, W. 2007. OLIGO 7 primer analysis software. *Methods Mol Biol,* 402**,** 35-60.

RYNDAK, M., WANG, S. & SMITH, I. 2008. PhoP, a key player in Mycobacterium tuberculosis virulence. *Trends Microbiol,* 16**,** 528-34.

SAHA, S. & RAGHAVA, G. P. 2006. VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition. *Genomics Proteomics Bioinformatics,* 4**,** 42-7.

SAIKRISHNAN, K., JEYAKANTHAN, J., VENKATESH, J., ACHARYA, N., SEKAR, K., VARSHNEY, U. & VIJAYAN, M. 2003. Structure of Mycobacterium tuberculosis single-stranded DNA-binding protein. Variability in quaternary structure and its implications. *J Mol Biol,* 331**,** 385-93.

SALEH, M. T. & BELISLE, J. T. 2000. Secretion of an acid phosphatase (SapM) by Mycobacterium tuberculosis that is similar to eukaryotic acid phosphatases. *J Bacteriol,* 182**,** 6850-3.

SALI, A. 1995. Modeling mutations and homologous proteins. *Curr Opin Biotechnol,* 6**,** 437-51.

SALI, A. & OVERINGTON, J. P. 1994. Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci,* 3**,** 1582-96.

SAMBANDAMURTHY, V. K., WANG, X., CHEN, B., RUSSELL, R. G., DERRICK, S., COLLINS, F. M., MORRIS, S. L. & JACOBS, W. R., JR. 2002. A pantothenate auxotroph of Mycobacterium tuberculosis is highly attenuated and protects mice against tuberculosis. *Nat Med,* 8**,** 1171-4.

SAMBROOK, J. & RUSSELL, D. 2001. *Molecular Cloning: A Laboratory Manual*, NY: Cold Spring Harbor Laboratory Press.

SANCHEZ-PADILLA, E., MERKER, M., BECKERT, P., JOCHIMS, F., DLAMINI, T., KAHN, P., BONNET, M. & NIEMANN, S. 2015. Detection of Drug-Resistant Tuberculosis by Xpert MTB/RIF in Swaziland. *New England Journal of Medicine,* 372**,** 1181-1182.

SANCHEZ, R. & SALI, A. 1997. Advances in comparative protein-structure modelling. *Curr Opin Struct Biol,* 7**,** 206-14.

SANCHEZ, R. & SALI, A. 1998. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proc Natl Acad Sci U S A,* 95**,** 13597-602.

SANDER, P., REZWAN, M., WALKER, B., RAMPINI, S. K., KROPPENSTEDT, R. M., EHLERS, S., KELLER, C., KEEBLE, J. R., HAGEMEIER, M., COLSTON, M. J., SPRINGER, B. & BOTTGER, E. C. 2004. Lipoprotein processing is required for virulence of Mycobacterium tuberculosis. *Mol Microbiol,* 52**,** 1543-52.

SANKOFF, D. & SELLERS, P. H. 1973. Shortcuts, diversions, and maximal chainsin partially ordered sets. *Discrete Mathematics,* 4**,** 287-293.

SAPAY, N. & TIELEMAN, D. P. 2011. Combination of the CHARMM27 force field with united-atom lipid force fields. *J Comput Chem,* 32**,** 1400-10.

SAQI, M. A., RUSSELL, R. B. & STERNBERG, M. J. 1998. Misleading local sequence alignments: implications for comparative protein modelling. *Protein Eng,* 11**,** 627-30.

SCHNEIDER, T. D., STORMO, G. D., GOLD, L. & EHRENFEUCHT, A. 1986. Information content of binding sites on nucleotide sequences. *J Mol Biol,* 188**,** 415-31.

SCHOMBURG, I., CHANG, A., EBELING, C., GREMSE, M., HELDT, C., HUHN, G. & SCHOMBURG, D. 2004. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res,* 32**,** D431-3.

SCHWEDE, T., KOPP, J., GUEX, N. & PEITSCH, M. C. 2003. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res,* 31**,** 3381-5.

SELLERS, P. H. 1980. The theory and computation of evolutionary distances: Pattern recognition. *Journal of Algorithms,* 1**,** 359-373.

SENARATNE, R. H., MOBASHERI, H., PAPAVINASASUNDARAM, K. G., JENNER, P., LEA, E. J. & DRAPER, P. 1998. Expression of a gene for a porin-like protein of the OmpA family from Mycobacterium tuberculosis H37Rv. *J Bacteriol,* 180**,** 3541-7.

SHAHBAAZ, M., BISETTY, K., AHMAD, F. & HASSAN, M. I. 2015a. In silico approaches for the identification of virulence candidates amongst hypothetical proteins of Mycoplasma pneumoniae 309. *Computational Biology and Chemistry,* 59, Part A**,** 67-80.

SHAHBAAZ, M., BISETTY, K., AHMAD, F. & HASSAN, M. I. 2015b. Towards New Drug Targets? Function Prediction of Putative Proteins of Neisseria meningitidis MC58 and Their Virulence Characterization. *OMICS,* 19**,** 416-34.

SHANNON, C. E. & WEAVER, W. 1963. *A Mathematical Theory of Communication*, University of Illinois Press.

SHEN, M. Y. & SALI, A. 2006. Statistical potential for assessment and prediction of protein structures. *Protein Sci,* 15**,** 2507-24.

SHI, S. & EHRT, S. 2006. Dihydrolipoamide acyltransferase is critical for Mycobacterium tuberculosis pathogenesis. *Infect Immun,* 74**,** 56-63.

SHOICHET, B. K., BAASE, W. A., KUROKI, R. & MATTHEWS, B. W. 1995. A relationship between protein stability and protein function. *Proceedings of the National Academy of Sciences of the United States of America,* 92**,** 452-456.

SILLITOE, I., LEWIS, T. E., CUFF, A., DAS, S., ASHFORD, P., DAWSON, N. L., FURNHAM, N., LASKOWSKI, R. A., LEE, D., LEES, J. G., LEHTINEN, S., STUDER, R. A., THORNTON, J. & ORENGO, C. A. 2015. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res,* 43**,** D376-81.

SIMOSSIS, V. A. & HERINGA, J. 2005. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res,* 33**,** W289-94.

SKJOT, R. L., OETTINGER, T., ROSENKRANDS, I., RAVN, P., BROCK, I., JACOBSEN, S. & ANDERSEN, P. 2000. Comparative evaluation of low-molecular-mass proteins from Mycobacterium tuberculosis identifies members of the ESAT-6 family as immunodominant T-cell antigens. *Infect Immun,* 68**,** 214-20.

SMITH, D. A., PARISH, T., STOKER, N. G. & BANCROFT, G. J. 2001. Characterization of auxotrophic mutants of Mycobacterium tuberculosis and their potential as vaccine candidates. *Infect Immun,* 69**,** 1142-50.

SMITH, I. 2003. Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence. *Clin Microbiol Rev,* 16**,** 463-96.

SMITH, T. F. & WATERMAN, M. S. 1981a. Comparison of biosequences. *Advances in Applied Mathematics,* 2**,** 482-489.

SMITH, T. F. & WATERMAN, M. S. 1981b. Identification of common molecular subsequences. *Journal of Molecular Biology,* 147**,** 195-197.

SMITH, T., WOLFF, K. A. & NGUYEN, L. 2013. Molecular Biology of Drug Resistance in Mycobacterium tuberculosis. Current topics in microbiology and immunology, 374, 53-80.

SMITH, W. & FORESTER, T. R. 1996. DL_POLY_2.0: a general-purpose parallel molecular dynamics simulation package. *J Mol Graph,* 14**,** 136-41.

SODING, J., BIEGERT, A. & LUPAS, A. N. 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res,* 33**,** W244-8.

SONNENBERG, M. G. & BELISLE, J. T. 1997. Definition of Mycobacterium tuberculosis culture filtrate proteins by two-dimensional polyacrylamide gel electrophoresis, N-terminal amino acid sequencing, and electrospray mass spectrometry. *Infect Immun,* 65**,** 4515-24.

SOWAJASSATAKUL, A., PRAMMANANAN, T., CHAIPRASERT, A. & PHUNPRUCH, S. 2014. Molecular characterization of amikacin, kanamycin and capreomycin resistance in M/XDR-TB strains isolated in Thailand. BMC Microbiology, 14, 165.

SREEVATSAN, S., PAN, X., STOCKBAUER, K. E., WILLIAMS, D. L., KREISWIRTH, B. N. & MUSSER, J. M. 1996a. Characterization of rpsL and rrs mutations in streptomycin-resistant Mycobacterium tuberculosis isolates from diverse geographic localities. *Antimicrob Agents Chemother,* 40**,** 1024-6.

SREEVATSAN, S., PAN, X., STOCKBAUER, K. E., WILLIAMS, D. L., KREISWIRTH, B. N. & MUSSER, J. M. 1996b. Characterization of rpsL and rrs mutations in streptomycin-resistant Mycobacterium tuberculosis isolates from diverse geographic localities. *Antimicrobial Agents and Chemotherapy,* 40**,** 1024-1026.

SREEVATSAN, S., STOCKBAUER, K. E., PAN, X., KREISWIRTH, B. N., MOGHAZEH, S. L., JACOBS, W. R., TELENTI, A. & MUSSER, J. M. 1997. Ethambutol resistance in Mycobacterium tuberculosis: critical role of embB mutations. *Antimicrobial Agents and Chemotherapy,* 41**,** 1677-1681.

STADEN, R. 1984. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res,* 12**,** 505-19.

STANLEY, S. A. & COX, J. S. 2013. Host–Pathogen Interactions During Mycobacterium tuberculosis infections. *In:* PIETERS, J. & MCKINNEY, D. J. (eds.) *Pathogenesis of Mycobacterium*

tuberculosis and its Interaction with the Host Organism. Berlin, Heidelberg: Springer Berlin Heidelberg.

STILLINGER, F. H. & RAHMAN, A. 1974. Improved simulation of liquid water by molecular dynamics. *The Journal of Chemical Physics,* 60**,** 1545-1557.

STORMO, G. D., SCHNEIDER, T. D., GOLD, L. & EHRENFEUCHT, A. 1982. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res,* 10**,** 2997-3011.

TATAR, L. D., MAROLDA, C. L., POLISCHUK, A. N., VAN LEEUWEN, D. & VALVANO, M. A. 2007. An Escherichia coli undecaprenyl-pyrophosphate phosphatase implicated in undecaprenyl phosphate recycling. *Microbiology,* 153**,** 2518-29.

TATENO, Y., IMANISHI, T., MIYAZAKI, S., FUKAMI-KOBAYASHI, K., SAITOU, N., SUGAWARA, H. & GOJOBORI, T. 2002. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res,* 30**,** 27-30.

TAYLOR, W. R. 1986. The classification of amino acid conservation. *Journal of Theoretical Biology,* 119**,** 205-218.

TELENTI, A., IMBODEN, P., MARCHESI, F., MATTER, L., SCHOPFER, K., BODMER, T., LOWRIE, D., COLSTON, M. J. & COLE, S. 1993. Originally published as Volume 1, Issue 8846Detection of rifampicin-resistance mutations in Mycobacterium tuberculosis. *The Lancet,* 341**,** 647-651.

TENG, S., SRIVASTAVA, A. K. & WANG, L. 2010. Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *BMC Genomics,* 11 Suppl 2**,** S5.

TETTELIN, H., SAUNDERS, N. J., HEIDELBERG, J., JEFFRIES, A. C., NELSON, K. E., EISEN, J. A., KETCHUM, K. A., HOOD, D. W., PEDEN, J. F., DODSON, R. J., NELSON, W. C., GWINN, M. L., DEBOY, R., PETERSON, J. D., HICKEY, E. K., HAFT, D. H., SALZBERG, S. L., WHITE, O., FLEISCHMANN, R. D., DOUGHERTY, B. A., MASON, T., CIECKO, A., PARKSEY, D. S., BLAIR, E., CITTONE, H., CLARK, E. B., COTTON, M. D., UTTERBACK, T. R., KHOURI, H., QIN, H., VAMATHEVAN, J., GILL, J., SCARLATO, V., MASIGNANI, V., PIZZA, M., GRANDI, G., SUN, L., SMITH, H. O., FRASER, C. M., MOXON, E. R., RAPPUOLI, R. & VENTER, J. C. 2000. Complete genome sequence of Neisseria meningitidis serogroup B strain MC58. *Science,* 287**,** 1809-15.

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**,** 4673-80.

THORNTON, J. M. 1981. Disulphide bridges in globular proteins. *J Mol Biol,* 151**,** 261-87.

TINOCO, I., JR., UHLENBECK, O. C. & LEVINE, M. D. 1971. Estimation of secondary structure in ribonucleic acids. *Nature,* 230**,** 362-7.

TSARA, V., SERASLI, E. & CHRISTAKI, P. 2009. Problems in diagnosis and treatment of tuberculosis infection. *Hippokratia,* 13**,** 20-22.

UDA, A., SHARMA, N., TAKIMOTO, K., DEYU, T., KOYAMA, Y., PARK, E.-S., FUJITA, O., HOTTA, A. & MORIKAWA, S. 2016. Pullulanase Is Necessary for the Efficient Intracellular Growth of <italic>Francisella tularensis</italic>. *PLoS One,* 11**,** e0159740.

ULLMAN, J. D., AHO, A. V. & HIRSCHBERG, D. S. 1976. Bounds on the Complexity of the Longest Common Subsequence Problem. *J. ACM,* 23**,** 1-12.

UNGER, R., HAREL, D., WHERLAND, S. & SUSSMAN, J. L. 1989. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins,* 5**,** 355-73.

VALIM, A. R. M., ROSSETTI, M. L. R., RIBEIRO, M. O. & ZAHA, A. 2000. Mutations in the rpoB Gene of Multidrug-Resistant Mycobacterium tuberculosis Isolates from Brazil. *Journal of Clinical Microbiology,* 38**,** 3119-3122.

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**,** 1701-18.

VANDAL, O. H., NATHAN, C. F. & EHRT, S. 2009. Acid resistance in Mycobacterium tuberculosis. *J Bacteriol,* 191**,** 4714-21.

VANOMMESLAEGHE, K., HATCHER, E., ACHARYA, C., KUNDU, S., ZHONG, S., SHIM, J., DARIAN, E., GUVENCH, O., LOPES, P., VOROBYOV, I. & MACKERELL, A. D., JR. 2010. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem,* 31**,** 671-90.

VASQUEZ, M. 1996. Modeling side-chain conformation. *Curr Opin Struct Biol,* 6**,** 217-21.

VLAHOVICEK, K., KAJAN, L., AGOSTON, V. & PONGOR, S. 2005. The SBASE domain sequence resource, release 12: prediction of protein domain-architecture using support vector machines. *Nucleic Acids Res,* 33**,** D223-5.

VON HEIJNE, G. 1981. On the hydrophobic nature of signal sequences. *Eur J Biochem,* 116**,** 419-22.

WALKER, J. E., SARASTE, M., RUNSWICK, M. J. & GAY, N. J. 1982. Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J,* 1**,** 945-51.

WANG, F., LANGLEY, R., GULTEN, G., WANG, L. & SACCHETTINI, J. C. 2007a. Identification of a type III thioesterase reveals the function of an operon crucial for Mtb virulence. *Chem Biol,* 14**,** 543-51.

WANG, Q., CANUTESCU, A. A. & DUNBRACK, R. L., JR. 2008. SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nat Protoc,* 3**,** 1832-47.

WANG, R., YIN, Y. J., WANG, F., LI, M., FENG, J., ZHANG, H. M., ZHANG, J. P., LIU, S. J. & CHANG, W. R. 2007b. Crystal structures and site-directed mutagenesis of a mycothiol-dependent enzyme reveal a novel folding and molecular basis for mycothiol-mediated maleylpyruvate isomerization. *J Biol Chem,* 282**,** 16288-294.

WATSON, J. D. & CRICK, F. H. 1953. Genetical implications of the structure of deoxyribonucleic acid. *Nature,* 171**,** 964-7.

WAYNE, L. G. 1994. Dormancy of Mycobacterium tuberculosis and latency of disease. *Eur J Clin Microbiol Infect Dis,* 13**,** 908-14.

WAYNE, L. G. & LIN, K. Y. 1982. Glyoxylate metabolism and adaptation of Mycobacterium tuberculosis to survival under anaerobic conditions. *Infect Immun,* 37**,** 1042-9.

WEISBURG, W. G., TULLY, J. G., ROSE, D. L., PETZEL, J. P., OYAIZU, H., YANG, D., MANDELCO, L., SECHREST, J., LAWRENCE, T. G. & VAN ETTEN, J. 1989. A phylogenetic analysis of the mycoplasmas: basis for their classification. *Journal of Bacteriology,* 171**,** 6455-6467.

WEISS, L. A., HARRISON, P. G., NICKELS, B. E., GLICKMAN, M. S., CAMPBELL, E. A., DARST, S. A. & STALLINGS, C. L. 2012. Interaction of CarD with RNA Polymerase Mediates Mycobacterium tuberculosis Viability, Rifampin Resistance, and Pathogenesis. *Journal of Bacteriology,* 194**,** 5621-5631.

WEST, N. P., CERGOL, K. M., XUE, M., RANDALL, E. J., BRITTON, W. J. & PAYNE, R. J. 2011. Inhibitors of an essential mycobacterial cell wall lipase (Rv3802c) as tuberculosis drug leads. *Chemical Communications,* 47**,** 5166-5168.

WILLIAMS, A., GÜTHLEIN, C., BERESFORD, N., BÖTTGER, E. C., SPRINGER, B. & DAVIS, E. O. 2011. UvrD2 Is Essential in Mycobacterium tuberculosis, but Its Helicase Activity Is Not Required. *Journal of Bacteriology,* 193**,** 4487-4494.

WILSON, D., PETHICA, R., ZHOU, Y., TALBOT, C., VOGEL, C., MADERA, M., CHOTHIA, C. & GOUGH, J. 2009. SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res,* 37**,** D380-6.

WORTH, C. L., PREISSNER, R. & BLUNDELL, T. L. 2011. SDM--a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res,* 39**,** W215-22.

WU, S. & ZHANG, Y. 2008. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins,* 72**,** 547-56.

WUTHRICH, K. 1990. Protein structure determination in solution by NMR spectroscopy. *J Biol Chem,* 265**,** 22059-62.

XU, D. & ZHANG, Y. 2012. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins,* 80**,** 1715-35.

XU, X. L. & GRANT, G. A. 2014. Regulation of Mycobacterium tuberculosis D-3-phosphoglycerate dehydrogenase by phosphate-modulated quaternary structure dynamics and a potential role for polyphosphate in enzyme regulation. *Biochemistry,* 53**,** 4239-49.

YAMAMOTO, K. & YOSHIKURA, H. 1986. A new representation of protein structure: vector diagram. *Comput Appl Biosci,* 2**,** 83-8.

YANG, Y., AUGUIN, D., DELBECQ, S., DUMAS, E., MOLLE, G., MOLLE, V., ROUMESTAND, C. & SAINT, N. 2011a. Structure of the Mycobacterium tuberculosis OmpATb protein: a model of an oligomeric channel in the mycobacterial cell wall. *Proteins,* 79**,** 645-61.

YANG, Y., FARAGGI, E., ZHAO, H. & ZHOU, Y. 2011b. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics,* 27**,** 2076-82.

YANG, Y., LABESSE, G., CARRÈRE-KREMER, S., ESTEVES, K., KREMER, L., COHEN-GONSAUD, M. & BLANC-POTARD, A.-B. 2012. The C-Terminal Domain of the Virulence Factor MgtC Is a Divergent ACT Domain. *Journal of Bacteriology,* 194**,** 6255-6263.

YANO, T., RAHIMIAN, M., ANEJA, K. K., SCHECHTER, N. M., RUBIN, H. & SCOTT, C. P. 2014. Mycobacterium tuberculosis type II NADH-menaquinone oxidoreductase catalyzes electron transfer through a two-site ping-pong mechanism and has two quinone-binding sites. *Biochemistry,* 53**,** 1179-90.

YATIME, L., BUCH-PEDERSEN, M. J., MUSGAARD, M., MORTH, J. P., LUND WINTHER, A. M., PEDERSEN, B. P., OLESEN, C., ANDERSEN, J. P., VILSEN, B., SCHIOTT, B., PALMGREN, M. G., MOLLER, J. V., NISSEN, P. & FEDOSOVA, N. 2009. P-type ATPases as drug targets: tools for medicine and science. *Biochim Biophys Acta,* 1787**,** 207-20.

YUAN, Y., CRANE, D. D., SIMPSON, R. M., ZHU, Y. Q., HICKEY, M. J., SHERMAN, D. R. & BARRY, C. E., 3RD 1998. The 16-kDa alpha-crystallin (Acr) protein of Mycobacterium tuberculosis is required for growth in macrophages. *Proc Natl Acad Sci U S A,* 95**,** 9578-83.

ZAHRT, T. C. & DERETIC, V. 2001. Mycobacterium tuberculosis signal transduction system required for persistent infections. *Proc Natl Acad Sci U S A,* 98**,** 12706-11.

ZHANG, M., WANG, J. D., LI, Z. F., XIE, J., YANG, Y. P., ZHONG, Y. & WANG, H. H. 2005. Expression and characterization of the carboxyl esterase Rv3487c from Mycobacterium tuberculosis. *Protein Expr Purif,* 42**,** 59-66.

ZHANG, M. M., ONG, C. L., WALKER, M. J. & MCEWAN, A. G. 2016. Defence against methylglyoxal in Group A Streptococcus: a role for Glyoxylase I in bacterial virulence and survival in neutrophils? *Pathog Dis,* 74.

ZHANG, Y. & SKOLNICK, J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*, 33**,** 2302-9.

ZHU, P. P., LI, W. C., ZHONG, Z. J., DENG, E. Z., DING, H., CHEN, W. & LIN, H. 2015. Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition. *Mol Biosyst*, 11**,** 558-63.

ZUCKERKANDL, E. & PAULING, L. 1965a. Evolutionary Divergence and Convergence in Proteins. *Evolving Genes and Proteins.* Academic Press.

ZUCKERKANDL, E. & PAULING, L. 1965b. Molecules as documents of evolutionary history. *J Theor Biol*, 8**,** 357-66.

Appendix A

Mutation studies

Figure A.1: The 3-D structure of arabinosyl transferase B involved in Ethambutol (EMB) resistance.

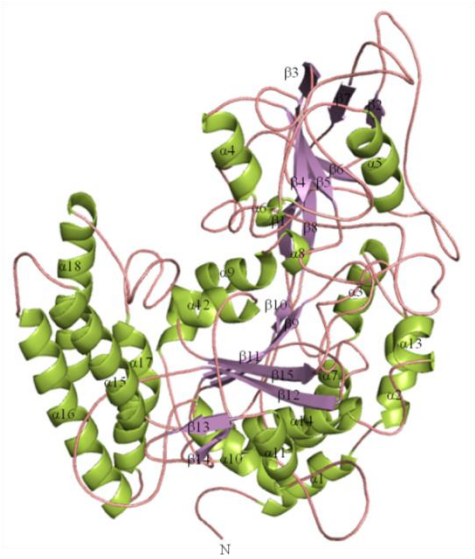
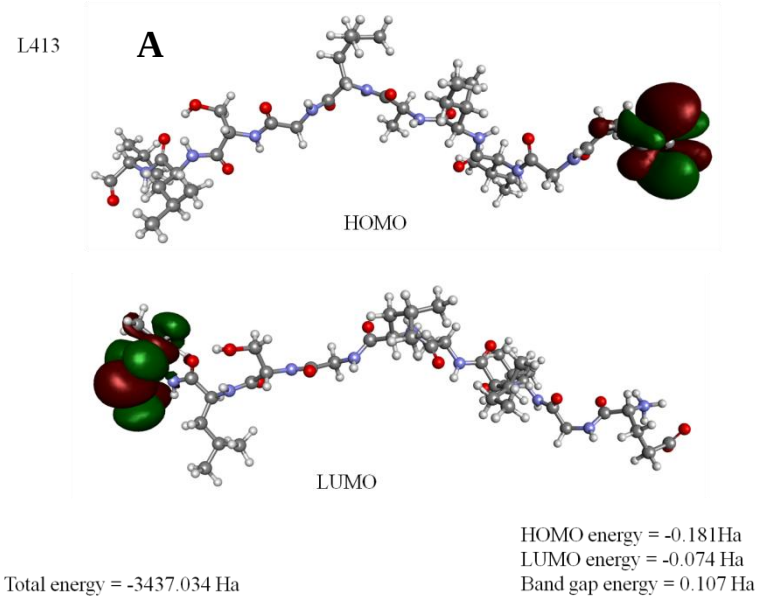
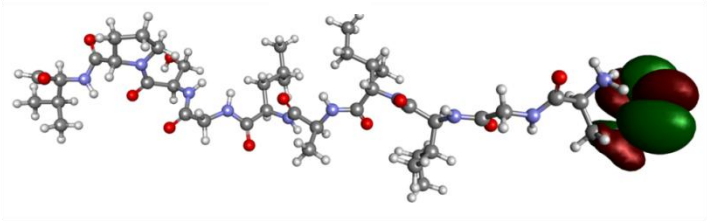


Figure A.2: Ethambutol (EMB) resistance-associated mutation. **(A)** The quantum calculations of the wild type Leu413**. (B)** Pro413 which is an experimentally validated mutation that may result in severe infection conditions in patients showed increase in the energy. **(C)** The quantum calculations of the wild type Trp332**. (D)** Asn332 is an *in silico* predicted mutation that may cause the EMB resistance in *M. tuberculosis* showed increase in energy.

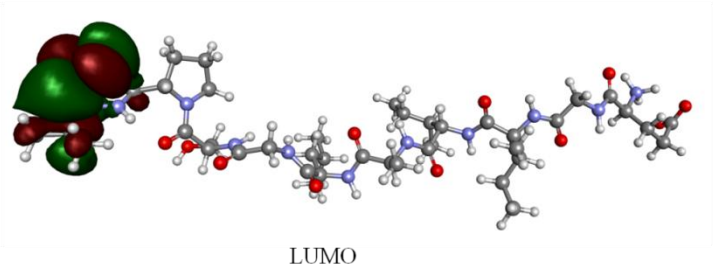


B

P413



HOMO

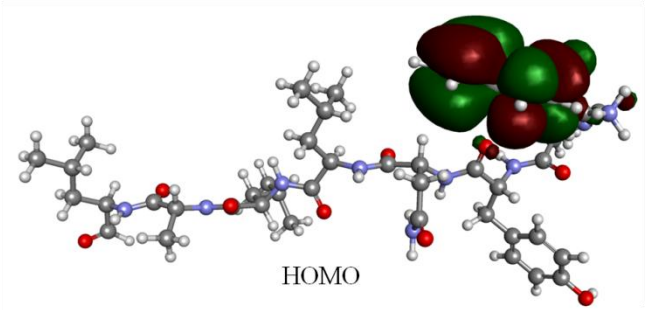


LUMO

HOMO energy = -0.181 Ha
LUMO energy = -0.04 Ha
Band gap energy = 0.141 Ha

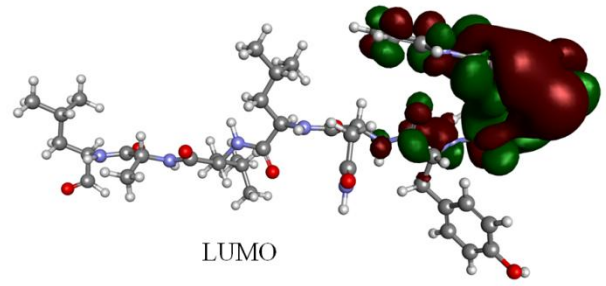
Total energy = -3394.097 Ha

W332



HOMO

C



LUMO

HOMO energy = -0.287 Ha
LUMO energy = -0.117 Ha
Band gap energy = 0.170 Ha

Total energy = -3681.87 Ha

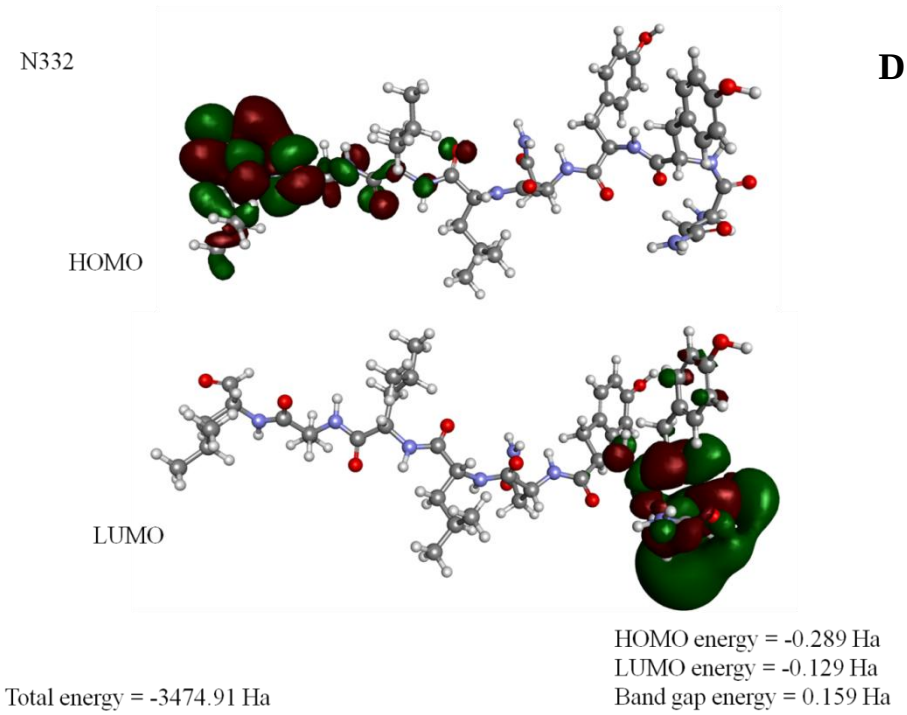
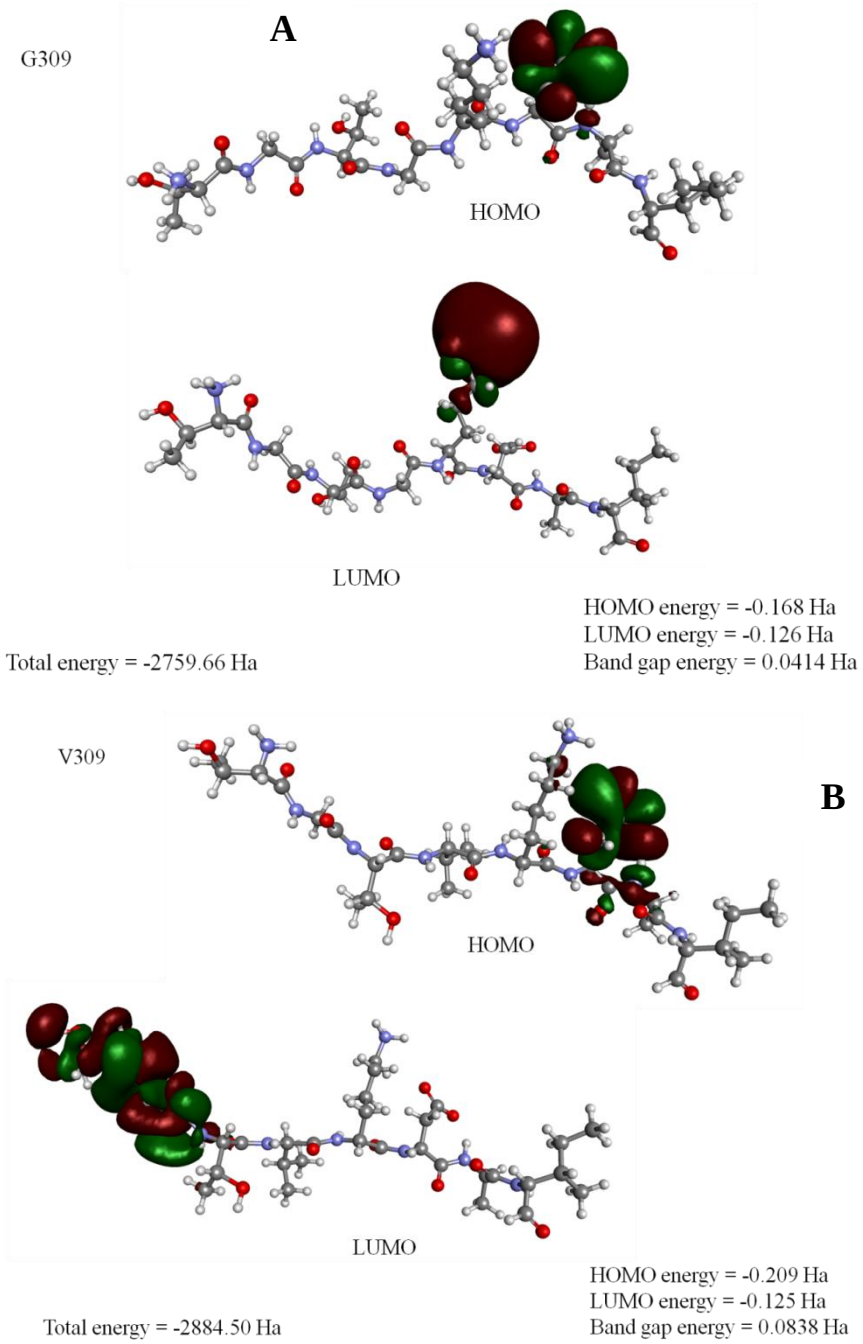


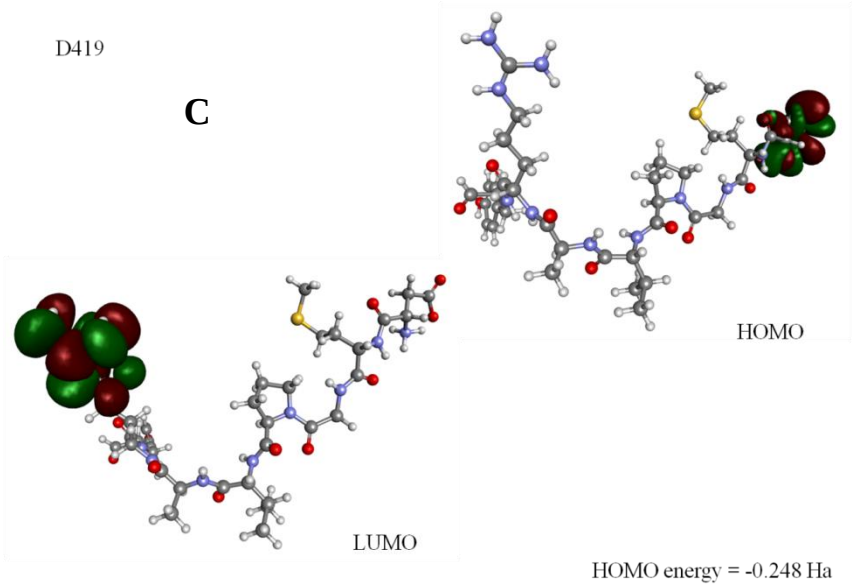
Figure A.3: The 3-D structure of catalase - peroxidase (KatG) enzyme involved in isoniazid resistance

Figure A.4: Point-mutations associated with isoniazid resistance. **(A)** The quantum calculations of the wild type Gly309**. (B)** Gly309Val which is an experimentally validated mutation that may result in severe infection conditions in patients showed increase in the energy. **(C)** The quantum calculations of the wild type Asp419**. (D)** Asp419Trp is an *in silico* predicted mutation that may cause the INH resistance in *M. tuberculosis* showed decrease in energy.



D419

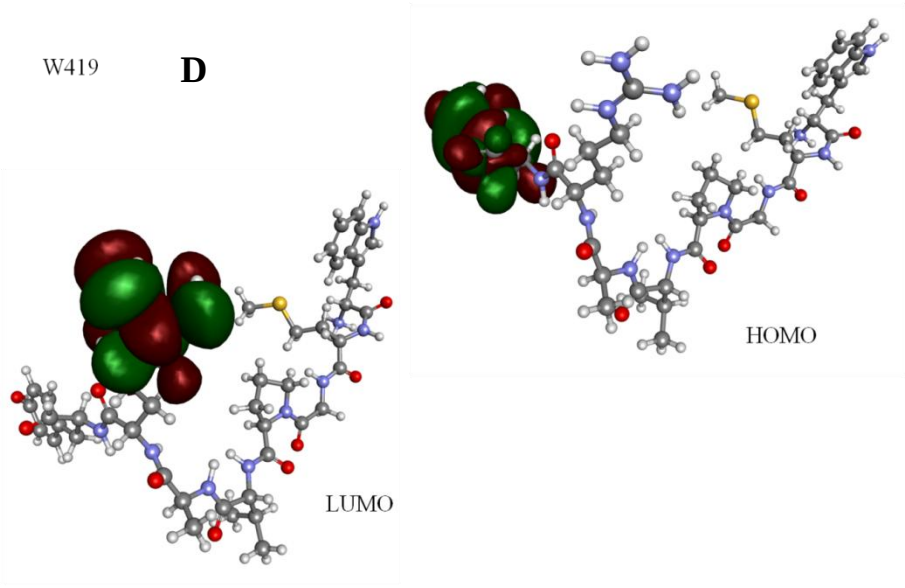
C



HOMO energy = -0.248 Ha
LUMO energy = -0.187 Ha
Band gap energy = 0.0601 Ha

Total energy = -3535.86 Ha

W419 **D**

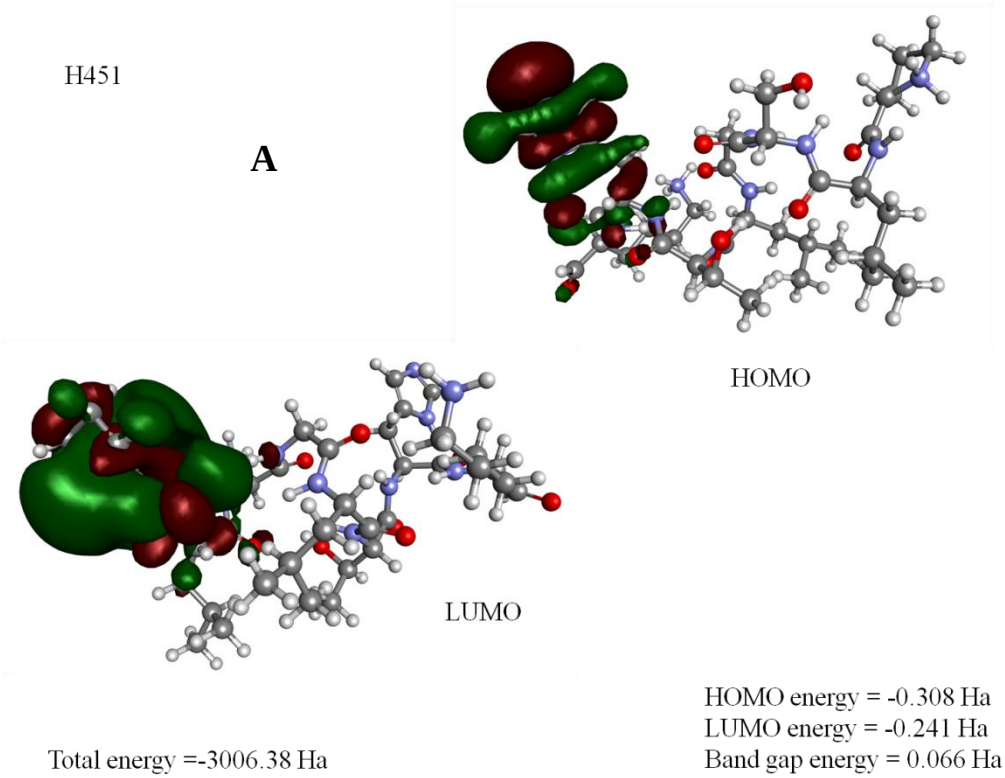


HOMO energy = -0.31 Ha
LUMO energy = -0.21Ha
Band gap energy = 0.105 Ha

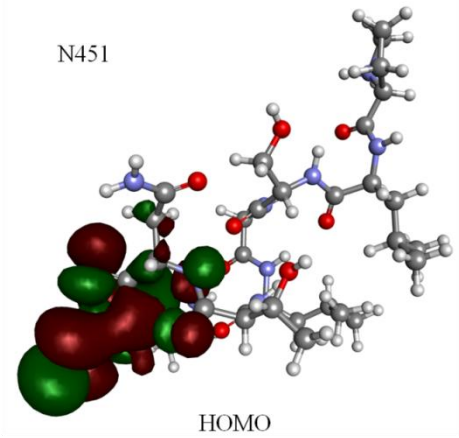
Total energy = -3722.18 Ha

Figure A.5: The 3-D structure of catalase - peroxidase (KatG) enzyme involved in isoniazid resistance

Figure A.6: Point-mutations leading to Rifampicin resistance in *M. tuberculosis*. **(A & B)** His451Asn is an experimentally validated mutation. **(C & D)** The Leu500Lys is computationally predicted mutation that may increases the severity of the infection in patients.

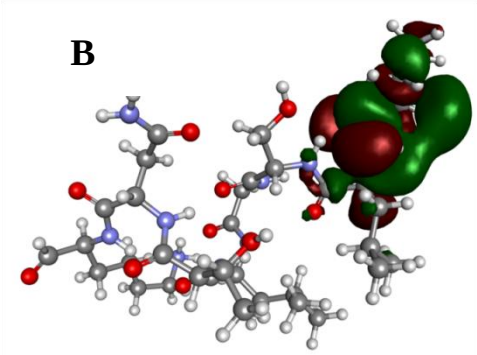


N451



HOMO

B



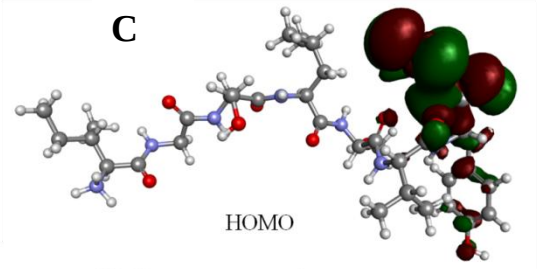
LUMO

Total energy =-2946.56 Ha

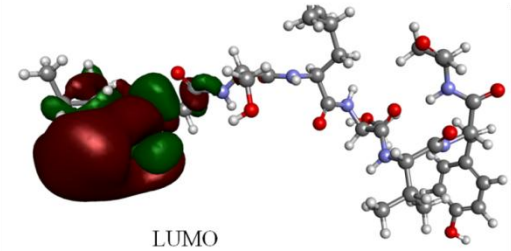
HOMO energy = -0.33 Ha
LUMO energy = -0.203 Ha
Band gap energy = 0.129 Ha

L500

C



HOMO



LUMO

Total energy = -2869.12 Ha

HOMO energy = -0.259 Ha
LUMO energy = -0.138 Ha
Band gap energy = 0.121 Ha

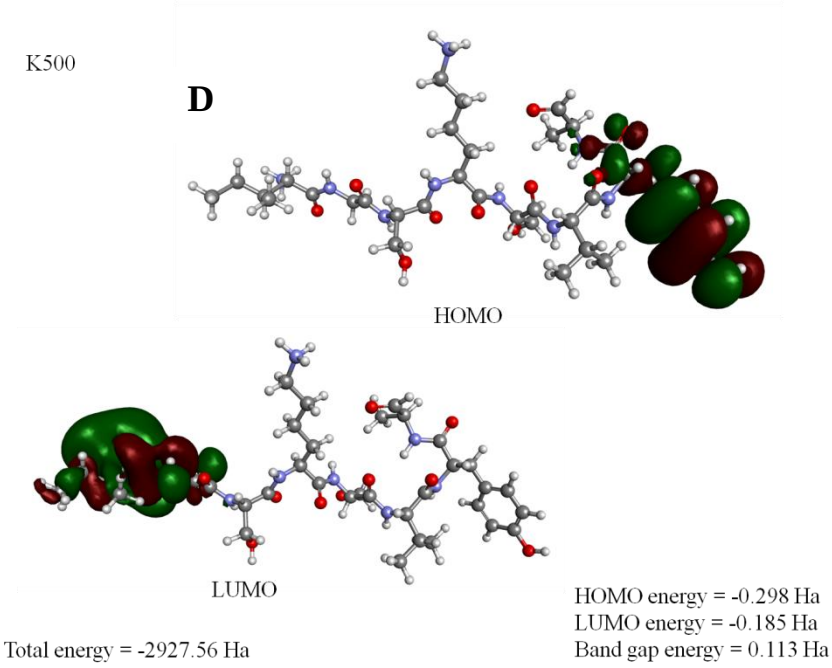


Figure A.7: The 3-D structure of ribosomal protein expressed by rpsL gene involved in Streptomycin resistance

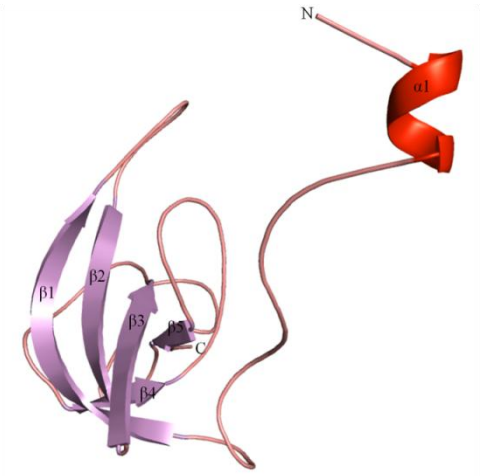
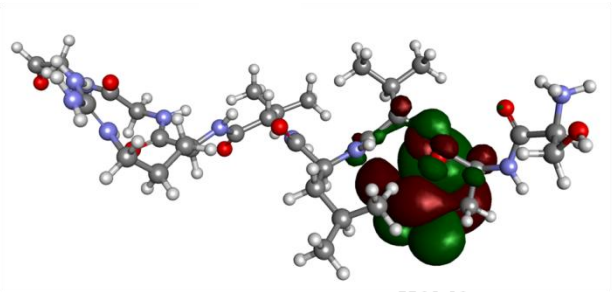


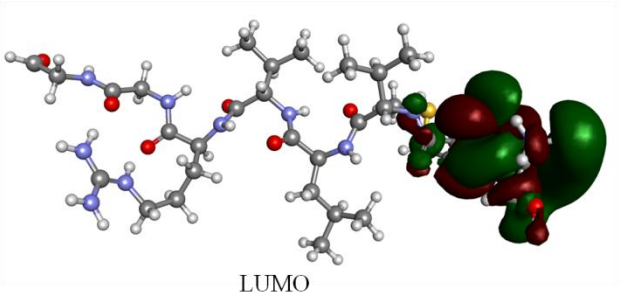
Figure A.8: The substitution in the amino acid residues leading to Streptomycin resistance in *M. tuberculosis*. **(A & B)** Val80Gly is an experimentally validated mutation in rpsL gene. **(C & D)** The Leu102Trp is predicted mutation in the rpsL gene that may enhances the severity of the disease.

A

V80



HOMO



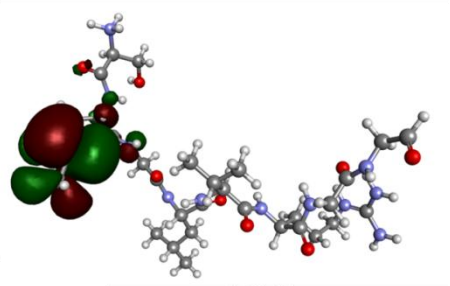
LUMO

HOMO energy = -0.34 Ha
LUMO energy = -0.19 Ha
Band gap energy = 0.15 Ha

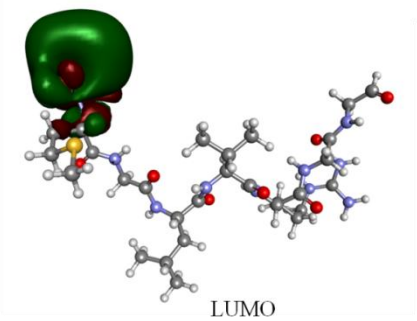
Total energy = -3176.45 Ha

B

G80



HOMO



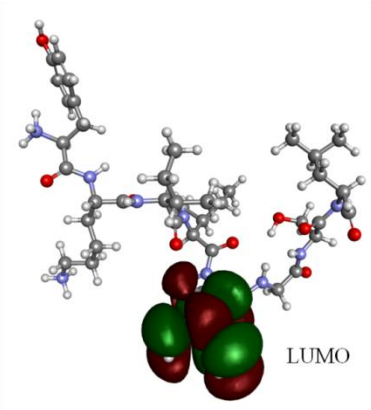
LUMO

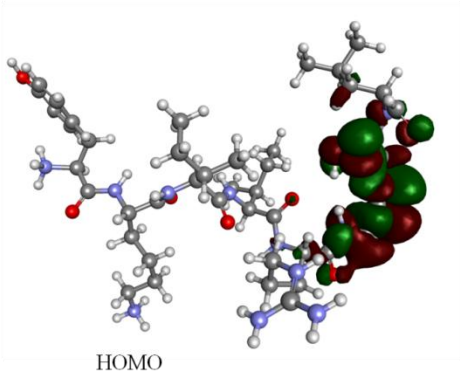
HOMO energy = -0.317 Ha
LUMO energy = -0.21 Ha
Band gap energy = 0.107 Ha

Total energy = -3051.13 Ha

L102

C



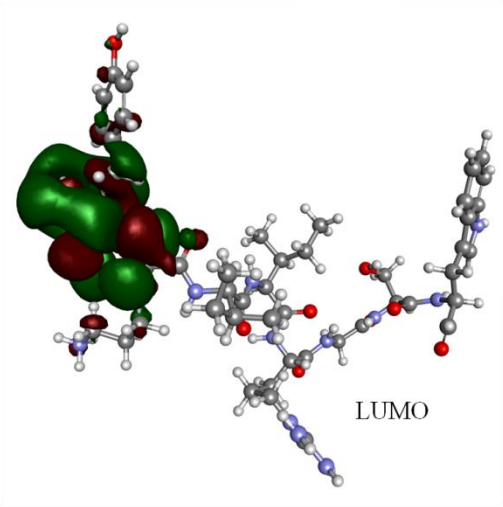


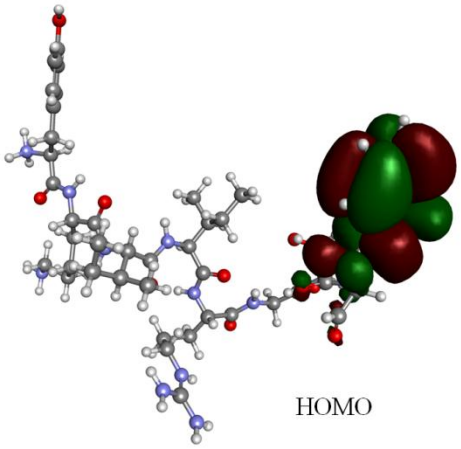
Total energy = -3314.16 Ha

HOMO energy = -0.38 Ha
LUMO energy = -0.269 Ha
Band gap energy = 0.116 Ha

W102

D





Total energy = -3573.95 Ha

HOMO energy = -0.31 Ha
LUMO energy = -0.27 Ha
Band gap energy = 0.042 Ha

Appendix B

In silico predicted drug resistance associated mutations

Table B.1: List of predicted mutations that may result in Ethambutol resistance in the *M. tuberculosis*.

S. No.	PMut[a] + SIFT [b] Pathological mutations	RI	Protein Stability							
			I-MUTANT 2.0 (Sequence based)		MuStab server		EASE-MM		SDM server	
			Predictions (Stability)	ΔΔG	Predictions (Stability)	confidence	Predictions (Stability)	ΔΔG	Predictions (Stability/effect)	ΔΔG
1)	**Cys4Lys**	8	(-)	-1.56	(+)	22.68%	Neutral	0.11	N + N	0.17
2)	**Arg8Trp**	8	(-)	-0.46	(+)	30.36%	LS	0.96	S + N	1.29
3)	**Arg27Cys**	8	(-)	-0.11	(-)	82.21%	DS	-1.23	N + N	-0.36
4)	**Ala30Asn**	8	(-)	-1.08	(-)	81.43%	LD	-0.67	HD + D	-3.50
5)	**Val49Arg**	8	(-)	-3.40	(-)	91.25%	DS	-1.31	S + N	1.36
6)	**Trp58Arg**	9	(-)	-1.94	(-)	88.93%	DS	-2.94	HD + D	-2.48
7)	**Leu71Asn**	8	(-)	-3.43	(-)	85.89%	DS	-1.73	HD + D	-3.39
8)	**Cys85Tyr**	8	(-)	-0.54	(-)	87.86%	DS	-1.35	HD + D	-2.46
9)	**Cys145Lys**	8	(-)	-0.44	(-)	92.86%	DS	-1.82	HD + D	-3.21
10)	**Phe161Asn**	8	(-)	-2.13	(-)	89.82%	DS	-3.05	HD + D	-3.29
11)	**Gly169Lys**	8	(-)	-2.22	(+)	27.14%	LD	-0.52	HD + D	-3.32
12)	**Phe189Pro**	8	(-)	-2.02	(-)	80.71%	DS	2.30	DS + N	-1.73
13)	**Leu192Asn**	8	(-)	-3.16	(-)	85.54%	DS	-2.44	HD + D	-2.02
14)	**Leu200Asn**	8	(-)	-3.53	(-)	81.79%	DS	-2.52	SS + N	0.69
15)	**Arg209Trp**	8	(-)	-0.53	(-)	84.54%	LS	0.89	S + N	1.21
16)	**Lys218Pro**	8	(+)	0.86	(+)	25.18%	DS	-1.45	SD + N	-0.87
17)	**Leu236Asn**	8	(-)	-2.99	(+)	25.18%	DS	-1.99	DS + N	-1.10
18)	**Leu239Asn**	8	(-)	-3.09	(-)	93.75%	DS	-1.87	DS + N	-1.81
19)	**Arg267Trp**	8	(-)	-0.89	(-)	81.25%	LS	0.89	SS + N	0.55
20)	**Val283Arg**	8	(-)	-3.84	(-)	91.79%	LD	-0.75	DS + N	-1.36
21)	**Trp290Arg**	9	(-)	-2.38	(-)	88.39%	DS	-2.17	N + N	-0.31
22)	**Arg308Cys**	8	(-)	-1.01	(+)	25.36%	LD	-0.61	SS + N	1.00
23)	**Gly314Ile**	8	(-)	0.29	(-)	78.04%	Neutral	-0.15	N + N	0.02
24)	**Arg321Cys**	8	(-)	-0.87	(-)	87.86%	DS	-2.28	SD + N	-0.93
25)	**Trp322Asn**	8	(-)	-1.84	(-)	88.04%	DS	-3.33	HD + D	-4.50
26)	**Trp332Asn**	9	(-)	-2.49	(-)	92.32%	DS	-3.79	HD + D	-5.57
27)	**Leu337Asn**	8	(-)	-3.46	(-)	92.32%	DS	-2.35	HD + D	-4.61
28)	**Trp349Asn**	9	(-)	-2.09	(-)	91.79%	DS	-3.09	HD + D	-4.50
29)	**Arg351Trp**	8	(-)	-0.96	(-)	79.64%	LD	-0.98	HS + D	2.14
30)	**Trp362Asn**	9	(-)	-1.70	(-)	91.07%	DS	-3.12	HD + D	-4.16
31)	**Val369Lys**	8	(-)	-2.67	(-)	94.46%	DS	-2.19	DS + N	-1.20
32)	**Arg372Cys**	8	(-)	-1.00	(-)	87.86%	LS	0.85	SS + N	0.66
33)	**Ala383Asp**	8	(-)	-1.23	(-)	80.81%	DS	-1.28	HD + D	-2.48
34)	**Ala387Trp**	8	(-)	-1.27	(-)	81.79%	LD	-0.50	SD +N	-0.55
35)	**Trp395Arg**	9	(-)	-1.01	(-)	87.86%	DS	-2.57	HD + D	-2.14
36)	**Leu402Asn**	8	(-)	-3.15	(-)	91.07%	DS	-2.14	SD + N	-0.90
37)	**Arg403Trp**	9	(-)	-1.09	(-)	81.61%	LS	0.83	HS + D	2.47
38)	**Ile408Lys**	8	(-)	-3.03	(-)	91.25%	DS	-2.11	HD + D	-2.74
39)	**Ala409Tyr**	8	(-)	-0.84	(+)	25.18%	LD	-0.53	N + N	-0.24
40)	**Arg427Cys**	8	(-)	-0.89	(-)	88.57%	LS	0.84	SD + N	-0.98
41)	**Ala432Tyr**	8	(-)	-0.87	(-)	81.79%	Neutral	-0.47	N +N	0.05
42)	**Ala434Trp**	8	(-)	-0.61	(-)	79.82%	Neutral	-0.45	HD + D	-2.39
43)	**Leu455Asn**	8	(-)	-3.48	(-)	91.07%	DS	-2.08	DS + N	-1.04
44)	**Ala484Asn**	8	(-)	-0.90	(-)	81.61%	LD	-0.61	N + N	-0.35
45)	**Ala485Asn**	8	(-)	-1.03	(-)	89.64%	LD	-0.51	SD + N	-0.76
46)	**Phe494Asn**	8	(-)	-1.75	(-)	92.32%	DS	-2.40	DS + N	-1.28
47)	**Asp534Trp**	8	(-)	-1.52	(-)	83.57%	Neutral	-0.31	S + N	1.30
48)	**Gly535Glu**	8	(+)	-0.21	(+)	25.54%	DS	-1.35	S + N	0.67
49)	**Arg539Trp**	9	(-)	-0.95	(-)	79.82%	DS	-1.18	SS + N	1.00

50)	**Arg540Trp**	9	(-)	-0.95	(-)	81.61%	DS	-1.13	SS + N	0.54
51)	**Leu558Asn**	8	(-)	-3.14	(-)	91.79%	LD	-0.84	HD + D	-3.88
52)	**Arg559Trp**	9	(-)	-0.95	(-)	85.54%	LD	-0.55	HD + D	2.65
53)	**Arg573Trp**	8	(-)	-1.05	(-)	79.11%	DS	-1.35	HD + D	3.38
54)	**Thr590Tyr**	8	(+)	1.40	(-)	81.79%	LD	-0.76	SD + N	-0.51
55)	**Lys591Pro**	8	(+)	-0.04	(-)	92.32%	Neutral	0.25	SS + N	0.82
56)	**Phe596Asn**	8	(-)	-0.88	(-)	92.50%	DS	-2.59	DS + N	-1.30
57)	**Ala600Tyr**	8	(-)	-0.91	(-)	80.54%	Neutral	-0.42	DS + N	-1.02
58)	**Phe634Arg**	8	(-)	-0.61	(-)	81.79%	Neutral	0.09	HD + D	-3.01
59)	**Trp646Arg**	9	(-)	-1.87	(-)	90.54%	DS	-2.80	HD + D	-2.33
60)	**Val649Arg**	8	(-)	-2.22	(-)	88.39%	DS	-1.63	HD + D	-3.21
61)	**Ala683Tyr**	9	(-)	-0.68	(-)	79.82%	LD	-0.64	HD + D	-2.21
62)	**His687Trp**	9	(-)	0.05	-	-	Neutral	-0.00	HD + D	2.66

Table B.2: List of *in silico* predicted mutations that leads to isoniazid resistance, observed in the patients of *M. tuberculosis*.

S. No.	**PMut + SIFT classification of pathological Amino acid Mutation**	**RI**	**Protein Stability**							
			I-MUTANT 2.0 (Sequence based)		**MuStab server**		**EASE-MM**		**SDM server**	
			Predictions (Stability)	**ΔΔG**	**Predictions (Stability)**	**confidence**	**Predictions (Stability class)**	**ΔΔG**	**Predictions (Stability/effect)**	**ΔΔG**
1)	**Pro21Lys**	8	(-)	-1.41	(-)	85.71%	LD	-0.5	DS + N	-1.34
2)	**Trp38Arg**	9	(-)	-1.79	(-)	87.86%	DS	-2.64	DS + N	-1.91
3)	**Trp39Arg**	8	(-)	-1.77	(-)	84.29%	DS	-2.41	HD + D	-3.94
4)	**Ile71Lys**	8	(-)	-2.19	(-)	91.79%	DS	-1.24	HD + D	-3.25
5)	**Thr85Lys**	8	(-)	-0.12	(-)	77.68%	LD	-0.73	SB + N	1.38
6)	**Trp91Asn**	9	(-)	-2.31	(-)	89.64%	DS	-4.02	HD + D	-2.23
7)	**Pro92Lys**	8	(-)	-1.25	(-)	88.39%	DS	-1.21	DS + N	-1.53
8)	**Ala93Asn**	8	(-)	-1.12	(-)	80.54%	DS	-1.08	HD + D	-4.22
9)	**Gly96Lys**	8	(-)	-1.01	(-)	92.86%	N	0.36	HS + D	2.78
10)	**Arg104Cys**	8	(-)	-1.07	(-)	83.75%	DS	-2.57	SS + N	0.67
11)	**Trp107Lys**	8	(-)	-2.01	(-)	93.04%	DS	-3.73	HD + D	-3.62
12)	**His108Lys**	8	(-)	-1.05	(-)	84.46%	DS	-1.45	HD + D	-3.91
13)	**Gly111Lys**	8	(-)	-1.86	(-)	90.54%	LD	-0.53	HD + D	-2.44
14)	**Gly121Glu**	8	(+)	0.06	(+)	26.25%	LD	-0.96	HD + D	-4.06
15)	**Arg128Cys**	8	(-)	-0.01	(-)	83.04%	LS	0.52	N + N	0.03
16)	**Asn138Phe**	8	(-)	-0.40	(+)	28.57%	N	0.17	HS + D	2.36
17)	**Lys143Pro**	8	(+)	0.94	(+)	23.57%	N	-0.39	S + N	1.10
18)	**Arg145Trp**	8	(-)	-0.33	(-)	78.93%	DS	-1.22	HS + D	2.67
19)	**Lys154Pro**	8	(+)	1.27	(-)	91.79%	N	0.22	DS + N	-1.10
20)	**Asp163Lys**	8	(-)	-1.05	(-)	88.93%	LD	-0.58	HD + D	-3.19
21)	**Gly169Lys**	8	(-)	-2.60	(-)	86.79%	N	-0.40	HD + D	-2.44
22)	**Gly186Lys**	8	(-)	-2.15	(-)	85.71%	LD	-0.73	HD + D	-2.68
23)	**Arg187Cys**	8	(-)	-0.88	(+)	22.86%	LS	0.65	SD + N	-0.93
24)	**Asp189Lys**	8	(-)	-1.08	(-)	82.32%	LD	-0.79	DS + N	-1.95
25)	**Trp198Arg**	9	(-)	-1.56	(-)	78.57%	DS	-2.67	HD + D	-3.44
26)	**Gly199Phe**	8	(-)	-1.10	(-)	88.57%	LD	-0.95	HS + D	2.08
27)	**Ala222Arg**	8	(-)	-0.63	(-)	79.64%	LD	-0.59	N + N	0.44
28)	**Met225Arg**	8	(-)	-1.39	(-)	85.71%	DS	-1.08	SD + N	-0.69
29)	**Gly226Gln**	8	(-)	-0.84	(-)	81.61%	LD	-0.95	HD + D	-2.99
30)	**Ile228Lys**	8	(-)	-2.71	(-)	92.32%	DS	-3.01	HD + D	-3.73
31)	**Pro232Lys**	8	(-)	-0.90	(+)	26.25%	DS	-1.21	N + N	0.28
32)	**Gly234Gln**	8	(-)	-0.71	(-)	81.25%	LD	-0.76	HD + D	-2.96
33)	**Met255Arg**	8	(-)	-1.37	(-)	81.43%	DS	-1.59	HD + D	-2.83
34)	**Gly268Gln**	8	(-)	-0.75	(-)	91.79%	LD	-0.58	N + N	-0.02
35)	**His270Lys**	9	(-)	-0.94	(-)	79.29%	LD	-0.91	DS + N	-1.65

36)	**His276Lys**	9	(-)	-0.92	(-)	80.89%	LD	-0.95	HD + D	-3.05
37)	**Gly285Arg**	8	(-)	-0.99	(-)	89.64%	DS	-1.19	HS + D	2.62
38)	**Pro288Lys**	8	(-)	-0.35	(-)	79.46%	LD	-0.62	HD + D	-2.59
39)	**Gly297Glu**	9	(+)	0.23	(-)	85.71%	LD	-0.75	HD + D	-2.81
40)	**Gly299Gln**	8	(-)	-0.44	(-)	87.5%	DS	-1.08	SD + N	-0.88
41)	**Trp300Arg**	8	(-)	-1.65	(-)	90.71%	DS	-2.55	HD + D	-2.02
42)	**Gly307Arg**	8	(-)	-2.00	(+)	28.04%	LD	-0.54	HD + D	-2.98
43)	**Ser315Arg**	8	(+)	-0.58	(+)	25.18%	LD	-0.81	S + N	1.93
44)	**Trp321Arg**	9	(-)	-1.88	(-)	79.64%	DS	-2.91	HD + D	-2.64
45)	**Trp341Arg**	9	(-)	-1.87	(-)	92.86%	DS	-2.90	HD + D	-2.14
46)	**Pro347Trp**	8	(-)	-1.33	(+)	25.71%	DS	-1.41	SS + N	0.74
47)	**Gly349Gln**	8	(-)	-1.48	(+)	23.04%	DS	-1.04	HD + D	-2.43
48)	**Asp366Lys**	8	(+)	-1.03	(-)	78.57%	LD	-0.62	N + N	-0.03
49)	**Met377Asn**	8	(-)	-0.86	(-)	92.5%	DS	-2.26	HD + D	-4.69
50)	**Asp381Lys**	8	(-)	-1.41	(+)	27.14%	LD	-0.79	HD + D	-3.19
51)	**Asp387Trp**	8	(-)	-1.24	(-)	80.89%	N	-0.42	SS + N	0.70
52)	**Phe408Asn**	8	(-)	-1.50	(-)	94.11%	DS	-2.90	HD + D	-4.37
53)	**Lys414Trp**	8	(-)	-1.17	(-)	77.68%	LD	-0.74	HS + D	3.55
54)	**Leu415Asn**	8	(-)	-2.90	(-)	86.79%	DS	-2.19	HD + D	-4.61
55)	**Arg418Trp**	8	(-)	-0.91	(-)	81.43%	SB	1.12	SS + N	0.80
56)	**Asp419Trp**	8	(-)	-1.22	(-)	83.57%	N	-0.03	HS + D	5.40
57)	**Gly421Tyr**	9	(-)	-0.90	(-)	81.61%	DS	-1.30	DS + N	-1.15
58)	**Pro422Phe**	9	(-)	-1.03	(-)	82.32%	DS	-1.00	SD + N	-0.72
59)	**Arg425Trp**	9	(-)	-1.08	(-)	83.93%	SB	1.07	S + N	1.64
60)	**Gly428Phe**	9	(-)	-0.28	(-)	83.75%	DS	-1.12	HS + N	0.85
61)	**Pro432Trp**	9	(-)	-1.70	-	-	DS	-1.53	N + N	-0.14

Table B.3: List of predicted mutations that may result in rifampicin resistance in the *M. tuberculosis*.

S. No.	**PMut [a]+ SIFT predicted disease associated mutations**	**RI**	**Protein Stability**							
			I-MUTANT 2.0 [d] (Sequence based)		**MuStab server [e]**		**EASE-MM**		**SDM server [g]**	
			Predictions (Stability)	**ΔΔG**	**Predictions (Stability)**	**confidence**	**Predictions (Stability class)**	**ΔΔG**	**Predictions (Stability/effect)**	**ΔΔG**
1)	**Gly4Gln**	8	(-)	-0.67	(-)	84.82%	N	-0.43	SD + N	-0.78
2)	**Arg38Trp**	8	(-)	-0.90	(-)	81.61%	LD	-0.74	SB + N	1.79
3)	**Pro51Lys**	8	(-)	-1.10	(-)	81.61%	DS	-2.17	HD +D	-2.82
4)	**Gly85Gln**	8	(-)	-1.00	(-)	87.86%	N	-0.28	N + N	-0.02
5)	**Pro95Leu**	8	(-)	-0.22	(-)	79.64%	DS	-1.13	SS + N	0.81
6)	**Gly145Cys**	8	(-)	-0.76	(+)	25%	DS	-1.22	N + N	-0.44
7)	**Met154Asp**	8	(+)	-0.03	(-)	91.96%	LD	-0.67	HD +D	-5.36
8)	**Gly155Lys**	8	(-)	-0.87	(-)	87.86%	SB	1.05	SS + N	0.68
9)	**Met160Arg**	8	(-)	-1.82	(-)	79.64%	LD	-0.99	DS + N	-1.56
10)	**Gly164Glu**	9	(+)	-0.85	(-)	92.68%	N	-0.43	HD + D	-3.36
11)	**Gly170Glu**	8	(+)	-0.00	(-)	89.64%	LD	-0.90	HD + D	-3.41
12)	**Glu172Ile**	8	(-)	0.32	(-)	87.86%	LD	-0.68	HS + D	4.10
13)	**Arg173Trp**	8	(-)	-0.65	(-)	82.14%	N	0.23	HS + D	2.02
14)	**Gln178Trp**	8	(-)	-0.09	(-)	80.89%	DS	-1.00	HS + D	2.55
15)	**Arg181Cys**	8	(-)	-1.57	(-)	83.93%	LD	-0.67	SS + N	0.51
16)	**Ser182Trp**	8	(+)	-0.09	(-)	81.61%	DS	-1.64	HD + D	-4.31
17)	**Pro183Lys**	8	(-)	-1.93	(-)	81.61%	N	-0.35	HD + D	-3.69
18)	**Gly209Ile**	8	(-)	-0.80	(-)	79.82%	LD	-0.93	SD + N	-0.26
19)	**Trp211Arg**	9	(-)	-2.32	(-)	88.93%	DS	-1.16	DS + N	-1.79
20)	**Asp217Lys**	8	(-)	-2.67	(-)	88.04%	N	-0.25	N + N	-0.03
21)	**Ala240Arg**	8	(-)	-0.53	(-)	82.86%	N	-0.42	HD + D	-2.61
22)	**Gly242Lys**	8	(-)	-1.35	(-)	92.86%	LD	-0.78	HD + D	-3.37
23)	**Ala273Lys**	8	(-)	-1.45	(-)	79.46%	LD	-0.80	HD + D	-5.17
24)	**Arg282Cys**	8	(-)	-1.08	(-)	80.54%	N	0.49	N + N	0.97
25)	**Phe301Asn**	9	(-)	-1.80	(-)	94.82%	DS	-1.29	SD + N	-1.28

26)	**Leu308Lys**	8	(-)	-4.10	(-)	94.11%	DS	-2.40	HD + D	-4.02
27)	**Val311Lys**	8	(-)	-2.11	(-)	93.93%	DS	-1.23	N + N	0.51
28)	**Gly312Cys**	8	(-)	-1.35	(-)	78.75%	DS	-1.96	HS + D	2.91
29)	**Arg313Trp**	9	(-)	-0.22	(-)	88.57%	LD	-0.52	HS + D	2.53
30)	**Asn317Trp**	8	(+)	-0.06	(-)	79.64%	N	-0.06	HS + D	4.16
31)	**Leu320Lys**	8	(-)	-3.00	(-)	92.86%	LD	-0.82	HD + D	-4.53
32)	**Asn375Trp**	8	(-)	0.06	(-)	79.64%	DS	-1.58	HS + D	2.55
33)	**Gly392Lys**	8	(-)	-0.82	(-)	89.64%	N	-0.13	HD +D	-2.44
34)	**Arg395Trp**	9	(-)	-0.56	(-)	77.32%	DS	-1.33	SS + N	1.53
35)	**Arg398Trp**	9	(-)	-0.43	(-)	81.79%	LD	-0.54	SB + N	2.5
36)	**Asn419Trp**	9	(-)	0.75	(+)	24.11%	N	0.50	SB + N	2.85
37)	**Phe430Asn**	8	(-)	-1.76	(-)	92.32%	DS	-2.03	HD + D	-3.74
38)	**Phe431Asn**	8	(-)	-1.80	(-)	93.57%	DS	-2.56	HD + D	-2.98
39)	**Gly432Glu**	8	(+)	-0.36	(-)	86.07%	N	0.18	HD + D	-2.89
40)	**Ser434Trp**	8	(+)	0.58	(-)	83.57%	N	-0.10	SD + N	-1.95
41)	**Leu436Lys**	8	(-)	-1.84	(-)	92.86%	N	0.49	HD + D	-2.12
42)	**Ser437Lys**	8	(+)	0.03	(-)	86.07%	DS	-1.01	DS + N	-1.58
43)	**Met440Arg**	8	(-)	-0.24	(-)	92.5%	LD	-0.99	DS + N	-1.96
44)	**Asp441Lys**	8	(-)	-0.29	(-)	94.11%	DS	-1.63	HD + D	-2.6
45)	**Asn444Trp**	8	(+)	0.24	(+)	29.11%	DS	-1.22	SS + N	1.68
46)	**Lys452Trp**	8	(-)	0.06	(+)	27.14%	DS	-1.29	HS + D	2.43
47)	**Arg453Trp**	9	(-)	0.04	(+)	24.82%	LD	-0.79	HS + D	5.73
48)	**Arg454Trp**	9	(-)	0.18	(+)	30.36%	LD	-0.57	N + N	0.55
49)	**Ala457Asp**	8	(+)	0.88	(-)	78.57%	LS	0.86	SD + N	-1.12
50)	**Gly459Asp**	8	(-)	-0.35	(-)	78.57%	N	0.21	HD + D	-2.58
51)	**Gly461Phe**	8	(-)	-0.50	(-)	79.46%	LD	-0.63	SB + N	2.08
52)	**Gly462Glu**	8	(+)	0.53	(+)	26.25%	DS	-1.24	HS + D	2.33
53)	**Leu463Lys**	8	(-)	-2.80	(-)	87.86%	DS	-1.01	HD + D	-2.12
54)	**Arg467Trp**	9	(-)	-0.46	(-)	77.68%	DS	-1.06	SS + N	1.82
55)	**Ala468Gln**	8	(-)	-0.66	(-)	78.57%	LD	-0.68	N + N	-0.13
56)	**Val472Lys**	9	(-)	-2.90	(-)	93.21%	LD	-0.93	SD + N	-1.2
57)	**Arg473Cys**	8	(-)	-1.30	(-)	80.71%	N	-0.35	SB + N	1.11
58)	**His476Lys**	9	(-)	-0.38	(-)	81.07%	DS	-1.25	DS + N	-1.86
59)	**His479Lys**	9	(-)	-0.04	(-)	80.54%	N	-0.33	SD + N	-1.65
60)	**Gly481Asp**	8	(-)	-0.50	(-)	82.32%	LD	-0.99	HD + D	-2.77
61)	**Arg482Cys**	8	(-)	-0.89	(-)	85.71%	DS	-1.15	SS + N	1.4
62)	**Cys484Tyr**	9	(-)	-0.50	(-)	81.61%	N	-0.26	HD + D	-4.45
63)	**Pro485Lys**	8	(-)	-0.51	(-)	89.64%	N	0.17	N + N	0.28
64)	**Ile486Asn**	9	(-)	-0.55	(-)	90.71%	N	-0.29	HD + D	-2.69
65)	**Thr488Arg**	8	(-)	0.53	(-)	90.71%	LD	-0.58	SS + N	1.28
66)	**Pro489Lys**	8	(-)	-0.65	(-)	87.86%	N	0.31	HD + D	-3.69
67)	**Gly491Glu**	8	(+)	-0.21	(-)	78.04%	LD	-0.74	N + N	0.01
68)	**Pro492Lys**	9	(-)	-0.77	(+)	25.18%	LD	-0.79	N + N	0.54
69)	**Asn493Trp**	8	(-)	0.36	(-)	82.32%	DS	-1.44	SS + N	1.57
70)	**Gly495Asp**	8	(-)	-1.31	(-)	85.54%	DS	-1.27	HD + D	-2.37
71)	**Ile497Lys**	8	(-)	-3.09	(-)	78.04%	DS	-2.49	HD + D	-5.39
72)	**Leu500Lys**	8	(-)	-3.36	(-)	88.93%	DS	-1.61	HD + D	-5.45
73)	**Asn507Trp**	8	(-)	-0.71	(-)	78.57%	DS	-2.48	HS + D	3.56
74)	**Gly510Phe**	8	(-)	-0.40	(-)	78.57%	LD	-0.88	HS + D	3.9
75)	**Phe511Asn**	8	(-)	-2.87	(-)	94.11%	LD	-0.60	SD + N	-1.2
76)	**Pro515Lys**	8	(-)	-1.68	(-)	92.68%	N	-0.34	DS + N	-2.24
77)	**Ala533Gln**	8	(-)	-0.68	(-)	92.5%	DS	-1.51	N + N	-0.21
78)	**Asp534Trp**	8	(-)	-0.48	(+)	23.04%	DS	-1.84	N + N	-0.25
79)	**Ala542Lys**	8	(-)	-1.28	(-)	93.75%	LS	0.66	HD + D	-4.23
80)	**Asp577Trp**	9	(-)	-0.46	(+)	24.29%	LD	-0.69	SS + N	1.33
81)	**Ser582Trp**	8	(+)	0.02	(-)	82.14%	LD	-0.84	SS + N	1.38
82)	**Pro583Lys**	8	(-)	-1.71	(-)	81.61%	N	-0.42	N + N	-0.64
83)	**Ile594Lys**	9	(-)	-2.79	(-)	92.86%	LD	-0.71	HD + D	-2.14
84)	**Glu598Trp**	8	(-)	-0.41	(-)	78.57%	N	-0.04	N + N	-0.5
85)	**Asp601Trp**	8	(-)	-0.53	(+)	27.68%	SB	1.03	SS + N	1.81
86)	**Asn603Ile**	9	(+)	0.72	(-)	78.04%	DS	-1.01	SS + N	1.4
87)	**Arg604Trp**	9	(-)	-0.81	(+)	28.75%	DS	-1.50	N + N	1.18

88)	**Ala605Gln**	9	(-)	-1.56	(-)	84.11%	DS	-1.12	SD + N	-1.91
89)	**Leu606Asn**	8	(-)	-2.13	(-)	92.86%	DS	-2.93	HD + D	-4.37
90)	**Met607Asp**	8	(-)	-0.82	(-)	90%	LS	0.77	HD + D	-2.26
91)	**Gly608Pro**	9	(-)	-0.64	(-)	82.32%	N	-0.06	N + N	0.89
92)	**Asn610Trp**	9	(-)	-0.69	(+)	24.46%	SB	1.03	SS + N	2.8
93)	**Met611Arg**	9	(-)	-0.08	(+)	22.86%	LD	-0.62	N + N	-0.93
94)	**Gln612Trp**	9	(-)	-0.12	(+)	30.36%	N	-0.47	HS + D	3.08
95)	**Arg613Trp**	9	(-)	-0.32	(-)	85.89%	N	-0.01	SS + N	1.74
96)	**Gln614Trp**	9	(-)	-0.28	(-)	78.75%	LD	-0.52	HS + D	2.63
97)	**Ala615Pro**	9	(-)	-0.72	(-)	88.04%	N	0.46	SD + N	-1.81
98)	**Val616Trp**	9	(-)	-1.00	(-)	85%	LD	-0.66	SS + N	1.71
99)	**Pro617Phe**	9	(-)	-0.48	(-)	79.64%	DS	-1.88	HS + D	5.1
100)	**Leu618Asn**	8	(-)	-2.84	(-)	91.79%	DS	-1.15	SD + N	-1.57
101)	**Pro624Trp**	9	(-)	-1.76	(-)	80.54%	LD	-0.79	N + N	0.27
102)	**Gly627Phe**	9	(-)	0.00	(-)	83.75%	DS	-2.68	HS + D	2.97
103)	**Thr628Tyr**	9	(-)	1.48	-	-	DS	-2.45	N + N	0.93
104)	**Gly629Cys**	9	(-)	-0.06	-	-	DS	-2.37	HS + D	4.23

Table B.4: List of predicted mutations leading to the streptomycin resistance in the *M. tuberculosis*.

S. No.	**PMut + SIFT based disease causing mutations**	**RI**	**Protein Stability**							
			I-MUTANT 2.0 (Sequence based)		**MuStab server**		**EASE-MM**		**SDM server**	
			Predictions (Stability)	**ΔΔG**	**Predictions (Stability)**	**confidence**	**Predictions (Stability)**	**ΔΔG**	**Predictions (Stability/effect)**	**ΔΔG**
1)	**Met1Asn**	9	(-)	0.06	-	-	N	-0.49	DS + N	-1.58
2)	**Thr3Tyr**	9	(-)	0.88	(-)	79.11%	LD	-0.67	SS + N	-0.53
3)	**Gln6Trp**	9	(-)	0.19	(+)	24.11%	N	-0.19	HS + D	2.39
4)	**Leu7Trp**	9	(-)	-0.08	(-)	81.25%	N	-0.26	SS + N	0.87
5)	**Arg9Trp**	9	(-)	0.02	(+)	25.36%	SB	1.14	SB + N	1.18
6)	**Arg12Trp**	9	(-)	0.22	(+)	28.57%	LS	0.97	SB + N	1.78
7)	**Ala22Trp**	9	(-)	-0.70	(+)	26.25%	N	-0.45	N + N	-0.02
8)	**Ala23Gln**	9	(-)	-1.18	(-)	89.64%	LD	-0.67	N + N	0.23
9)	**Leu24Trp**	9	(-)	-0.35	(-)	79.82%	LD	-0.51	HD + D	-7.36
10)	**Ser27Trp**	9	(+)	0.16	(+)	22.68%	LD	-0.94	N + N	0.37
11)	**Pro28Trp**	9	(-)	-1.09	(-)	77.5%	LD	-0.86	N + N	1.32
12)	**Gln29Lys**	9	(-)	-0.46	(-)	79.82%	LD	-0.60	N + N	-0.45
13)	**Arg30Trp**	9	(-)	-0.02	(-)	80.54%	SB	1.09	HS + D	2.02
14)	**Arg31Trp**	9	(-)	-0.02	(-)	78.57%	LS	0.95	HS + D	2.14
15)	**Gly32Phe**	9	(-)	-0.45	(+)	25.71%	DS	-1.09	HS + D	2.13
16)	**Val33Tyr**	9	(-)	-2.38	(-)	92.32%	DS	1.29	N + N	-0.40
17)	**Cys34Tyr**	9	(-)	0.41	(-)	82.68%	DS	-1.15	HD + D	-4.08
18)	**Arg36Trp**	9	(-)	-0.29	(-)	81.07%	N	0.25	N + N	0.27
19)	**Val37Trp**	9	(-)	-1.93	(-)	90%	DS	-1.25	SD + N	-0.51
20)	**Thr39Tyr**	9	(-)	0.76	(-)	82.32	N	-0.46	N + N	0.03
21)	**Thr41Tyr**	9	(-)	0.83	(-)	80.71%	LD	-0.91	SS + N	0.61
22)	**Pro42Trp**	9	(-)	-1.5	(+)	23.04%	DS	-1.51	DS + N	-1.48
23)	**Lys44Trp**	9	(-)	-0.1	(+)	28.21%	LS	0.50	SB + N	1.39
24)	**Pro45Trp**	9	(-)	-1.46	(+)	23.57%	DS	-1.19	SB + N	1.32
25)	**Asn46Trp**	9	(+)	0.53	(+)	25.89%	N	-0.37	HS + D	2.85
26)	**Ser47Trp**	8	(+)	-0.1	(+)	26.61%	N	0.25	DS + N	-1.49
27)	**Arg50Trp**	8	(-)	-0.50	(+)	26.07%	LD	-0.87	HS + D	2.46
28)	**Lys51Trp**	8	(+)	0.46	(+)	26.79%	LD	-0.69	HS + D	3.47
29)	**Val52Trp**	8	(-)	-1.26	(-)	82.32%	DS	-1.55	DS + N	-1.56
30)	**Lys56Pro**	8	(+)	1.78	(-)	81.25%	N	0.25	DS + N	-1.14
31)	**Leu57Trp**	8	(-)	-1.02	(-)	88.21%	N	-0.23	N + N	-0.35
32)	**Glu62Trp**	8	(-)	-0.21	(+)	28.57%	LS	0.72	SS + N	0.90
33)	**Ala65Phe**	9	(-)	0.38	(-)	83.39%	N	-0.17	SD + N	0.59
34)	**Tyr66Asp**	9	(-)	0.02	(-)	91.96%	DS	-2.77	DS + N	-1.49

35)	**Gly71Phe**	9	(-)	0.35	(-)	78.75%	LD	-0.96	HS + D	2.86
36)	**His72Trp**	9	(-)	0.33	(+)	29.11%	N	-0.07	HS + D	4.22
37)	**Asn73Leu**	9	(+)	1.39	(+)	31.07%	N	0.23	SS + N	0.88
38)	**Gln75Trp**	9	(-)	-0.14	(-)	79.64%	N	0.23	N + N	0.07
39)	**Glu76Trp**	9	(-)	0.15	(+)	25.36%	LS	0.55	SD + N	-0.53
40)	**His77Trp**	9	(+)	0.66	(+)	23.93%	N	-0.21	SB + N	1.47
41)	**Ser78Trp**	9	(+)	-0.03	(-)	78.04%	N	0.05	HD + D	-2.38
42)	**Val82Lys**	9	(-)	-2.80	(-)	90%	DS	-2.64	HD + D	-5.15
43)	**Gly84Phe**	9	(-)	0.47	(-)	79.64%	DS	-1.26	N + N	0.49
44)	**Val87Trp**	9	(-)	-2.45	(-)	78.04%	N	0.45	SS + N	0.64
45)	**Lys88Pro**	9	(+)	0.49	(-)	89.64%	N	-0.11	N + N	-0.37
46)	**Asp89Trp**	9	(-)	-1.79	(-)	79.64%	N	-0.15	SB + N	1.74
47)	**Pro91Trp**	9	(-)	-1.45	(-)	81.61%	DS	-1.50	SD + N	-0.96
48)	**Val93Trp**	9	(-)	-2.14	(-)	88.21%	DS	-1.03	SB + N	1.78
49)	**Arg94Trp**	9	(-)	-0.73	(-)	80.54%	N	0.45	SB + N	1.89
50)	**Lys96Pro**	8	(+)	0.35	(-)	79.82%	N	0.28	SB + N	1.37
51)	**Arg99Val**	9	(-)	-0.54	(-)	85.71%	N	-0.18	N + N	-0.21
52)	**Gly100Phe**	9	(-)	-1.05	(+)	24.82%	DS	-1.01	N + N	-0.07
53)	**Leu102Trp**	9	(-)	-0.62	(-)	78.75%	N	-0.09	HD + D	-7.43
54)	**Asp103Trp**	9	(-)	-1.31	(+)	26.25%	N	-0.07	N + N	0.07
55)	**Gly106Phe**	9	(-)	-1.23	(+)	28.57%	LD	-0.82	N + N	0.39
56)	**Val107Trp**	9	(-)	-0.41	(-)	87.86%	N	-0.37	N + N	0.28
57)	**Arg110Trp**	9	(-)	0.04	(-)	78.04%	LD	-0.85	HS + D	2.87
58)	**Arg114Met**	9	(-)	-0.90	(+)	26.07%	N	-0.06	N + N	-0.37
59)	**Ser115Trp**	9	(+)	-0.24	(-)	77.32%	N	-0.34	DS + N	-1.07
60)	**Arg116Trp**	9	(-)	-0.48	(+)	27.14%	LD	-0.57	SB + N	1.44
61)	**Tyr117Cys**	9	(+)	-0.22	(+)	25.18%	LD	-0.83	HS + D	2.13
62)	**Gly118Phe**	9	(-)	-0.43	(+)	27.14%	LD	-0.57	N + N	-0.33
63)	**Lys120Pro**	9	(+)	0.43	(-)	88.75%	N	0.24	SB + N	1.35
64)	**Lys123Pro**	9	(+)	0.46	-	-	N	0.37	SB + N	1.35

Appendix C

Outcomes of the function prediction of Hypothetical Proteins (HPs)

Table C.1: List of HPs predicted to be enzymes in the group of HPs obtained from *M. tuberculosis* (The class of enzyme is indicated with Tan background)

S. No.	Uniprot ID	Function
		Transferase
1.	I6Y6S3	Class I glutamine amidotransferase-like
2.	P96407	glycosyl transferase
3.	P95226	DNA polymerase
4.	O53699	Sulfotransferase
5.	O07250	S-adenosyl-L-methionine-dependent methyltransferases
6.	O07251	S-adenosyl-L-methionine-dependent methyltransferases
7.	O07253	S-adenosyl-L-methionine-dependent methyltransferases
8.	O06296	molybdopterin cytidylyltransferase
9.	O06315	Glycosyltransferase
10.	I6WY86	MobA-like NTP transferase domain
11.	P95198	Rhodanese-related sulfurtransferase
12.	L7N651	1,4-dihydroxy-2-naphthoate prenyltransferase
13.	O53756	1,4-dihydroxy-2-naphthoate prenyltransferase
14.	P9WKT1	Acyltransferase
15.	P9WFK3	MaoC dehydratase
16.	P9WHK1	Phosphoribosyltransferase
17.	O07763	galactose-1-phosphate uridylyltransferase
18.	P95073	Leucine carboxyl methyltransferase 1
19.	I6XWA9	Transglutaminase-like superfamily
20.	I6X9V3	Aminomethyltransferase beta-barrel domains
21.	O53848	PLP-dependent transferases
22.	I6XWK6	Gamma-glutamylcyclotransferase
23.	O53409	Glutamine amidotransferases class-II (GATase)
24.	O53414	sulfurtransferase
25.	O06539	Isoprenylcysteine carboxyl methyltransferase (ICMT) family
26.	O06547	Geranyl diphosphate 2-C-methyltransferase
27.	P9WM13	diguanylate cyclase
28.	P9WLZ5	Nucleotidyltransferase (NT) domain of DNA polymerase beta and similar

		proteins
29.	P9WMU5	2-phospho-L-lactate transferase
30.	O06830	Glycerol-3-phosphate (1)-acyltransferase
31.	O53147	glutathione synthase/ribosomal protein S6 modification enzyme (glutaminyl transferase)
32.	P71780	glycosyltransferase (part1)
33.	P71784	sialic acid O-acetyltransferase NeuD
34.	P71785	SAM-dependent methyltransferase
35.	P9WLW3	methyltransferase domain protein
36.	P71792	S-adenosyl-L-methionine-dependent methyltransferases
37.	P71794	S-adenosyl-L-methionine-dependent methyltransferases
38.	P9WLV9	Glycosyl transferase family 2
39.	P9WLV7	Aspartate Aminotransferase, domain 1
40.	O06625	Acyltransferase family
41.	O06593	S-adenosyl-L-methionine-dependent methyltransferases
42.	O06145	OB fold (Dihydrolipoamide Acetyltransferase, E2P)
43.	O53920	Transglutaminase-like superfamily
44.	O33186	Multifunctional methyltransferase subunit Trm112p-like protein
45.	P9WLS7	Chloramphenicol Acetyltransferase
46.	P95151	MOSC (MOCO sulfurase C-terminal) domain
47.	P95149	Crotonobetainyl-coa:carnitine coa-transferase; domain 1
48.	P95148	acetyl-CoA acetyltransferase
49.	P95284	glutamine amidotransferase
50.	O53979	methyltransferase type 11
51.	P9WJZ5	SAM dependent methyltransferase
52.	P9WLN3	Aminoglycoside phosphotransferase family enzyme
53.	P9WLL9	S-adenosyl-L-methionine-dependent methyltransferases
54.	P9WLL5	Phosphatidylinositol (PI) phosphodiesterase
55.	O06232	Phosphoribosyl transferase domain
56.	P9WLG1	Sulfotransferase family
57.	P9WLD1	Chloramphenicol Acetyltransferase
58.	P71889	L-arginine/glycine Amidinotransferase; Chain A
59.	P71734	Bacterial transglutaminase-like N-terminal region
60.	P71730	DNA polymerase III, delta subunit
61.	P71916	methyltransferase type 11
62.	I6YDJ7	Thymidylate Synthase; Chain A
63.	Q50732	Bacterial transglutaminase-like N-terminal region
64.	P9WL93	Bacterial transglutaminase-like N-terminal region
65.	P9WL79	radical SAM protein

66.	I6Y1G8	Thiopurine S-methyltransferase
67.	O07191	S-adenosylmethionine-dependent methyltransferases (SAM or AdoMet-MTase), class I
68.	I6YEA3	Leucine carboxyl methyltransferase 1
69.	P71626	Nucleotidyl transferase AbiEii toxin, Type IV TA system
70.	P71625	Nucleotidyl transferase AbiEii toxin, Type IV TA system
71.	I6X5U4	S-adenosyl-L-methionine-dependent methyltransferases
72.	P95137	S-adenosyl-L-methionine-dependent methyltransferases
73.	I6Y242	S-adenosyl-L-methionine-dependent methyltransferases
74.	I6YEW1	S-adenosyl-L-methionine-dependent methyltransferases
75.	I6YAW3	Cobalamin-independent methionine synthase
76.	I6XFY8	Glycerol-3-phosphate (1)-acyltransferase
77.	I6YAZ1	Geranyl diphosphate 2-C-methyltransferase
78.	O05796	S-adenosyl-L-methionine-dependent methyltransferases
79.	O07206	glutathione S-transferase
80.	O05876	L-lysine-epsilon aminotransferase lat'
81.	O05887	Phosphoribosyltransferase
82.	P96877	CoA transferase
83.	O53356	gamma-glutamylcyclotransferase
84.	O50419	DNA polymerase
85.	P9WGJ7	Type I PLP-dependent aspartate aminotransferase-like (Major domain)
86.	P9WKZ3	Formyltransferase
87.	P9WKY7	ribosomal-protein-alanine acetyltransferase rimI
88.	O53566	OB fold (Dihydrolipoamide Acetyltransferase, E2P)
89.	I6YCC4	Sulfotransferase family
90.	I6YGI1	arylamine N-acetyltransferase
91.	I6X831	glutamyl-tRNA amidotransferase
92.	O69667	Thiopurine S-methyltransferase
93.	O69689	Aspartate Aminotransferase, domain 1&2
94.	P9WKW9	Glycosyl transferase family 2
95.	P96246	Guanidinoacetate N-methyltransferase B
96.	O53594	Acetyltransferase
		Peptidase
97.	O07436	L,D-transpeptidase 4
98.	O07236	L,D-transpeptidase catalytic domain protein
99.	O06308	Peptidase M50
100.	P71560	Glycyl-glycine endopeptidase LytM
101.	P96358	Trypsin-like peptidase domain
102.	O05316	X-Pro dipeptidyl-peptidase

103.	P9WM47	Peptidase family C39
104.	O06825	L,D-transpeptidase catalytic domain-like
105.	O06624	endopeptidase domain like (from Nostoc punctiforme)
106.	O06803	Leucine aminopeptidase 1
107.	P9WLR9	Glutamyl endopeptidase 2/ Trypsin-like serine proteases
108.	O53978	Peptidase family M48
109.	L7N684	Zn peptidases
110.	I6X4J0	Peptidase family M20/M25/M40
111.	I6Y9M6	Type IV leader peptidase family
112.	P9WL95	Putative zinc-binding metallo-peptidase
113.	P9WL33	M23 peptidase domain-containing protein
114.	O53341	Zinicin-like metallopeptidase
115.	O05859	Zinicin-like metallopeptidase
116.	O05872	SOS response associated peptidase (SRAP)
117.	O53352	Zinicin-like metallopeptidase
118.	O06381	Zinicin-like metallopeptidase
119.	P9WKX1	Peptidase family M23
120.	P96242	Zinicin-like metallopeptidase
121.	P96230	Peptidase
		Peroxidase
122.	P9WL19	OsmC-like protein
		Oxidoreductase
123.	P71591	2-nitropropane dioxygenase
124.	P9WMA5	Pyridoxamine 5'-phosphate oxidase
125.	P9WM65	Phosphopantetheine attachment site
126.	O07171	Pyridoxamine 5'-phosphate oxidase
127.	P9WI85	quercetin 2,3-dioxygenase/ Pirin-related protein
128.	O53680	glyoxalase/bleomycin resistance protein/dioxygenase
129.	O53707	Xanthine and CO dehydrogenases maturation factor, XdhC/CoxF family - like domain
130.	O53711	Xanthine and CO dehydrogenases maturation factor, XdhC/CoxF family - like domain
131.	P9WKW3	pyridoxamine 5'-phosphate oxidase-like protein
132.	O53734	Amine oxidase
133.	P9WKS9	Glutaredoxin
134.	O06389	Deazaflavin-dependent oxidoreductase, nitroreductase family protein
135.	O06412	Glyoxalase/Bleomycin resistance protein/Dioxygenase superfamily
136.	P9WIR3	Glyoxalase/Bleomycin resistance protein/Dioxygenase superfamily protein
137.	I6WZ39	glyoxalase/bleomycin resistance protein/dioxygenase

138.	I6Y8K5	Quinoprotein amine dehydrogenase/ PE-PGRS family protein
139.	I6X9T8	Pyrimidine monooxygenase RutA
140.	I6WZG6	Encapsulating protein for peroxidase
141.	I6Y4U9	Dyp-type peroxidase-like
142.	O06633	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase
143.	P9WKQ5	Flavin-containing monooxygenase
144.	P9WKQ3	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase
145.	I6XA34	2,3-Dihydroxybiphenyl 1,2-Dioxygenase; domain 1
146.	I6Y946	NADPH-dependent FMN reductase
147.	I6Y946	dihydrodipicolinate reductase
148.	L7N6A4	2OG-Fe(II) oxygenase superfamily
149.	O53405	Quinoprotein amine dehydrogenase
150.	O53407	Dihydrodipicolinate reductase
151.	O53413	Cysteine dioxygenase type I
152.	O06569	Antibiotic biosynthesis monooxygenase
153.	O06552	pyridoxamine 5'-phosphate oxidase
154.	P9WP13	Electron Transport, Fmn-binding Protein; Chain A/Pnp Oxidase; Chain A
155.	L7N6B1	2,3-Dihydroxybiphenyl 1,2-Dioxygenase, domain 1
156.	P9WFP7	Fe-S cluster assembly protein SufB
157.	P9WFP5	Fe-S cluster assembly protein SufD
158.	O53157	Fe-S cluster assembly (FSCA)/ Phenylacetate-CoA oxygenase, PaaJ subunit
159.	P9WI91	Phytanoyl-CoA dioxygenase
160.	P9WLU9	ketoacyl reductase
161.	P9WP11	Pyridoxine 5'-phosphate (PNP) oxidase-like and flavin reductase-like proteins
162.	O06592	Pyridoxine 5'-phosphate (PNP) oxidase-like and flavin reductase-like proteins
163.	O53923	AhpC/TSA family (alkyl hydroperoxide reductase (AhpC) and thiol specific antioxidant (TSA))
164.	P71990	alkyl hydroperoxide reductase
165.	O06789	Trans-1,2-dihydrobenzene-1,2-diol dehydrogenase
166.	O07754	Pyridoxamine 5'-phosphate oxidase
167.	O07738	Nitronate monooxygenase
168.	P9WLN7	Nucleotide-binding domain/ N-terminal domain of adrenodoxin reductase-like
169.	O86340	Pyridoxamine 5'-phosphate oxidase
170.	O06216	Luciferase
171.	P9WMN5	Iron-sulphur cluster biosynthesis
172.	P9WLH7	Oxygenase, catalysing oxidative methylation of damaged DNA
173.	P9WLE7	DSBA-like thioredoxin domain
174.	P9WLB3	Formate dehydrogenase/DMSO reductase, domains 1-3/ NADPH-cytochrome p450 reductase FAD-binding domain-like

175.	Q79FF7	Ferredoxin reductase-like, C-terminal NADP-linked domain
176.	P95233	FMN-dependent nitroreductase-like
177.	O53193	DSBA-like thioredoxin domain
178.	P9WLA5	antibiotic biosynthesis monooxygenase family protein
179.	P9WQ03	Archaeal ammonia monooxygenase subunit A (AmoA)
180.	P9WL45	chlorite dismutase
181.	I6X540	Inosine-5-monophosphate dehydrogenase related
182.	O33291	antibiotic biosynthesis monooxygenase
183.	O33313	NADPH-dependent FMN reductase
184.	O33272	Glutaredoxin
185.	I6XFT2	NADP-dependent oxidoreductase
186.	O53240	Pyridoxamine 5'-phosphate oxidase
187.	I6XFZ8	Methanol Dehydrogenase; Chain A
188.	P95105	NAD(P)H-dependent FMN reductase
189.	P95086	Beta subunit luciferase
190.	I6XG43	Luciferase-like monooxygenase
191.	O05773	Acyl-CoA dehydrogenase
192.	P9WL07	FMN-dependent nitroreductase-like
193.	O05856	tRNA--hydroxylase - like domain
194.	O05897	2-polyprenyl-6-methoxyphenol hydroxylase-like oxidoreductase
195.	O50380	D-Lactate dehydrogenase/ Berberine and berberine like
196.	O50382	trimethylamine N-oxide reductase I catalytic subunit
197.	O50398	Pyridoxamine 5'-phosphate oxidase
198.	I6X7D4	Luciferase-like monooxygenase
199.	I6XHG6	3-ketosteroid-9-alpha-hydroxylase oxygenase subunit domain protein
200.	P9WI89	Phytanoyl-CoA dioxygenase family protein
		Hydrolase
201.	O53605	Appr-1-p processing enzyme family
202.	O53619	Imidazolonepropionase
203.	Q50655	HNH nuclease
204.	P9WM61	S-adenosyl-L-homocysteine hydrolase
205.	P96819	3-methyladenine DNA glycosylase
206.	P96391	P-loop containing nucleoside triphosphate hydrolases
207.	P96392	P-loop containing nucleoside triphosphate hydrolases
208.	P95220	Allophanate hydrolase
209.	P95229	alpha/beta hydrolase
210.	O53697	alpha/beta hydrolase family protein
211.	O07246	Exported protein NLP/P60 family member
212.	L0T643	N-acetylglucosaminyl deacetylase, LmbE family

213.	O33266	HNH nucleases
214.	P95201	HNH nucleases
215.	O53701	AAA - ATPases
216.	P95204	Phospholipase D/nuclease
217.	P95205	HNH endonuclease
218.	P96267	Alpha/beta hydrolase
219.	P96280	ATP-dependent protease La (LON) domain protein
220.	O33360	HNH endonuclease domain protein
221.	O33363	GDSL-like Lipase/Acylhydrolase family
222.	O06396	P-loop containing nucleoside triphosphate hydrolases
223.	O06406	putative endoIII-related endonuclease
224.	O06418	Amidohydrolase
225.	O53776	Mut7-C RNAse domain
226.	I6WYU2	ATPase AAA+ superfamily protein
227.	I6Y897	Serine Threonine Protein Phosphatase 5, Tetratricopeptide repeat
228.	I6X9N8	DEAD/DEAH box helicase
229.	I6X9Q1	Integrase
230.	P9WML3	Diadenosine tetraphosphate (Ap4A) hydrolase and other HIT family hydrolases - like domain
231.	I6Y8R4	Carbohydrate acetyl esterase/feruloyl esterase
232.	I6Y4S9	RelA/SpoT domain-containing protein
233.	P71837	Glycoside hydrolase/deacetylase
234.	P71839	Metallo-beta-lactamase, chain A
235.	P71840	ATPase/ metallo-beta-lactamase superfamily protein
236.	P0CG96	Carboxypeptidase regulatory domain-like
237.	P9WKH9	Integrase, catalytic region - like domain
238.	O53874	Helicase conserved C-terminal domain
239.	P9WKQ1	Endonuclease/Exonuclease/phosphatase family
240.	P9WKP3	Metallo-beta-lactamase; Chain A
241.	O05900	DD-peptidase/beta-lactamase superfamily
242.	P9WKM7	alpha/beta-Hydrolases
243.	O86320	Exopolyphosphatase
244.	P96356	AAA ATPase
245.	O53399	AAA ATPases / Helix-Turn-Helix domain
246.	O53404	acyl-CoA thioesterase
247.	O53417	Alpha/beta-hydrolase family N-terminus
248.	O53421	VSR Endonuclease
249.	O06567	glycoside hydrolase
250.	O06570	endopeptidase domain like (from Nostoc punctiforme)

251.	P9WM57	HNH nucleases
252.	P9WM55	HNH nucleases
253.	O06554	DNA-glycosylase/Endonuclease III
254.	O50435	DNA or RNA helicase of superfamily II
255.	O50440	PE-PPE domain
256.	O86348	Alpha/beta hydrolase fold
257.	O05293	Alpha/beta hydrolase fold
258.	O05294	Alpha/beta hydrolase fold
259.	O05305	P-loop containing nucleotide triphosphate hydrolases
260.	O33221	Haloacid dehalogenase-like hydrolase
261.	O5046	DNA helicase II / ATP-dep6endent DNA helicase PcrA/ Exodeoxyribonuclease V alpha subunit
262.	P9WML1	HIT family hydrolase
263.	P9WM41	P-loop containing nucleotide triphosphate hydrolases
264.	P9WM39	Gamma-D-glutamyl-L-diamino acid endopeptidase 1/Carbohydrate acetyl esterase/feruloyl esterase
265.	P9WGC1	Metallo-beta-lactamase; Chain A
266.	P9WLZ3	DD-peptidase/beta-lactamase superfamily
267.	P71806	His-Me finger endonucleases
268.	P9WFQ3	P-loop containing nucleotide triphosphate hydrolases
269.	O06815	6-phosphogluconolactonase/ ATPase
270.	P71767	Membrane-bound serine protease NfeD
271.	P9WPR9	stomatin/prohibitin-family membrane protease subunit
272.	P9WLW7	Glycosyl hydrolase domain; family 43
273.	O06603	HNH nucleases
274.	O06597	Nudix hydrolase
275.	O06154	Metallo-beta-lactamase, chain A
276.	P9WLT3	HNH nucleases
277.	P71976	alpha/beta-Hydrolases
278.	P71986	glycoside hydrolase
279.	O06798	His-Me finger endonucleases
280.	P9WLR5	DNase-RNase Family
281.	P95145	A/G-specific adenine DNA glycosylase
282.	O07751	Metal-dependent hydrolases
283.	O07720	Metallo-beta-lactamase, chain A
284.	P95270	Preprotein translocase SecA subunit-related protein
285.	P9WLQ5	HNH nucleases
286.	P95267	Helicase
287.	O53977	YacP-like NYN domain (Nucleases)

288.	P9WLM9	P-loop containing nucleotide triphosphate hydrolases
289.	O53461	HNH nucleases
290.	P9WLM1	erythromycin esterase
291.	O53479	ATPase
292.	O53494	Amidohydrolase family
293.	O86353	Dienelactone hydrolase family
294.	P9WLK7	alpha/beta-Hydrolases
295.	P9WLJ9	IS1556 transposase
296.	P64937	Transposase
297.	P9WLJ3	HNH nucleases
298.	O33254	Lambda Exonuclease; Chain A
299.	O53518	P-loop containing nucleotide triphosphate hydrolases
300.	P9WLJ1	Exonuclease
301.	P9WLG5	VSR Endonuclease
302.	O53531	DD-peptidase/beta-lactamase superfamily
303.	O53534	Metallo-beta-lactamase, chain A
304.	P9WLC7	alpha/beta superfamily hydrolase
305.	P9WLC1	P-loop containing nucleotide triphosphate hydrolases
306.	O05833	cytidine deaminase
307.	P71926	Ribonuclease
308.	P71922	P-loop containing nucleotide triphosphate hydrolases
309.	O53196	HNH endonuclease
310.	I6Y0X6	P-loop containing nucleotide triphosphate hydrolases
311.	I6XEI5	Endo-N-acetylmuramidases (muramidases) are lysozymes
312.	P95024	Restriction endonuclease-like
313.	P95011	Alpha/beta hydrolase
314.	P9WL67	Alpha/beta hydrolase
315.	I6Y1G3	P-loop containing nucleotide triphosphate hydrolases
316.	I6Y1I1	Alpha/beta hydrolase
317.	I6YA17	dUTPase
318.	I6YA21	endoribonuclease L-PSP
319.	P9WFG7	fatty acid-binding protein/ Putative transposase
320.	O33319	Predicted metal-dependent hydrolase
321.	P71654	Alpha/beta hydrolase
322.	P71646	Integrase
323.	P71644	Integrase
324.	I6XFD1	P-loop containing nucleotide triphosphate hydrolases
325.	P9WPR1	magnesium chelatase
326.	O53470	Restriction endonuclease

327.	P9WFM9	Endonuclease
328.	P9WL23	Amidohydrolase family
329.	I6X5W6	HNH nucleases
330.	O53287	NUDIX hydrolase
331.	I6YF16	Endonuclease
332.	I6XG38	HNH endonuclease
333.	I6YB54	Glycosyl hydrolases family 2
334.	O05777	DinB/YfiT-like putative metal-dependent hydrolase
335.	O05791	Cytidine Deaminase, domain 2
336.	O53329	P-loop containing nucleotide triphosphate hydrolases
337.	O53335	Type III restriction enzyme, res subunit
338.	O05873	P-loop containing nucleotide triphosphate hydrolases
339.	L7N658	ATPase AAA+ superfamily protein
340.	O53363	alpha/beta fold family hydrolase
341.	I6X7B3	Alpha/beta hydrolase
342.	I6Y3Q7	Amidohydrolase
343.	P96837	VSR Endonuclease
344.	P96859	Metallo-beta-lactamase; Chain A
345.	I6Y3Z2	1,6-anhydro-N-acetylmuramyl-L-alanine amidase AmpD
346.	O06380	DD-peptidase/beta-lactamase superfamily
347.	I6XHV9	Helicase
348.	P9WKX7	P-loop containing nucleotide triphosphate hydrolases
349.	P9WGJ1	Haloacid Dehalogenase-like Hydrolases
350.	I6XHX8	Nucleoside Triphosphate Pyrophosphohydrolase
351.	I6YGW9	YjgF/chorismate_mutase-like, putative endoribonuclease
352.	O69681	VSR Endonuclease
353.	I6Y4D2	N-acetylmuramoyl-L-alanine amidase
354.	O69700	Nucleoside Triphosphate Pyrophosphohydrolase
355.	O69720	tRNA adenosine deaminase
356.	O69731	Glycosidases
357.	P72042	His-Me finger endonucleases
358.	P72062	Metallo-beta-lactamase; Chain A
359.	Q79F96	N-acetylmuramoyl-L-alanine amidase-like
360.	O07810	Cof-like hydrolase family protein
361.	P9WH21	Metallo-beta-lactamase, chain A
362.	P96233	Transposase
363.	P96217	P-loop containing nucleotide triphosphate hydrolases
364.	P96216	ATPase
365.	O05439	Pullulanase

		Lipase
366.	O53410	Patatin-like phospholipase
367.	O53423	GDSL-like Lipase/Acylhydrolase family
368.	P9WK89	Secretory lipase
369.	P94973	Esterases and lipases
370.	P71725	Lysophospholipase
371.	O05805	Lysophospholipase
372.	I6YB49	Patatin-like phospholipase
		Phosphatase
373.	P9WHV5	Exopolyphosphatase
374.	P9WM79	CapA and related proteins, metallophosphatase domain
375.	P96374	Ppx/GppA phosphatase family
376.	P9WGF9	phosphohistidine phosphatase
377.	Q50699	Purple Acid Phosphatase; chain A, domain 2
378.	P9WM07	EAL domain (diguanylate phosphodiesterase)
379.	O06240	Histidine phosphatase superfamily
380.	P9WLH9	CYTH-like phosphatases
381.	P9WL81	Purple acid phosphatase, N-terminal domain
382.	I6YEE1	Calcineurin-like phosphoesterase
383.	I6X827	Purple Acid Phosphatase; chain A, domain 2
384.	P96221	Phosphatase YcdX
		Sulfatase
385.	I6Y8I5	Sulfatase-modifying factor 1 (EC 1.8.99.-) (C-alpha-formylglycine-generating enzyme 1)
		Thioesterase
386.	P96817	Acyl-CoA thioesterase
387.	O07408	4-hydroxybenzoyl-CoA thioesterase
388.	O06307	Thioesterase
389.	O53751	Acyl-ACP thioesterase
390.	P9WKT9	Acyl-CoA thioesterase II (EC 3.1.2.-)
391.	O06178	Thioesterase superfamily - like domain
392.	O53917	acyl-CoA thioesterase
393.	P9WLN5	Acyl-ACP thioesterase
394.	I6Y9E8	Thioesterase-like superfamily
395.	I6XHI0	Thioesterase/thiol ester dehydrase-isomerase
396.	I6YGF8	Thioesterase/thiol ester dehydrase-isomerase
		Synthase
397.	P9WFP9	secondary thiamine-phosphate synthase enzyme

		Ligase
398.	I6Y8S6	Phosphoribosylformylglycinamidine synthase PurS subunit
399.	O05575	5-formyltetrahydrofolate cyclo-ligase
400.	P71763	cullin%2C a subunit of E3 ubiquitin ligase
401.	O06619	2'-5' RNA ligase superfamily
402.	P9WLP9	biotin--[acetyl-CoA-carboxylase] synthetase
403.	P71733	glutamate--cysteine ligase, Gcs2
404.	P9WLA9	carboxylate-amine ligase%2C YbdK family
405.	P9WPQ1	acetyl-CoA carboxylase biotin carboxyl carrier protein subunit
406.	P96873	acetyl-CoA synthetase
407.	O53562	cullin%2C a subunit of E3 ubiquitin ligase
408.	I6YCS6	Glutamate-cysteine ligase family 2 (GCS2)
409.	O69697	ATP-dependent DNA ligase
		Lyase
410.	P95216	cobalamin biosynthesis protein CbiX
411.	O53749	4-carboxymuconolactone decarboxylase
412.	P71813	3,4-dihydroxy-2-butanone 4-phosphate synthase
413.	P9WMV1	Adenylate and Guanylate cyclase catalytic domain
414.	O86370	Hydroxyneurosporene synthase (CrtC)
415.	O05580	isoprene synthase, chloroplastic-like
416.	O06571	Adenylyl and guanylyl cyclase catalytic domain
417.	O06572	Adenylyl Cyclase, chain A
418.	O06556	alpha-ketoglutarate decarboxylase
419.	O05306	lysine decarboxylase
420.	Q79FN0	GDP-D-mannose dehydratase
421.	O53905	Carboxymuconolactone decarboxylase family
422.	O53929	Cyclase Family
423.	O06800	Carboxymuconolactone decarboxylase family
424.	O06218	Carboxymuconolactone decarboxylase family
425.	P9WLB7	Carboxymuconolactone decarboxylase - like domain
426.	I6YF40	HpcH/HpaI aldolase/citrate lyase family
427.	O53342	bacteriocin biosynthesis cyclodehydratase domain-containing protein
428.	O53564	Acetoacetate decarboxylase-like
429.	P96239	prephenate dehydratase
430.	O05448	Transglycosylase
		Phosphorylase
431.	P9WLE3	Phosphorylase superfamily
432.	P9WLE1	Phosphorylase superfamily

Alanine racemase		
433.	P9WLY5	alanine racemase
434.	P9WLY3	alanine racemase
435.	O06802	Alanine racemase
436.	P9WFQ7	Alanine racemase
Kinase		
437.	P9WQI1	Homoserine kinase (HK) (HSK) (EC 2.7.1.39)
438.	O33198	thiamine pyrophosphokinase
439.	O06242	Phosphatidylinositol 3-and 4-kinase
440.	P9WMT7	Glycerate kinase family
441.	P71737	Histidine kinase
442.	I6YA47	signal transduction histidine kinase
443.	P95120	phosphatase domain of the dihydroxyacetone kinase
444.	O05861	Protein kinase-like (PK-like)
445.	O05848	Diacylglycerol kinase catalytic domain
446.	O50392	histidine kinase-, DNA gyrase B-, and HSP90-like ATPase family protein
447.	Q93IG6	histidine kinase
448.	O69702	Adenosine specific kinase family protein
Isomerase		
449.	P9WM93	SnoaL-like polyketide cyclase
450.	P9WM91	mycothiol-dependent maleylpyruvate isomerase
451.	P96815	polyketide cyclase
452.	P96818	Ketosteroid isomerase-related protein
453.	L7N657	Polyketide cyclase/dehydrase
454.	O07237	Polyketide cyclase SnoaL-like domain
455.	O07256	Mycothiol-dependent maleylpyruvate isomerase, metal-binding domain
456.	O33273	SnoaL-like domain protein
457.	P9WKU1	Xylose isomerase-like
458.	O06391	Phosphoglycerate mutase
459.	P9WKS3	Mycothiol maleylpyruvate isomerase N-terminal domain
460.	O53868	Polyketide cyclase / dehydrase and lipid transport
461.	O53408	Polyketide cyclase / dehydrase and lipid transport
462.	O53432	N-acylglucosamine 2-epimerase (GlcNAc 2-epimerase)
463.	O06820	polyketide cyclase
464.	L7N6A8	Methylmalonyl-CoA mutase large subunit
465.	P9WLU7	Polyketide cyclase
466.	P71985	Mycothiol maleylpyruvate isomerase N-terminal domain
467.	P95147	UDP-glucose 4-epimerase GalE4

468.	O07748	Polyketide cyclase
469.	O53961	Polyketide cyclase
470.	P9WLN9	PEP phosphonomutase-like enzyme
471.	O53480	Mycothiol maleylpyruvate isomerase N-terminal domain
472.	P9WIH5	Epimerase/NADH dehydrogenase (ubiquinone)/ PEP-utilising enzyme, mobile domain
473.	O53519	polyketide cyclase
474.	O53520	polyketide cyclase
475.	P71898	Putative modulator of DNA gyrase
476.	P71897	Putative modulator of DNA gyrase
477.	P9WL87	polyketide cyclase
478.	Q79FC4	triosephosphate isomerase
479.	P9WL43	Diaminopimelate Epimerase; Chain A, domain 1
480.	I6Y1P2	polyketide cyclase
481.	I6X666	polyketide cyclase
482.	O07166	Pseudouridine synthase
483.	I6XI16	polyketide cyclase

Table C.2: List of different classes of cellular activities observed in the group of HPs obtained from *M. tuberculosis*

S. No	Uniprot ID	Predicted function
		Lipoprotein
1.	P9WM15	Lipoprotein
2.	O07726	Lipoprotein
3.	P9WLF7	lipoprotein LppN
4.	O50383	Lipoprotein
5.	I6X7F2	LppP/LprE lipoprotein
		Transcriptional proteins
6.	P9WFK5	transcriptional regulator
7.	P71704	PadR family transcriptional regulator
8.	O07434	CopY family transcriptional regulator
9.	P95215	TetR family transcriptional regulator
10.	O07254	Tetracycline Repressor, domain 2
11.	P96276	Fis family transcriptional regulator
12.	P71830	TetR family transcriptional regulator
13.	I6WZI4	LytR family transcriptional regulator
14.	O53836	TetR family transcriptional regulator
15.	Q6GZX4	Poxvirus Late Transcription Factor VLTF3 like
16.	O06534	transcription initiation protein
17.	O50432	Virulence activator alpha C-term/ Helix-turn-helix domains
18.	O50442	PucR C-terminal helix-turn-helix domain
19.	O05296	Homeodomain-like
20.	O06829	Homeodomain fold
21.	O06799	Metal-sensitive transcriptional repressor
22.	O06805	ANTAR (AmiR and NasR transcription antitermination regulators) domain
23.	P9WLR3	HTH-type transcriptional regulator
24.	P9WLM5	"Winged helix" DNA-binding domain/MarR-like transcriptional regulators
25.	O53464	Homeodomain-like
26.	P9WPH5	PucR C-terminal helix-turn-helix domain
27.	P9WLC5	Homeodomain-like/ "winged helix" repressor DNA binding domain
28.	P71885	MarR family transcriptional regulator
29.	O05828	transcriptional activator protein
30.	O05823	Transcription regulator of the Arc/MetJ class
31.	P95021	transcription antitermination protein NusB
32.	P96872	Cell envelope-related transcriptional attenuator domain

33.	O06347	LytR family transcriptional regulator
34.	I6X7F9	PadR family transcriptional regulator
35.	P9WKW7	Transcription elongation factor, GreA/GreB, C-term
		Ribosomal protein
36.	I6X8G2	Sigma 54 modulation/S30EA ribosomal protein, C-terminal
37.	P9WMA9	Sigma 54 modulation/S30EA ribosomal protein C terminus
38.	P9WL63	50S ribosomal protein L30e-like
39.	O05886	Sigma 54 modulation protein / S30EA ribosomal protein
		Bacteriophage related proteins
40.	P9WLM3	Phage envelope protein
41.	O53468	Phage derived protein Gp49-like
42.	I6YA42	phage Gp37Gp68 family protein
43.	O53332	Phage derived protein Gp49-like
		DNA recombination/Replication protein
44.	O53500	SWIM zinc finger
45.	I6XEF6	DNA recombination-mediator protein A
46.	P9WL29	DNA recombination-mediator protein A
		Translation protein
47.	P71804	TfuA-like protein
48.	O86327	ribosome silencing factor RsfS
		Protein binding
49.	P9WKN1	Von Willebrand factor type A
50.	O53703	VWA domain-containing protein
51.	P9WLX5	VWA domain-containing protein
52.	P9WLQ9	Von Willebrand factor (vWF)
53.	O33260	PAC2 (Proteasome assembly chaperone) family
54.	P71923	VWA domain-containing protein
55.	P71640	sigma-70 family RNA polymerase sigma factor
56.	O53260	amino acid-binding protein
57.	O50389	Forkhead-associated domain (phosphopeptide recognition domain)
58.	O69686	FAD binding domain-containing protein
59.	P72038	sigma-70 family RNA polymerase sigma factor
		Structural motifs
60.	O05574	Zinc ribbon domain
		Metal binding protein
61.	O53728	dinb family like domain
62.	P9WF27	YcaO cyclodehydratase, ATP-ad Mg2+-binding
63.	I6XY36	Flavin-binding protein dodecin

64.	O86372	Cupin domain
65.	O06797	Putative heavy-metal-binding
66.	P95285	DinB superfamily
67.	P9WLP3	Hemerythrin HHE cation binding domain-containing protein
68.	O06194	Cupin superfamily barrel domain protein
69.	P9WL59	Four Helix Bundle (Hemerythrin (Met), subunit A)
70.	P9WGC3	Fe-S metabolism associated domain
71.	O06336	Cupin domain
72.	P72040	DinB superfamily
		Cell synthesis and regulation protein
73.	O05573	Flp pilus assembly protein RcpC/CpaB
74.	O53400	cell filamentation protein Fic
75.	O05310	cell division protein DivIVA
76.	O33184	DivIVA Family
77.	P9WK27	septum formation inhibitor Maf
78.	O69623	pilus biosynthesis protein TadE
79.	I6XI06	pknH-like extracellular domain protein
80.	P9WKW5	Septum formation
		Virulence related protein
81.	P9WM99	Excreted virulence factor EspC, type VII ESX diderm
82.	I6X8R5	Hemophore
83.	I6Y4Y1	Toxin-antitoxin system
84.	P96357	Toxin-antitoxin system
85.	O53943	EspG family (These proteins are involved in the ESAT-6 secretion system 1 (ESX-1) of Mycobacterium tuberculosis)
86.	P9WLQ7	Hemolysins and related proteins containing CBS domains - like domain
87.	O53976	cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins (CAP domain)
88.	P9WIP5	protein mbtH
89.	P71738	PemK-like, MazF-like toxin of type II toxin-antitoxin system
90.	O53199	Antitoxin
91.	P9WL71	Toxin
92.	I6X520	toxin-antitoxin system toxin component
93.	I6YBX3	Pentapeptide repeat-like
		Tetratricopeptide related protein
94.	O05304	Serine Threonine Protein Phosphatase 5, Tetratricopeptide repeat
95.	O33193	TPR repeat-containing protein
96.	I6YG27	Serine Threonine Protein Phosphatase 5, Tetratricopeptide repeat

	Degradation protein	
97.	O07755	Ethyl tert-butyl ether degradation EthD - like domain
	Lipoplysacaride/ polysaccharide associated protein	
98.	P9WLX9	Ricin-type beta-trefoil lectin domain
99.	O86358	Putative glycolipid-binding
	Immunity protein	
100.	P9WLW5	WbqC-like protein
101.	O05443	Immunity protein 61
	Regulatory protein	
102.	P9WMA1	Biofilm regulator BssS
103.	P9WMB1	Biofilm regulator
104.	P95225	Type II toxin-antitoxin system, antitoxin Phd/YefM
105.	O53702	ParD-like antitoxin of type II bacterial toxin-antitoxin system
106.	O50393	Roadblock/LC7 domain
	Nucleotide Binding Proteins	
107.	P9WPI5	CobW/HypB/UreG, nucleotide-binding domain
108.	O07239	Molecular chaperone DnaK
109.	P9WFK9	YajQ family cyclic di-GMP-binding protein
110.	P9WFN5	phosphatidylethanolamine-binding protein
111.	P9WFN1	Phosphatidylethanolamine-binding protein
112.	P9WLH3	Zn-ribbon protein%2C possibly nucleic acid-binding protein
113.	Q6MWZ8	ATP-binding protein
114.	P9WKZ5	FAD/NAD(P)-binding domain
	DNA Binding Proteins	
115.	O07173	DNA-binding protein
116.	O53652	lambda repressor-like DNA-binding domains
117.	P9WKS5	antitoxin VapB
118.	I6WZ26	"winged helix" repressor DNA binding domain
119.	I6Y8F7	"winged helix" repressor DNA binding domain
120.	P9WKR9	Copper fist DNA binding domain
121.	O06578	"winged helix" repressor DNA binding domain
122.	P9WM37	site specific recombinase
123.	P71983	"Winged helix" DNA-binding domain
124.	O06243	DNA-binding protein, CopG family
125.	O53205	single-strand DNA-binding protein
126.	I6XEH5	lambda repressor-like DNA-binding domains
127.	I6YDM0	"winged helix" repressor DNA binding domain
128.	O06195	"winged helix" repressor DNA binding domain

129.	O07196	DNA-binding protein
130.	I6XFF7	DNA-binding protein
131.	P9WL17	DNA-binding protein
132.	O53334	lambda repressor-like DNA-binding domains
133.	P9WNR9	DNA-binding protein%2C YbaB/EbfC family
134.	L0TGF0	DNA-binding protein
135.	O53598	single-stranded DNA-binding protein
		RNA Binding Proteins
136.	P9WFL1	RNA-binding protein
137.	P96389	NYN domain-containing protein
138.	P9WFM7	K homology RNA-binding domain, type I
		Signal transduction protein
139.	P9WKU9	Putative bacterial sensory transduction regulator
140.	P9WKS7	FIST C & N domain
141.	O05446	Tat (twin-arginine translocation) pathway signal sequence containing protein

Table C.3: List of predicted transport mechanisms related HPs present in the genome of *M. tuberculosis*

S. No	Uniprot ID	Predicted functions
		Transporters and Carrier Proteins
1.	P96827	PPE family protein
2.	I6Y870	ABC transporter permease
3.	O07790	ABC transporter permease
4.	I6WZD7	Nuclear Transport Factor 2; Chain: A,
5.	I6XW93	Nuclear Transport Factor 2; Chain: A,
6.	P9WFG9	Lipocalin (Iron Transport)
7.	P71568	Sulphate Transporter and Anti-Sigma factor antagonist domain
8.	O53892	multidrug resistance membrane efflux protein emrB
9.	P9WIM7	RDD family
10.	P9WFM3	AI-2E family transporter
11.	O86315	The CBS domain/MgtE intracellular N domain
12.	P9WM31	H+-ATPase
13.	L7N692	DoxX Superfamily
14.	P71796	Sulfite exporter TauE/SafE
15.	O33182	Phosphonate ABC transporter, periplasmic phosphonate binding protein
16.	O33187	Tetracycline Repressor, domain 2
17.	Q79FL4	multidrug resistance protein EmrB
18.	P9WFP3	CBS-domain pair/ FAD-binding/transporter-associated domain-like
19.	O07742	Lipocalin
20.	O07728	STAS (after Sulphate Transporter and AntiSigma factor antagonist) domain
21.	O53486	Nuclear Transport Factor 2; Chain: A,
22.	O07257	Zinc-uptake complex component A periplasmic (ABC transporter)
23.	P9WLD9	Na+/H+ antiporter
24.	P9WPI7	ABC-type cobalt transport system, permease component
25.	P71757	Na+-dependent bicarbonate transporter superfamily
26.	P71754	Mce associated membrane protein
27.	P71729	ComEC/Rec2-related protein - like domain
28.	P71728	competence protein ComEA helix-hairpin-helix repeat region
29.	I6X4W0	STAS (after Sulphate Transporter and AntiSigma factor antagonist) domain
30.	P9WL25	Nuclear transport factor 2 (NTF2-like) superfamily
31.	P9WL15	F1F0 ATP synthase subunit B, membrane domain

32.	P95244	polymorphic PE/PPE family protein
33.	I6YAV3	DoxX Superfamily
34.	O05882	potassium transporter TrkA
35.	I6YG83	Nuclear Transport Factor 2; Chain: A,
36.	O06349	DoxX Superfamily
37.	P72035	Nuclear Transport Factor 2; Chain: A,
38.	O07801	PE-PPE domain-containing protein

Appendix D

Function prediction of HPs with low precision

Table D.1: List of less precisely annotated HPs present in the genome of *M. tuberculosis*

S. No	Uniprot ID	Protein function
1.	P9WM97	Spermidine/putrescine-binding periplasmic protein - like domain
2.	P71599	Carbamoyl-phosphate synthetase large chain, oligomerisation - like domain
3.	P9WM95	Transposase
4.	P9WM85	ABC-type sugar transport system, permease component - like domain
5.	P9WM77	Fe-S oxidoreductase - like domain
6.	O53604	Glutathione S-transferase - like domain
7.	I6X8E6	CBS domain
8.	L7N686	Glycosyl transferase, group 1 - like domain
9.	Q10891	ATP-binding region, ATPase-like - like domain
10.	O53630	Glycosyltransferase - like domain
11.	O07172	adenosylmethionine--8-amino-7-oxononanoate transaminase
12.	I6WXK8	Papillomavirus major capsid L1 (late) protein - like domain
13.	O07428	Prephenate dehydrogenase - like domain
14.	O07429	putative metallohydrolase
15.	O07437	Phytanoyl-CoA dioxygenase family protein
16.	P9WLB1	SecY protein - like domain
17.	O53672	Peptidase S1 and S6, chymotrypsin/Hap - like domain
18.	L7N694	Formate dehydrogenase/DMSO reductase, domains 1-3
19.	O53682	PEP-utilising enzyme, mobile region - like domain
20.	O07234	Protein kinase - like domain
21.	O07238	3-oxoacyl-[acyl-carrier protein] reductase - like domain
22.	P9WL03	RmlC-like cupins
23.	O33270	Transglutaminase, two C-terminal domains
24.	O06300	glutamate 5-kinase
25.	O06310	Nuclease subunit of the excinuclease complex - like domain
26.	O53713	Collagen triple helix repeat - like domain
27.	O53716	Amino acid transporters - like domain
28.	O53718	Amidohydrolase 1 - like domain
29.	P95202	von Willebrand factor, type A - like domain
30.	P95203	Glycosyl transferase, family 28
31.	I6Y3N9	Outer membrane protein - like domain
32.	P95206	Acyl CoA:acetate/3-ketoacid CoA transferase, alpha subunit - like domain
33.	P96270	DEAD/DEAH box helicase, N-terminal - like domain
34.	O53733	P-loop containing nucleoside triphosphate hydrolases

35.	O53740	Helicase, C-terminal - like domain
36.	O53744	UDP-glucose 4-epimerase
37.	O53745	Putative holin
38.	P9WKV9	ADP-ribosylation
39.	P9WKV5	Alpha-D-Glucose-1,6-Bisphosphate; Chain A, domain 4
40.	Q6MX36	Glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain
41.	P9WKU7	Positive stranded ssRNA viruses
42.	P9WKU5	SH2 domain
43.	V5QPR5	Terpenoid synthases
44.	O06409	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase
45.	P9WKL3	Phosphomannomutase - like domain
46.	P9WM83	Cell growth inhibitor/plasmid maintenance toxic component
47.	P9WM81	Alkaline phosphatase - like domain
48.	O53777	Acetyl-CoA synthase (ACS)
49.	O07771	PRTase-like
50.	Q79FY5	Thiamin diphosphate-binding fold
51.	I6Y481	Glutathione synthetase ATP-binding domain-like
52.	I6XVR9	Peptidase C19, ubiquitin carboxyl-terminal hydrolase 2 - like domain
53.	I6X9E8	GHMP Kinase, C-terminal domain
54.	P96923	Bacterial regulatory protein, LuxR - like domain
55.	I6WZ30	Nitrous oxide reductase, N-terminal domain
56.	I6Y4G1	ClpP/crotonase
57.	P95044	Carbohydrate binding domain
58.	P95046	Sialidases
59.	P9WKS1	E set domains
60.	O53803	Tetrahydrobiopterin biosynthesis enzymes-like
61.	I6WZ83	Nitrile hydratase alpha chain
62.	O53813	Tetrahydrobiopterin biosynthesis enzymes-like
63.	I6Y4U0	Putative bacterial sensory transduction regulator YbjN
64.	I6Y8U3	Sterol carrier protein, SCP
65.	I6XWB9	"Winged helix" DNA-binding domain
66.	O53837	Immunoglobulin-like beta-sandwich
67.	O53842	Ribonuclease H-like
68.	P9WKR7	Bacterial transcription activator, effector binding - like domain
69.	P9WKR3	Molybdenum cofactor biosynthesis proteins
70.	P9WKP5	Phosphoenolpyruvate/pyruvate domain
71.	O05910	Toxin-antitoxin system antidote Rhh family
72.	P9WKN9	Positive stranded ssRNA viruses

73.	P9WKM5	Zn-binding ribosomal proteins
74.	P9WKM3	Sterile alpha motif SAM - like domain
75.	P9WKM1	Amidase - like domain
76.	P9WKL9	beta and beta-prime subunits of DNA dependent RNA-polymerase
77.	O05582	Six-hairpin glycosidases
78.	O05592	Beta-galactosidase GanA
79.	O05597	Enolase C-terminal domain-like
80.	P96375	septum formation initiator subfamily protein
81.	P96353	"winged helix" repressor DNA binding domain
82.	O53431	Ribonucleotide reductase Class Ib, NrdI - like domain
83.	O53448	Concanavalin A-like lectins/glucanases
84.	P9WM59	RNA-binding region RNP-1 (RNA recognition motif) - like domain
85.	O86351	Haem peroxidase, plant/fungal/bacterial - like domain
86.	O06568	Heme-dependent catalase-like
87.	O06577	Condensation domain
88.	O06583	putative transmembrane protein
89.	O06555	P-loop containing nucleoside triphosphate hydrolases
90.	O50427	Anthranilate synthase component I and chorismate binding protein - like domain
91.	O05312	Cell division GTPase - like domain
92.	O86316	X-Pro dipeptidyl-peptidase
93.	P9WM49	Chaperone J-domain
94.	P9WM45	Phosphatidylinositol 3- and 4-kinase, catalytic - like domain
95.	P9WM43	Thiamin diphosphate-binding fold (THDP-binding)
96.	P9WM35	Deoxyhypusine synthase - like domain
97.	Q79FQ6	N-terminal nucleophile aminohydrolases (Ntn hydrolases)
98.	P9WM33	beta and beta-prime subunits of DNA dependent RNA-polymerase
99.	P9WM29	Trypsin-like serine proteases
100.	P9WM27	"Helical backbone" metal receptor
101.	P9WM19	Homing endonucleases
102.	P9WM17	P-loop containing nucleoside triphosphate hydrolases
103.	P9WM09	Aminopeptidase/ Ribosomal-protein-alanine N-acetyltransferase
104.	V5QQR7	Anticodon-binding - like domain
105.	P71802	Nucleotide-diphospho-sugar transferases
106.	P71810	D-threonine aldolase, metal-activated pyridoxal enzyme
107.	O06827	SecA DEAD-like - like domain
108.	O06824	ABC transporter related - like domain
109.	O06823	proline and glycine and valine rich secreted protein
110.	O06816	Alpha-D-Glucose-1,6-Bisphosphate; Chain A, domain 4

111.	P9WLX3	Phosphoenolpyruvate/pyruvate domain
112.	L7N6B6	FAD/NAD(P)-binding domain
113.	P9WLW1	Cysteine proteinases
114.	O06180	PHD finger superfamily
115.	P9WLT9	REP13E12 repeat protein
116.	P9WLT7	Zinc beta-ribbon
117.	O06133	ABC transporter involved in vitamin B12 uptake, BtuC
118.	O06136	Phosphatidylglycerol lysyltransferase
119.	O06149	FomD barrel-like domain
120.	L7N673	FMN-linked oxidoreductases
121.	P94979	P-loop containing nucleoside triphosphate hydrolases
122.	O86371	FAD/NAD(P)-binding domain
123.	O33195	Myosin tail - like domain
124.	P71982	Nucleic acid-binding proteins
125.	P9WLS3	Phospholipase D/nuclease
126.	O33271	Restriction endonuclease-like
127.	P72005	S-adenosyl-L-methionine-dependent methyltransferases
128.	O06790	FAD linked oxidase, C-terminal - like domain
129.	O06796	N-terminal domain of bifunctional PutA protein
130.	O33178	HNH nuclease - like domain
131.	O33181	Glutamate-ammonia ligase adenylyltransferase - like domain
132.	O53931	Ribosomal protein L18P/L5E - like domain
133.	O53953	glycine cleavage system protein P
134.	O53954	EF-Tu/eEF-1alpha/eIF2-gamma C-terminal domain
135.	O07222	SecB-like
136.	P9WLS1	Immunoglobulin V-set - like domain
137.	P9WFG1	Trimeric LpxA-like enzymes
138.	P9WLR7	Respiratory nitrate reductase 1 gamma chain
139.	P9WFG3	Bacterial hemolysins
140.	O07756	Calpastatin
141.	O07745	Ethanolamine ammonia lyase large subunit - like domain
142.	O07741	Pyruvate-ferredoxin oxidoreductase, PFOR, domain III
143.	O07739	"Winged helix" DNA-binding domain
144.	P9WFQ1	Alpha-Beta Plaits
145.	O07719	Ribulose-phoshate binding barrel
146.	P95289	Organic anion transporter polypeptide OATP - like domain
147.	P95287	Acetoacetate decarboxylase-like
148.	P95267	Orn/DAP/Arg decarboxylase 2 - like domain
149.	P95266	Cytochrome P450 - like domain

150.	L0T9Q6	Iojap-related protein - like domain
151.	P95264	Transcriptional regulator, Rrf2 - like domain
152.	P95263	Cytochrome c oxidase, subunit I - like domain
153.	P9WLQ3	Aconitate hydratase, N-terminal - like domain
154.	P95256	Cytochrome P450 - like domain
155.	P95253	C-type lectin - like domain
156.	P9WLP5	Organic radical activating enzymes - like domain
157.	O53462	Fibronectin, type III - like domain
158.	O53465	Alpha-2-macroglobulin, N-terminal 2 - like domain
159.	O53469	P-loop containing Nucleoside Triphosphate Hydrolases
160.	O53477	von Willebrand factor, type A - like domain
161.	O53487	Aldehyde dehydrogenase - like domain
162.	O53491	Response regulator receiver - like domain
163.	P9WLL3	Rhodopsin-like GPCR superfamily - like domain
164.	L7N6B8	Coagulation factor 5/8 type, C-terminal - like domain
165.	P9WLK9	Repulsive guidance molecule, N-terminal - like domain
166.	Q10690	Methyltransferase type 12 - like domain
167.	P9WLK3	WD-40 repeat - like domain
168.	P9WLK1	Peptidase M23B - like domain
169.	O33249	Fibronectin, type III - like domain
170.	O33252	Alpha-Beta Plaits
171.	O06241	Endonuclease/exonuclease/phosphatase - like domain
172.	O06238	Predicted enzyme with a TIM-barrel fold - like domain;
173.	O53506	Methylenetetrahydrofolate reductase
174.	O53517	UvrD/REP helicase - like domain
175.	O53523	PAP/25A-associated - like domain
176.	P9WLH1	AAA ATPase - like domain
177.	I6XDU8	Formylmethionine deformylase - like domain
178.	P9WLG9	AAA ATPase - like domain
179.	P9WLG7	ADP-ribosylglycohydrolase - like domain
180.	O53527	Aldo/keto reductase - like domain
181.	O53530	Glutamyl-tRNA reductase - like domain
182.	P9WLF9	AIR (aminoimidazole ribonucleotide) synthase related protein
183.	P9WLE9	Mediator of RNA pol II transcription subunit 19
184.	P9WLE5	Isocitrate dehydrogenase NADP-dependent, monomeric type - like domain
185.	P9WFL7	Protein of unknown function UPF0167 - like domain
186.	P9WLD5	cytochrome P450
187.	P9WLD3	AAA ATPase, central region - like domain

188.	L7N683	Ribosomal protein L25 (general stress protein Ctc) - like domain
189.	L7N666	Harpin-induced 1 - like domain
190.	P9WLB9	Sigma-70 region 3 - like domain
191.	P95232	Bacterial chemotaxis sensory transducer - like domain
192.	P95238	Formylmethionine deformylase - like domai
193.	O05838	Catechol dioxygenase, N-terminal - like domain
194.	L0TC46	ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase
195.	O86326	Peptidase dimerisation - like domain
196.	P71917	Exoribonuclease - like domain
197.	Q79FD9	Cellulose synthase operon C, C-terminal - like domain
198.	O53179	Transmembrane domain
199.	P9WLA7	Glucose-6-phosphate dehydrogenase - like domain
200.	I6YDH3	MHC class I, alpha chain, alpha1 and alpha2 - like domain
201.	I6X4D6	EPSP synthase (3-phosphoshikimate 1-carboxyvinyltransferase) - like domain
202.	O53206	Aldose 1-epimerase - like domain
203.	O06175	AAA ATPase - like domain
204.	I6Y0Y0	Spectrin repeat - like domain
205.	O53222	WD-40 repeat - like domain
206.	I6XEK2	Hemagglutinin
207.	P9WLA3	Alpha-Beta Plaits
208.	P9WL91	YjbR
209.	O06198	PAS - like domain
210.	P9WL65	ComEC/Rec2-related protein - like domain
211.	P9WL61	Formyl transferase, N-terminal - like domain
212.	P9WL57	Bacterial NAD-glutamate dehydrogenase - like domain
213.	P9WL53	Phage tail assembly chaperone gp38 - like domain
214.	P9WL51	Nuclear protein SET - like domain
215.	I6Y1F5	Superfamily I DNA and RNA helicases and helicase subunits - like domain
216.	P71958	Phosphopentomutase - like domain
217.	P71959	Molecular chaperone (small heat shock protein) - like domain
218.	I6Y9Z5	Transposase, IS4 - like domain
219.	O86317	CheW-like protein - like domain
220.	I6XF31	Adenylylsulfate kinase - like domain
221.	O07207	Sushi/SCR/CCP - like domain
222.	I6X562	Shikimate kinase - like domain
223.	I6YE70	Arginine biosynthesis protein ArgJ - like domain
224.	O33227	Phosphatidylinositol-specific phospholipase C, Y domain - like domain

225.	I6Y1K4	Protein kinase - like domain
226.	I6Y1K7	S-adenosylmethionine-dependent methyltransferases (SAM or AdoMet-MTase), class I
227.	I6X5B0	Winged helix DNA-binding domain
228.	O33316	AAA ATPase - like domain
229.	P71653	HEC/Ndc80p - like domain
230.	I6XFC2	Baculovirus Y142 protein - like domain
231.	I6YEE9	EGF-like, laminin - like domain
232.	P71643	Fungal transcriptional regulatory protein, N-terminal - like domain
233.	I6YEF3	Ribonuclease HII/HIII - like domain
234.	P71627	Proteinase inhibitor I2, Kunitz metazoa - like domain
235.	I6X5G8	Glyoxal oxidase, N-terminal - like domain
236.	I6YAC9	Alpha/beta hydrolase fold-3 - like domain
237.	I6Y1W7	Glycyl-glycine endopeptidase LytM
238.	P9WFM9	Restriction endonuclease-like
239.	P9WL27	GMP Synthetase; Chain A, domain 3
240.	P9WL13	Low molecular weight phosphotyrosine protein phosphatase - like domain
241.	P95133	Carboxylesterase, type B - like domain
242.	P95115	MutS III - like domain
243.	O53245	Formate dehydrogenase accessory protein - like domain
244.	I6YEY1	MgtC/SapB transporter - like domain
245.	I6X630	Ferredoxin - like domain
246.	I6YAY5	Aldose 1-epimerase - like domain
247.	I6X642	Heat shock protein DnaJ, N-terminal - like domain
248.	I6YB21	Fimbrial assembly - like domain
249.	P95087	RhoGAP - like domain
250.	P9WL11	Protein of unknown function, DUF488
251.	O53298	Nucleotidyl transferase of unknown function (DUF2204)
252.	O05776	Curli production assembly/transport component CsgG - like domain
253.	O05785	ATP-dependent Zn proteases - like domain
254.	O07033	Excinuclease ABC, C subunit, N-terminal - like domain;
255.	O07034	Hpt - like domain
256.	P9WL09	Protein of unknown function DUF1577 - like domain
257.	I6Y2Q7	Transcriptional regulatory protein, C-terminal - like domain
258.	P95184	Peptidase M16, C-terminal - like domain
259.	O53315	Exonuclease - like domain
260.	O53316	Glycoside hydrolase, family 19 - like domain
261.	O53319	Sulfotransferase - like domain

262.	O53322	Glycoside hydrolase, family 32 - like domain
263.	O53336	Chorismate synthase - like domain
264.	I6XGJ1	ExsB - like domain
265.	L7N668	RNA polymerase I, subunit RPA14, fungi - like domain
266.	O05844	RNA binding S1 - like domain
267.	O05888	Lipoxygenase - like domain
268.	O05899	Glucose-6-phosphate isomerase like protein; domain 1
269.	O53351	Peptidase S21, herpesvirus maturational proteinase, assemblin - like domain
270.	P96874	Pimeloyl-CoA synthetase - like domain
271.	P9WL01	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase
272.	O53362	primosomal protein
273.	P9WKZ9	P-loop containing nucleotide triphosphate hydrolases
274.	P9WKY9	Histidine kinase A, N-terminal - like domain
275.	P9WKY3	Citrate synthase - like domain
276.	O06257	Adenylyl cyclase class-3/4/guanylyl cyclase - like domain
277.	P9WKY1	4-diphosphocytidyl-2C-methyl-D-erythritol synthase - like domain
278.	I6YC91	Predicted transcriptional regulators - like domain
279.	I6XHD1	Peroxiredoxin - like domain
280.	I6YGE4	Caffeic acid 3-O-methyltransferase 1
281.	I6XHH2	Signal peptide protein
282.	I6X7P2	Aldo/keto reductase - like domain
283.	I6YGR2	Relaxase - like domain
284.	O06270	Histidinol dehydrogenase - like domain
285.	I6YGU1	C-5 cytosine-specific DNA methylase - like domain
286.	I6Y464	NAD+ synthase - like domain
287.	O06364	Helix-turn-helix type 3 - like domain
288.	I6Y469	Wnt superfamily - like domain
289.	I6YCP0	AAA ATPase
290.	O69624	Predicted GTPases (dynamin-related) - like domain
291.	I6YCP7	Oxidoreductase
292.	I6X824	Glycosyltransferases involved in cell wall biogenesis - like domain
293.	O69654	Pyridine nucleotide-disulphide oxidoreductase, NAD-binding region - like domain
294.	O69659	Electron transport accessory proteins
295.	I6Y4C4	Response regulator receiver - like domain;
296.	O69699	RNA-binding region RNP-1 (RNA recognition motif) - like domain
297.	O69712	Cysteinyl-tRNA synthetase, class Ia - like domain
298.	O69714	Molybdopterin binding domain - like domain

299.	O69715	Lipoxygenase, LH2 - like domain
300.	P72036	6-Phosphogluconate Dehydrogenase, domain 3
301.	P9WKX3	Peptidase family M23 (zinc metallopeptidases)
302.	O07804	AAA ATPase - like domain
303.	P96247	Carbohydrate binding family 25 - like domain
304.	P96227	Secretion protein HlyD - like domain
305.	O07036	RNA-binding region RNP-1 (RNA recognition motif) - like domain
306.	O05447	RNA-binding region RNP-1 (RNA recognition motif) - like domain
307.	O05436	Protein-glutamine gamma-glutamyltransferase E

FINAL

Send Orders for Reprints to reprints@benthamscience.ae

Current Topics in Medicinal Chemistry, 2016, *16*, 000-000

Current Advances in the Identification and Characterization of Putative Drug and Vaccine Targets in the Bacterial Genomes

Mohd. Shahbaaz[a], Krishna Bisetty[a], Faizan Ahmad[b] and Md. Imtaiyaz Hassan[b,*]

[a]*Department of Chemistry, Durban University of Technology, Durban – 4000, South Africa;* [b]*Center for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, Jamia Nagar, New Delhi – 110025, India*


Md. Imtaiyaz Hassan

Abstract: The development in sequencing technologies over the past few decades have increased the pace of decoding genetic and functional information present in the genomes of pathogenic microorganisms. The knowledge obtained through sequencing projects facilitated the identification of genes that codes for virulence factors. A major portion of genomes of pathogenic of bacteria contains genes which are classified as "hypothetical or uncharacterized". Due to unavailability of precise information about the functionality of these genes, the pathogenic mechanisms utilized by varieties of microorganisms are not fully understood. This respective class of proteins draws a significant interest of pharmaceutical research as they have potential to provide new clues regarding the development of novel therapeutics particularly against the multidrug resistant strains of bacteria. The *in silico* identification of putative drug and vaccine targets in the set of uncharacterized proteins through comparative and subtractive genome analyses facilitates the increase usability and efficiency of the present drugs. The functional annotation of these characterized target proteins can uncover varieties of biochemical pathways important for the survival and pathogenesis of bacteria. This review focuses on the current protocols available for identification and functional annotations of these uncharacterized potential therapeutic targets.

Keywords: Hypothetical proteins, Drug targets, Genomic analyses, Sequence based function prediction, Structure-Function relationship, Molecular Dynamics simulations.

1. INTRODUCTION

The extensive use of antibiotics in various therapies leads to the development of resistance against multiple dugs in the pathogenic bacteria [1-4]. The multidrug resistance (MDR) primarily occur by the accumulation of multiple genes each resistant to a given drug on R plasmid and by over expression of genes encoding the multidrug efflux pumps which generally belong to Major Facilitator Superfamily [1-4]. Therefore, the genomic information of these pathogenic microorganisms is necessary in order to understand the mechanisms of MDR more precisely [1-4]. Through the commencement of high-throughput sequencing techniques, the generation of genomic data from clinically-significant microorganisms has become increasingly efficient [5]. However, the comparative genomic attempts are focused on microorganisms involved in the pathogenesis, deeper understanding of their biology as well as molecular mechanisms [1, 2, 5]. These analyses also enable the detailed comprehension of MDR mechanisms and the factors responsible for its occurrence [1, 2, 5, 6].

The bacterial genomes generally contain a circular chromosome of size ranges from 0.5-10 Mb [5]. However, pathogenic species such as *Vibrio*, *Leptospira* and *Burkholderia* contain more than one chromosome [5]. Such bacteria also contain plasmids of variable length harbor genes that may provide an advantage in the form of MDR to the bacteria carrying them [5]. The GC content in the bacterial genomes ranging from 25% in the genus of the *Mycoplasma* to roughly 75% in *Micrococcus* species [5]. The genomes of pathogenic bacteria are dynamic in nature due to frequent exposure to mutagens, chromosomal aberrations and horizontal gene transfer thereby resulting in the alteration in their metabolic capabilities [5]. The latter mechanism leads to the transfer of MDR genes and other virulent factors [5].

The major portion of the sequenced bacterial genomes was annotated as "hypothetical or uncharacterized" because of the unavailability of experimentally proven functionality despite their occurrence in varieties of phylogenetic lineages [7]. These hypothetical proteins (HPs) can be used as potential targets in the process of drug repositioning by utilizing the present drugs for the latest therapeutic indications [8-11]. Generally, most drugs molecules are designed to interact and inhibit the target proteins for a particular disease [12]. The drugs may also interact with off-target HPs that are not their primary targets which result in unanticipated side effects [12]. The side effects of a drug may yield a complex trends that can be understand by analyzing its interactions with off-targets like HPs [12].

This article reviews the developments in new methodologies for the characterization of potential drug targets among the group of HPs in bacterial genomes and various strategies available for functional annotations. The identification of

*Address correspondence to this author at the Center for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, Jamia Nagar, New Delhi – 110025, India; E-mail: mihassan@jmi.ac.in

1568-0266/16 $58.00+.00

OMICS A Journal of Integrative Biology
Volume 19, Number 7, 2015

DOI: 10.1089/omi.2015.0032

Towards New Drug Targets? Function Prediction of Putative Proteins of *Neisseria meningitidis* MC58 and Their Virulence Characterization

Mohd. Shahbaaz,[1] Krishna Bisetty,[1] Faizan Ahmad,[2] and Md. Imtaiyaz Hassan[2]

Abstract

Neisseria meningitidis is a Gram-negative aerobic diplococcus, responsible for a variety of meningococcal diseases. The genome of *N. meningitidis* MC58 is comprised of 2114 genes that are translated into 1953 proteins. The 698 genes (~35%) encode hypothetical proteins (HPs), because no experimental evidence of their biological functions are available. Analyses of these proteins are important to understand their functions in the metabolic networks and may lead to the discovery of novel drug targets against the infections caused by *N. meningitidis*. This study aimed at the identification and categorization of each HP present in the genome of *N. meningitidis* MC58 using computational tools. Functions of 363 proteins were predicted with high accuracy among the annotated set of HPs investigated. The reliably predicted 363 HPs were further grouped into 41 different classes of proteins, based on their possible roles in cellular processes such as metabolism, transport, and replication. Our studies revealed that 22 HPs may be involved in the pathogenesis caused by this microorganism. The top two HPs with highest virulence scores were subjected to molecular dynamics (MD) simulations to better understand their conformational behavior in a water environment. We also compared the MD simulation results with other virulent proteins present in *N. meningitidis*. This study broadens our understanding of the mechanistic pathways of pathogenesis, drug resistance, tolerance, and adaptability for host immune responses to *N. meningitidis*.

Introduction

Neisseria meningitidis MC58 belongs to the Gram-negative bacterial family of Neisseriaceae, and is an encapsulated and delicate aerobic diplococcus bacterium. Among the primary sources of bacterial meningitis worldwide, it is believed that only *N. meningitidis* can trigger epidemic conditions by causing manifestations such as pneumonia and sepsis (Jafri et al., 2013). This bacterium generally inhabits the cavity of the patient's nasopharynx. It causes life-threatening meningococcal diseases in children, especially in industrialized countries of Asia and Africa (Stephens et al., 2007). *N. meningitidis* is classified into 12 well-defined serogroups (viz., A, B, C, W, X, and Y) on the basis of the outcomes of genome typing techniques and their structural distinctions present in capsular polysaccharides, outer membrane proteins, as well as lipo-oligosaccharides (Broker et al., 2014; Rouphael and Stephens, 2012). Various recombinations and horizontal exchange of genes within the meningococcal genomes are responsible for the antigenic diversity among colonel complexes, leading to a distinctive clone expression (Read, 2014).

A variety of virulence factors are responsible for the pathogenesis of *N. meningitidis,* including capsular polysaccharides (CPS), lipo-oligosaccharide (LOS), and adhesins (Hill et al., 2010). The CPS, considered to be a key virulence factor, is comprised of N-acetyl-mannosamine-1-phosphate subunits with phosphodiester linkages (Fiebig et al., 2014). CPS enables this pathogen to escape the phagocytic complement-mediated mechanisms of the host and also forms the source for immunological serogrouping (Stephens et al., 2007). The sequenced genome of the MC58 strain of *N. meningitidis* contains 2,272,351 bps that are expressed in about 1953 proteins. It shows the presence of 2158 coding regions, of which 53.7% of the coding regions in the genome were allocated a biological function, while 35% were found to be ''hypothetical proteins (HPs)'' (Tettelin et al., 2000).

HPs are predicted proteins with no experimental validation at the biochemical level of protein expression (Shahbaaz et al., 2014a; 2014b). Approximately half of the proteins are not functionally characterized in the majority of the sequenced bacterial genomes. Identification of their natural functions will be useful in completing the available genomic information (Loewenstein et al., 2009; Nimrod et al., 2008).

[1]Department of Chemistry, Durban University of Technology, Durban, South Africa.
[2]Center for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, Jamia Nagar, New Delhi, 110025, India.

Computational Biology and Chemistry 59 (2015) 67–80

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research article

In silico approaches for the identification of virulence candidates amongst hypothetical proteins of *Mycoplasma pneumoniae* 309

Mohd. Shahbaaz[a], Krishna Bisetty[a], Faizan Ahmad[b], Md. Imtaiyaz Hassan[b,*]

[a] Department of Chemistry, Durban University of Technology, Durban 4000, South Africa
[b] Center for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, Jamia Nagar, New Delhi 110025, India

ARTICLE INFO

Article history:
Received 16 January 2015
Received in revised form 8 September 2015
Accepted 14 September 2015
Available online 18 September 2015

Keywords:
Hypothetical proteins
Mycoplasma pneumoniae
Function predictions
Sequence analyses
Virulence factors
Molecular dynamics Simulations

ABSTRACT

Mycoplasma pneumoniae type 2a strain 309 is a simplest known bacterium and is the primary cause of community acquired pneumonia in the children. It mainly causes severe atypical pneumonia as well as several other non-pulmonary manifestations such as neurological, hepatic, hemolytic anemia, cardiac diseases and polyarthritis. The size of *M. pneumoniae* genome (Accession number: NC_016807.1) is relatively smaller as compared to other bacteria and contains 707 functional proteins, in which 204 are classified as hypothetical proteins (HPs) because of the unavailability of experimentally validated functions. The functions of the HPs were predicted by integrating a variety of protein classification systems, motif discovery tools as well as methods that are based on characteristic features obtained from the protein sequence and metabolic pathways. The probable functions of 83HPs were predicted successfully. The accuracy of the diverse tools used in the adopted pipeline was evaluated on the basis of statistical techniques of Receiver Operating Characteristic (ROC), which indicated the reliability of the functional predictions. Furthermore, the virulent HPs present in the set of 83 functionally annotated proteins were predicted by using the Bioinformatics tools and the conformational behaviours of the proteins with highest virulence scores were studied by using the molecular dynamics (MD) simulations. This study will facilitate in the better understanding of various drug resistance and pathogenesis mechanisms present in the *M. pneumoniae* and can be utilized in designing of better therapeutic agents.

1. Introduction

Mycoplasma pneumoniae is one of the smallest self-replicating bacteria which belong to the family of Mycoplasmataceae. It causes acute respiratory tract infection known as atypical pneumonia, and various types of neurological, cardiac, hepatic and hemolytic manifestations (Razin et al., 1998). *M. pneumoniae* infection occurs worldwide in an endemic fashion and is occasionally epidemic. It accounts for more than 40% of community acquired pneumonia cases in children, and 18% of cases require hospitalization (Ferwerda et al., 2001). The *M. pneumoniae* also causes several extra-pulmonary infections and the central nervous system (CNS) manifestations like meningoencephalitis, encephalitis, polyradiculitis and aseptic meningitis, which are more common among patients suffering from pneumonia caused by *M. pneumoniae* (Leonardi et al., 2005). It also induces several chronic disease conditions in which the clearance of organism from host is a very complex process (Waites et al., 2008).

Infection of *M. pneumoniae* causes various chronic diseases such as juvenile idiopathic arthritis, rheumatoid arthritis, asthma and Crohn's disease (Waites et al., 2008). *M. pneumoniae* is an obligate surface pathogen which develop adherence to mucosal epithelium of respiratory and urogenital tracts of the host by using varieties of specialized surface proteins (Hu et al., 1977). The key adherence proteins present in the *M. pneumoniae* are P1 adhesins (Hu et al., 1977) which are the primary cause of its virulence (Layh-Schmitt and Herrmann, 1994). These adhesins are also significant in defending the parasite from mucociliary clearance (Krause, 1996). The current study aimed at the annotations of uncharacterized adhesins and virulence causing proteins in the genome of *M. pneumoniae*.

The genome of *M. pneumoniae* contains 8,17,176 base pairs that forms 750 genes, 36 tRNAs genes, 4 non-coding RNA genes, and 1 rRNA operon (Kenri et al., 2012). These genes are translated into 707 proteins (Kenri et al., 2012) which are involved in varieties of functions. Around 204 proteins in the respective set are listed as "Hypothetical Proteins (HPs)" as no biochemical functions were

* Corresponding author.
E-mail address: mihassan@jmi.ac.in (M. I. Hassan).

http://dx.doi.org/10.1016/j.compbiolchem.2015.09.007
1476-9271/

COMPUTATIONAL AND EXPERIMENTAL STUDIES OF PUTATIVE VIRULENCE FACTORS OF MYCOBACTERIUM TUBERCULOSIS H37Rv

ORIGINALITY REPORT

%5	%1	%4	%1
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Khan, Faez Iqbal, Dong-Qing Wei, Ke-Ren Gu, Md. Imtaiyaz Hassan, and Shams Tabrez. "Current updates on computer aided protein modeling and designing", International Journal of Biological Macromolecules, 2016. Publication	%1
2	Idrees, Danish, Sudhir Kumar, Syed Abdul Arif Rehman, Samudrala Gourinath, Asimul Islam, Faizan Ahmad, and Md. Imtaiyaz Hassan. "Cloning, expression, purification and characterization of human mitochondrial carbonic anhydrase VA", 3 Biotech, 2016. Publication	%1
3	Shahbaaz, Mohd., Krishna Bisetty, Faizan Ahmad, and Md. Imtaiyaz Hassan. "Towards New Drug Targets? Function Prediction of Putative Proteins of Neisseria meningitidis MC58 and Their Virulence Characterization", Omics A Journal of Integrative Biology, 2015.	%1